1968

# On Gravitational Collapse and Cosmology

by

Stephen Hawking

Department of Applied Mathematics and Theoretical Physics,
University of Cambridge.

and

Roger Penrose

Department of Mathematics, Birkbeck College,
University of London .

Abstract

We present a new theorem on space-time singularities. On the basis of the Einstein (or Brans-Dicke)theory, and without using any Cauchy surface assumption, we show (essentially from the property that gravitation is always attractive) that singularities will occur if there exists either a compact spacelike hypersurface or a closed trapped surface or a point whose past light-cone starts converging again. The first condition would be satisifed by any spatially closed universe, the second by a collapsing star and the third by the observable portion of our actual universe - as we shall show follows from observations of the microwave background radiation.

## 1. Introduction

The most immediately noticable fact about gravitation is that it always seems to be attractive. Thus, despite its extreme weakness compared to electromagnetism, it is nevertheless significant for a large body such as a star since all the gravitational fields of individual particles add up while the electromagnetic fields cancel each other out. Indeed on an even larger scale, that of the whole Universe, gravitation dominates over all other forces. This attractive character and its $r^2$ dependence make gravitation rather a 'dangerous' force since it is potentially unstable: if a star is compressed slightly, the particles move closer together and so the attractice force between them will increase. Normally, of course, the repulsive pressure forces would increase by a somewhat greater amount and so restore balance. However, Chandrasekhar[1] has shown that when a star of greater than 1.5 times the solar mass exhausts its nuclear fuel and cools, the pressure forces are insufficient to resist the gravitational attraction. What then would happen to such a star? Would it collapse to some sort of singularity or would it happen that the smallest departure from spherical symmetry would cause different parts not to fall exactly towards the centre and so miss each other.

In the reverse direction in time a similar question arises in respect of the whole Universe: Was there a singularity in the past or did the Universe manage somehow to pass smoothly from a contracting phase to the present expansion? In this essay we shall present a new and very general theorem which shows that singularities would be expected both in the collapse of a star and at the beginning of the expansion of the Universe. This theorem combines and extends two previous theorems of the authors[2,3]. It is based on general relativity as this the most satisfactory theory of gravitation so far proposed but similar results would probably hold in any relativistic theory in which gravitation is attractive. In particular the theorem applies also to the theory of Brans and Dicke[4] so long as gravitation remains attractive. (If the gravitational constant were to change sign, as is in principle allowed in the Brans-Dicke theory, this could only occur via a region where it becomes infinite. In such a region, the presence of any matter would cause infinite curvature).

2. The Theorem

One might define a space-time singularity as a point at which the metric is degenerate or at which the curvature is infinite. However, such a point ought not to be regarded as part of space-time since the known laws of physics could not hold there. We shall therefore adopt the view that space-time consists only of points at which the metric is Lorentzian and suitably differentiable (say $C^2$). We then detect singular points which have been "cut out" by the existence of incomplete geodesics. This approach is discussed further in references 5 and 6.

In the first theorem[2,7,8] to use this criterion for a singularity, it was shown that there must exist incomplete null geodesics if

1) The energy-momentum tensor obeys the inequality*

$$T_{ab} V^a V^b \geqslant \tfrac{1}{2} V_a V^a T$$

for any timelike vector $V^a$.

2) There exists a noncompact spacelike hypersurface H which is a Cauchy surface (i.e. it intersects every inextendible timelike or null curve once and once only).

3) There exists a closed trapped surface T(i.e. T is a compact spacelike 2-surface such that both the 'ingoing' and the 'outgoing' families of null geodesics orthogonal to T are converging: see Fig.1).

Condition 1) is a very reasonable inequality which is satisfied by all known forms of matter. It is the condition that gravity should always be attractive. Condition 3) will be satisfied when a star collapses inside its Schwarzschild radius. One would expect stars of greater than 1.5 - 15 solar masses to do this eventually provided that their angular momemtum is sufficiently small. (This uncertainty in the mass limit results from uncertainty as to how much material is ejected during the collapse process.

The theorem will also hold if condition 3) is replaced by 3'):

3' ) There is a point p such that the divergence of the system of null geodesics through p changes sign somewhere to the part of p (Fig.2).

---

* In refs. 2,7,8 the effectively weaker assumption $R_{ab}l^a l^b \leqslant 0$, for null vectors $l^a$, was all that was used. This has the advantage that the presence of a cosmological constant does not affect the discussion. However, it is hard to believe that a non-zero $\lambda$ can produce any qualatively noticeable affects in regions of high curvature. Here we assume $\lambda = 0$. (Note that $\lambda > 0$ implies that gravitation is not 'always attractive').

It is shown in the Appendix that observations of the cosmic background radiation indicate that there is indeed enough matter on our past light-cone to cause it converge again.

It this seems reasonable to assume that conditions 1) and 3) or 3′) are satisfied. However the weakness of the theorem lies in condition 2) which requires there to exist a noncompact Cauchy surface. That it should be non-compact is not much of a restriction since it was shown in reference 3 that there will be incomplete geodesics in any 'generic' solution which satisfied condition 1) and which had a compact Cauchy surface. However the requirement that there should exist any global Cauchy surface at all is a very strong one. For being a Cauchy surface is a property not of the surface itself but of the whole space-time in which it is imbedded. There are plenty of solutions known which do not have Cauchy surfaces. Thus one might feel that this theorem does not indicate that we must expect singularities in space-time, but merely that global Cauchy surfaces do no exist in our Universe. To remove this weakness we shall present a new version of the theorem which does not require existence of a Cauchy surface This new theorem also shows that there will be a singularity if the Universe is spatially closed, that is, if there exists a compact spacelike surface (not necessarily a Cauchy surface).

The precise statement of the theorem is : Space=time is not timelike and null geodesically complete if,

(a) condition 1) holds.

(b) Every timelike and null geodecis has a point on it at which

$$k_{[a}R_{b]cd}k_{[e}k_{f]}k^{c}k^{d}$$ is nonzero, where $k^{a}$ is the tangent vector to the geodesic.

(c) There are no closed timelike curves.*

(d) One of the three following conditions holds:

(i) condition 3).

(ii) condition 3′).

(iii) There is a compact spacelike hypersurface H.

* With condition (b) this implies strong causality[11].

Condition (b) would be expected to hold in any 'generic' space-time. (It fails in certain highly spacial exact solutions, however, but this is not of interest physically). In the presence of (a), condition (b) will follow if every timelike geodesic encounters some curvature and every null geodesic contains a point at which it fails to be directed along a principal null direction of the Weyl tensor. In a physically reasonable solution, the presence of randomly oriented gravitational field (e.g. radiation) is to be expected. This would be sufficient to ensure that (b) holds.

3. Proof

We shall give the proof for cases (ii) and (iii) that for (i) is similar to (ii).

For a set S we let $I^+(S)$ and $J^+(S)$ denote the sets of points which can be reached from S by future directed timelike and nonspacelike curves respectively. We define $E^+(S)$ to be $J^+(S) - I^+(S)$. Points of $E^+(S)$ will lie on the boundary of $J^+(S)$ and will be reached from S by null geodesics. All these definitions have duals in which future is replaced by past and '+' by '-'. We define $C^+(S)$ as the set of points $p \in J^+(S)$ such that $J^-(p) \cap J^+(S)$ is a non-empty compact set on which strong causality held. We denote by $D^+(S)$ the set of all points q such that every inextensible past directed nonspacelike curve through q intersects S.

We shall assume that space-time is timelike and null geodesically complete and show that this leads to a contradiction with conditions (a), (b), (c) and (d). Consider first case (ii). $\dot{J}^-(p)$, the boundary of $J^-(p)$ will be generated by null geodesic segments which may have a past end-point but which can have a future end-point only if they pass through p. Near p, $\dot{J}^-(p)$ will be generated by the past-directed null geodesics through p. As condition (a) holds and the past directed null geodesics from p start converging again, there would be a point conjugate to p on every such geodesic within a finite affine distance from p. Points on such a geodesic beyond the conjugate point could be connected to p by a timelike curve and so would lie in the interior of $J^-(p)$. Thus $E^-(p)$ would be compact.

Let A denote $E^-(p)$ in case (ii) and H in case (iii). Then in both cases $A = E^-(A)$ will be compact. This implies that $\dot{J}^-(A) - A$, if nonempty, will be generated by null geodesic segments which have no future end-point . Now consider $C^-(A)$ which will contain $D^-(A)$[5]. Suppose $q \in C^-(A) - D^-(A)$ exists. Then there would be an inextendible future directed curve $\lambda$ from $q$ , not meeting A. Since $J^+(q) \cap J^-(A)$ would be compact and strong causality holds, $\lambda$ would have to intersect $J^-(A) - A$ at some point r. But there would be an inextendible future directed with geodesic in $\dot{J}^-(A)$ through r. This would mean that $J^+(q) \cap J^-(A)$ was not compact. This shows that $C^-(A)$ equals $D^-(A)$.

Figure 3 shows possible forms for $C^-(A)$ in the two cases. However, it should be emphasized that the diagrams are meant for illustration only and that there are other possibilities.

By an argument similar to one in reference 5, $C^-(A)$ cannot have compact closure. For if it had, we could cover $C^-(A)$ with a finite number of local causality neighbourhoods $B_i$. Then if $p_1 \in J^-(A) \cap [B_1 - \overline{C}^-(A)]$ there would be a future directed nonspacelike curve $\lambda_1$ , from $p_1$ to A which intersected $\dot{C}^-(A)$ in some other neighbourhood $B_2$. Continuing this process would exhaust all the $B_i$ and so lead to a contradiction. There must be an inextendible past directed nonspacelike curve $\lambda$ from the compact set A which remains in $C^-(A)$ since if every such curve left $C^-(A)$ it would have compact closure. As $E^-(A) \cap A$ is empty in both cases, it is possible to choose $\lambda$ to be timelike. The boundary of $J^+(\lambda)$ will be generated by null geodesic segments which cannot have past end-points. By condition (b), each of these segments will therefore have to have a future end-point as otherwise there would be points of $\dot{J}^+(\lambda)$ which had timelike separation. Every inextendible future directed nonspacelike curve from $\lambda$ will intersect A. Thus if F denotes the compact set $A \cap \overline{J}^+(\lambda)$ then $E^+(F) - F$ will consist of those points of $\dot{J}^+(\lambda)$ through which there are null geodesic generating segments which intersect A to the past. Since each such segment must have a future end-point, it follows that the set $G = F \cup E^+(F)$ is compact,

We then use similar arguments to those above to show that $\overline{D}^+(G)$ is noncompact and that there is an inextensible timelike curve $\mu$ which remains in $D^+(G)$. There will be points $q \in \lambda$ and $r \in \mu$ such $r \in I^+(q)$. Thus there will be a curve $\alpha$ which

is inextensible in both future and past directions and which remains in $D^-(F) \cup D^+(G)$. Let $a_i$ and $b_i$ be sequences of points on $\alpha \cap I^+(F)$ and $\alpha \cap I^-(F)$ respectively such that any compact segment of $\alpha$ contains a finite number of each sequence. For each value of $i$ there will be a timelike geodesic segment $\gamma_i$ of maximum length between $a_i$ and $b_i$[5]. Each $\gamma_i$ will intersect the compact set F. Thus there will be a $q \in F$ which is a limit of $\gamma_i \cap F$ and a nonspacelike direction at q which is a limit of the directions of the $\gamma_i$. Let $\gamma$ be the inextendible geodesic through q with this direction. It will remain in $D^-(F) \cup D^+(G)$ both in the future and in the past. By condition (b), there will be conjugate points x and y of $\gamma$. As the positions of conjugate points on a geodesic are determined by an integral of the curvature along the geodesic, they can be chosen as a continuous function of the position of the intersection of the geodesic with F and of the direction at F. Thus if U and V are any neighbourhood of x and y respectively, there will be some $\gamma_i$ which intersects U and V and which contains conjugate points $x'$ and $y'$. But this is impossible[9] as $\gamma_i$ is supposed to be a geodesic segment of maximum length between $a_i$ and $b_i$. This establishes the desired contradiction which shows that the original assumption that space-time is timelike and null geodesically complete must be false.

Appendix

We wish to show that there is sufficient matter on our past light-cone to cause it to start converging again. A sufficient condition for this to be so is that there should be a distance R such that along every past directed null geodesic from us,

$$\frac{R}{3} \int_{R}^{\infty} \frac{8\pi G}{c^2} \; T_{ab} \, K^a \, K^b \, dr > 1 \qquad (I)$$

In this integral, $K^a = \frac{dx^a}{dr}$ is the tangent vector to the null geodesic and r is an affine parameter on the null geodesic normalised so that at is, r = 0 and $K^a U_a = \frac{1}{c}$ where $U^a$ is the past directed unit timelike vector representing the local standard of rest.

In a forthcoming paper[10], it is shown that, with certain assumptions, observations of the microwave background radiation indicate that not only do the past directed null geodesics from us start converging again but so also do the timelike ones, As we are concerned only with the null geodesics, the assumptions we shall need will be weaker.

The observations show that between the wavelengths of 20 cm and 2mm the background radiation is isotropic to within 1°/o and has the spectrum of a black body at 2.7°K. We shall assume that this black body spectrum indicates not that the radiation was necessarily created with this form, but that it has been thermalised by repeated scattering. Thus there much be sufficient matter on each past directed null geodesic from us to make the optical depth large in that direction. We shall show that this matter will be sufficient to cause the inequality I to be satisfied.

The smallest ratio of density to opacity at these wavelengths will be obtained if the matter consists of ionised hydrogen in which case there would be scattering by free electrons. The optical depth would be

$$\int_0^\infty \frac{\sigma}{m} \rho \, CK^a U_a \, dr \qquad \text{(II)}$$

where $\sigma$ is the Thomson scattering cross-section, m is the mass of a hydrogen atom, $\rho$ is the density of the ionised gas and $U^a$ is the local velocity of the gas. The red-shift Z of the gas is given by $(CK^aU^a - 1)$. We assume that this increases down out past-light cone. As galaxies are observed with red-shifts of 0.46 mostof the scattering must occur at red-shifts greater than this. (In fact if the quasars really are at cosmological distances, the scattering must occur at red-shifts of greater than 2). With a Hubble constant of 100 km./sec./ Megaparsec, a red-shift of 0.46 corresponds to a distance of about $3 \times 10^{27}$ cms. Taking R to be this distance, the contribution of the gas density to the integral in I is

$$1.8 \int_R^\infty \rho \, C^2(K^aU_a)^2 \, dr$$

while the optical depth of gas at red-shifts greater than 0.46 is

$$4 \int_R^\infty \rho \, C \, K^a \, U_a \, dr$$

As $C K^a U_a$ will be greater than 1.46 for r greater than R, it can be seen that the inequality I will be satisfied at an optical depth of about 1.6. If the optical depth of the Universe were less than this one would not expect a black body spectrum as the photons would not suffer sufficient collisions to thermalise them. Even if the radiation were to be created with a black body spectrum, what one would see would be a dilute 'grey' body spectrum which could agree with the observations between 20 cms. and 2 cms but which would not fit those at 9 mm and 2 mm. Thus we can be fairly certain that condition 3') is satisfied in the observed Universe.

References


1. Chandrasekhar, S. Mon Not. R.A.S. 95 207 (1935).
2. Penrose, R. Phys.Rev.Lett. 14 57 (1965).
3. Hawking, S.W. Proc.Roy.Soc. A, 295 490 (1966).
4. Brans, C. amd Dicke, R.H. Phys.Rev. 124 3 (1961).
5. Hawking, S.W. Proc. Roy.Soc. A. 300, 187 (1967).
6. Geroch, R.P. 'What is a singularity in general relativity' (Preprint).
7. Penrose, R. Lectures delivered at the 1967 Battelle Summer Rencontres. To be published by Benjamin Inc. New York.
8. Penrose, R. Adams Prize Essay 1966 University of Cambridge.
9. Hawking, S.W. Proc.Roy.Soc. A 294 511 (1966).
10. Hawking, S.W. and Ellis, G.F.R. Ap.J. to appear.
11. Hawking, S.W., Adams Prize Essay 1966 University of Cambridge.

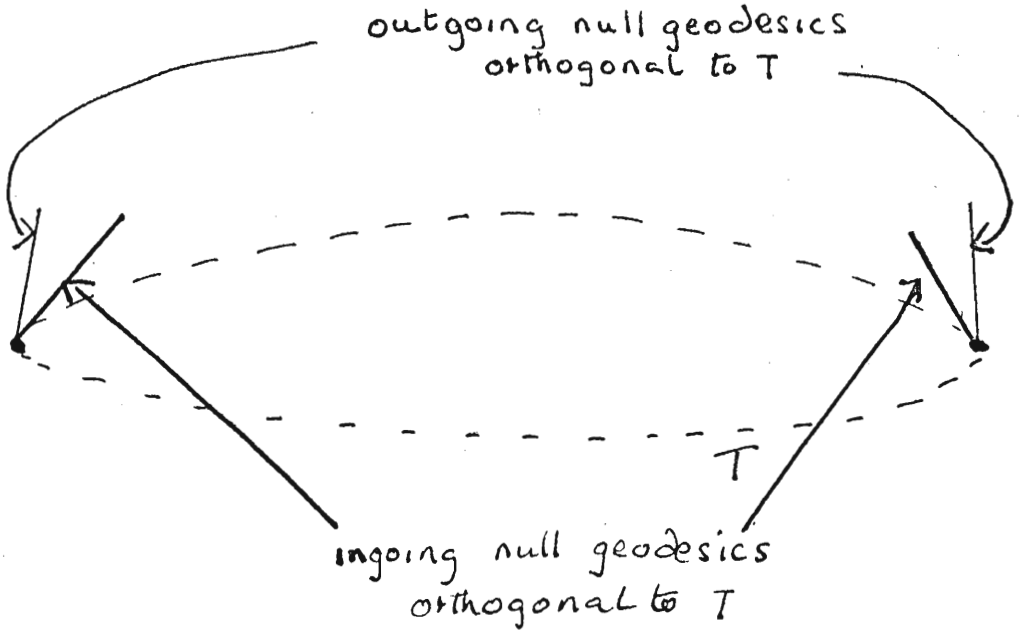



FIGURE I

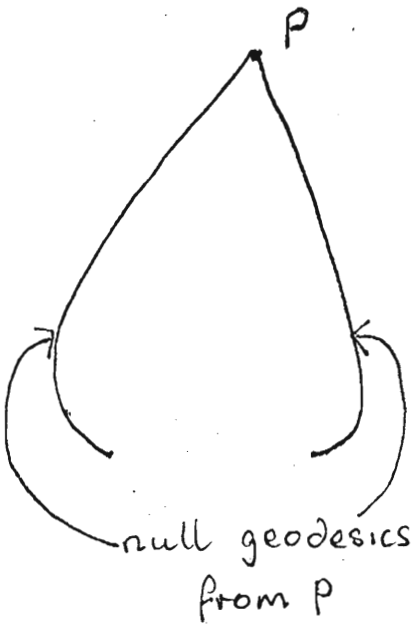
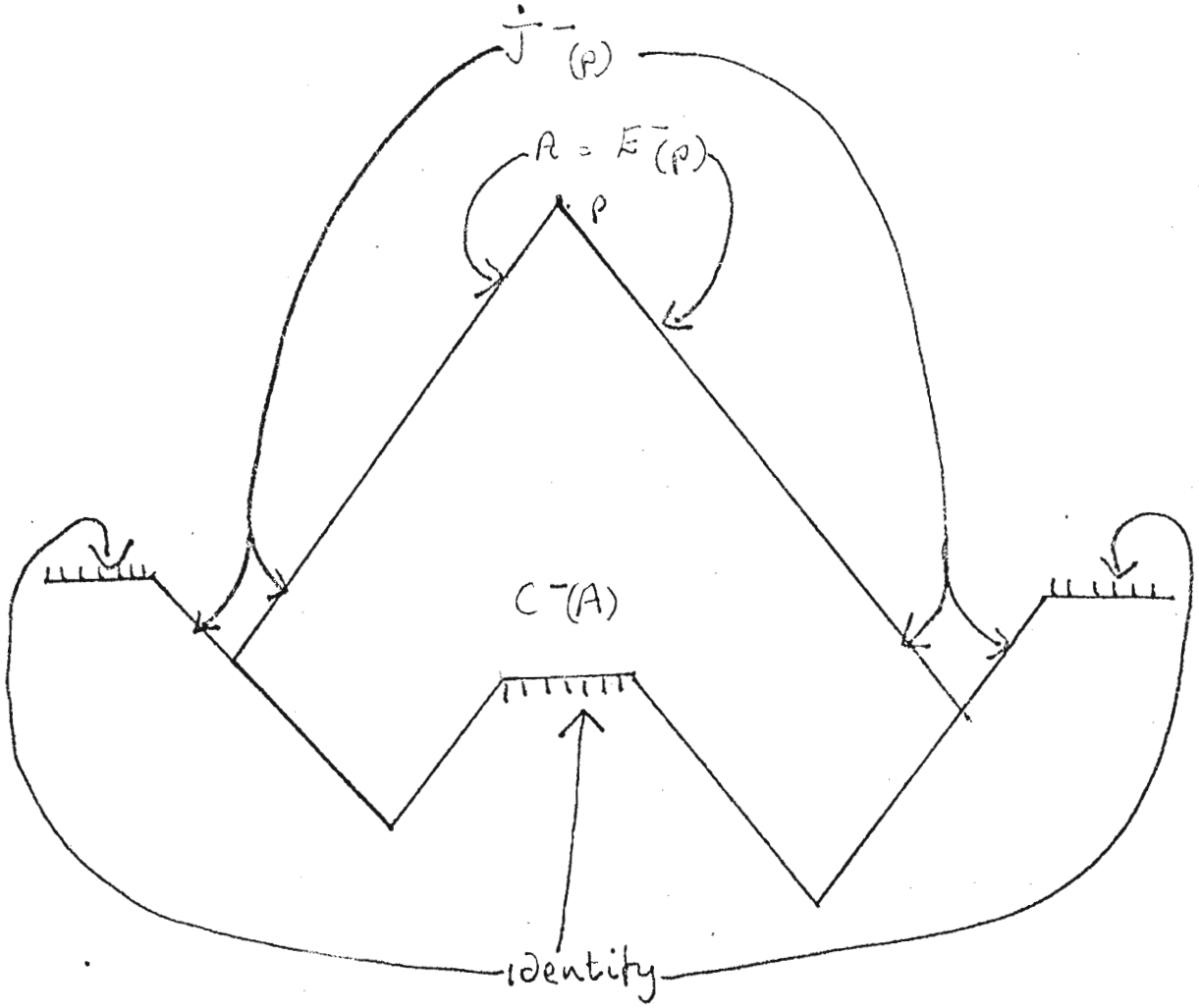



FIGURE 2.

case (ii)




case (iii)

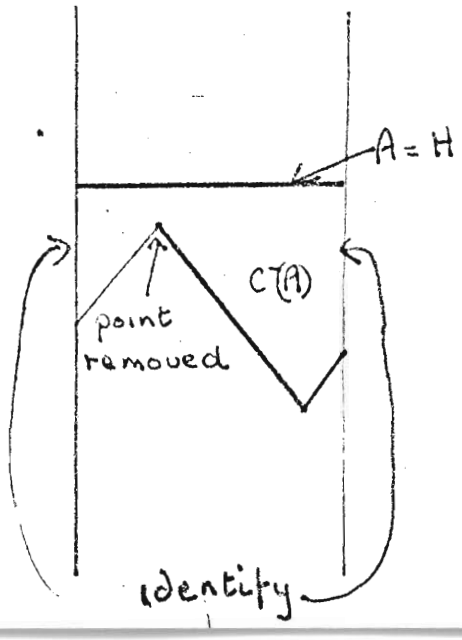



FIGURE 3.

I was born in 1942 and was educated at St. Albans School and University College, Oxford, where I took a B.A. in Physics. I then went to Cambridge to do research in Cosmology and General Relativity under the supervision of Dr. D. W. Sciama. I obtained my Ph.D. in 1966 and in the same year won an Adams Prize. Since 1965 I have been a Research Fellow of Gonville and Caius College, Cambridge.

I am married with one son.

Stephen Hawking

ROGER PENROSE

Born: 8th August 1931, Colchester, Essex, England.

Degrees: B.Sc. (Spec.) University College, London (1st class) 1952
Ph.D. Cambridge 1957

Employment, etc.:

Feb. 1956 - Aug. 1956: National Research Development Corporation, Tilney Street, London. (Mathematical work in connection with electronic computer programmes). Also consultant 1956-57.

1956-1957 Assistant Lecturer (Pure Mathematics) Bedford College, London.

1957-1960 Research Fellow, St. John's College, Cambridge.

1959-1961 N.A.T.O. Research Fellow, at Princeton, N. J., U.S.A. and also at Syracuse, N. Y. and Cornell, Ithaca, N.Y. (General Relativity and Quantum Theory)

1961-1963 Research Associate under U.S.A.F. contract, King's College, London (Mathematics Department)

1963-1964 Visiting Associate Professor (Mathematics and Physics), University of Texas, Austin 12, Texas, U.S.A.

1964-1966 Reader; 1966-Professor (Applied Mathematics), Birkbeck College, London.

During academic year 1966-1967, on leave in U.S.A. at Yeshiva University, New York; Princeton University; Cornell University; University of Chicago; Battelle Institute Seattle)