Astronomical and Astrophysical Transactions (*AApTr*), 2021, Vol. 32, Issue 3, pp. 153–162,
ISSN 1055-6796, Photocopying permitted by license only, ©Cambridge Scientific Publishers

Roger Penrose: the singularity theorem

A.N. Petrov*

Sternberg Astronomical Institute, M.V. Lomonosov Moscow State University, Universitetskii pr. 13, Moscow, 119992 Russia

Received 4 April, 2021

Our goal is to introduce the singularity theorem of Roger Penrose (Penrose 1965a) to a wide readership (not expert in gravitation physics). This paper is not a review, rather it is an essay given in a popular style. A short biography of Roger Penrose is presented and his scientific interests are outlined. More attention is paid to topological methods in general relativity, however there are no exact formulations and mathematical proofs; only the main notions are given and explained in order to be understood by an untrained reader. Finally Penrose's theorem on the unavoidability of singularity under collapse is derived and discussed.

Keywords: Nobel Prize, topological methods, collapse, singularity theorem

1 A short earlier biography and scientific interests

Roger was born 8 August 1931 in Colchester in Essex, UK, into an intellectual family. His mother, Margaret (Leathes), was a doctor and his father, Lionel Penrose, was a psychiatrist and a geneticist. His uncle and grandparents were artists and physiologists. His maternal grandmother, Sonia Marie Natanson, was a Jewish Russian who had left St. Petersburg in the late 1880s following Roger's grandfather. The older brother of Roger, Oliver Penrose, is a theoretical physicist; his sister Shirley Hodgson is a geneticist; his younger brother, Jonathan Penrose, is a chess grandmaster. For more detail, see Encyclopaedia Britannica (2021).

Roger Penrose began his higher education in London. He attended University College School and University College in London, where he graduated with bachelor's degree in mathematics. Later he went up to Cambridge University to complete a master's degree. He studied at St John's College under the supervision of the professor in geometry and astronomy Sir William Vallance Douglas Hodge. Already as a student Penrose began an active scientific study. For example, in 1955, he reintroduced Moore's generalised matrix known as the Moore–Penrose inverse (Penrose, 1955). His PhD thesis entitled "Tensor methods in algebraic geometry" was defended in 1958 under the supervision of algebraist and geometer professor John Arthur Todd.

*Email: alex.petrov55@gmail.com

In life, it is often the case that meeting with one or other person determines the future. Thus, the transition from pure mathematics to problems of astrophysics and cosmology was realized through his friendship with Dennis Sciama (Ellis&Penrose, 2010). He introduced Roger to the main problems of that time, the cosmological singularity, and as a very professional mathematician, Penrose through his study of singularities started the development of a separate and very effective mathematical formalism. It is so-called topological (or global) methods in general relativity; for a more detailed presentation, see the book (Hawking&Ellis, 1973). A quotation from Kip Thorne: "Roger Penrose revolutionised the tools that we use to analyse the properties of spacetime". We can tell that the development of this formalism has its origin in his famous work of 1965 entitled "Gravitational collapse and space–time singularities" (Penrose, 1965a). In fact, after 55 years the results of this paper (unavoidability of singularity under collapse) have been recognised by the Nobel Committee. Later together with Stephen Hawking this theorem has been represented on a more fundamental mathematical basis (Hawking&Penrose, 1969).

In addition to singularity theorems, Penrose was very successful in many other fields. For example, he invented twistors; he discovered the Penrose tilings; his invention of spin networks later came to formulate a loop quantum gravity. Penrose was also very successful in the popularization of science, and his books on popular science have engaged with a wide audience; he published "The Road to Reality: A Complete Guide to the Laws of the Universe", a comprehensive explanation of the laws of physics including his own theory; it is important to note that all Penrose's studies were quite fundamental.

During his long scientific career Roger Penrose has received many awards and prizes. Let us list some of them:

- 1971: the Dannie Heineman Prize for Astrophysics;
- 1972: the election for a Fellow of the Royal Society (FRS);
- 1975: the Eddington Medal of the Royal Astronomical Society (along with Stephen Hawking);
- 1985: the Royal Society Royal Medal;
- 1988: the prestigious Wolf Foundation Prize for Physics (along with Stephen Hawking);
- 1989: the Dirac Medal and Prize of the British Institute of Physics;
- 1990: the Albert Einstein Medal;
- 1991: the Naylor Prize of the London Mathematical Society;
- 1992–1995: the President of the International Society on General Relativity and Gravitation;
- 1994: was knighted for services to science;
- 1998: was elected Foreign Associate of the United States National Academy of Sciences;
- 2004: the De Morgan Medal;
- 2005: an honorary doctorate by Warsaw University and Katholieke Universiteit Leuven (Belgium);
- 2006: the Dirac Medal given by the University of New South Wales;
- 2008: the Copley Medal;

- 2011: the Fonseca Prize by the University of Santiago de Compostela;
- 2012: the Richard R. Ernst Medal;
- 2015: an honorary doctorate by CINVESTAV-IPN (Mexico);
- 2020: one half of the Nobel Prize in Physics for the discovery that black hole formation is a robust prediction of the general theory of relativity, a half-share also going to Reinhard Genzel and Andrea Ghez for the discovery of a supermassive compact object at the centre of our galaxy.

2 Problems of singularities

Singularity problems of the cosmological models are, first of all, problems of the singularities in the FLRW solutions. What is the content of these problems? Firstly, one has to note that metric coefficients lose their continuity; invariants of curvature, density and temperature of matter become infinite; secondly, spacetime in whole collapses to a point. In fact, at a singularity a state becomes undetermined and cannot be interpreted as physically meaningful. It looks paradoxical, but namely the physical meaninglessness of the situation stimulated Roger Penrose and his adherents for a deep research in this field in the 1960s.

Initially, it was thought that problems of singularities are far from real physics, but the development of the Big Bang model has shown that they are just the native problems of modern astrophysics and cosmology. Thus, the examination of the singularity problems becomes a quite actual task. Many works were devoted to this research, see references in book (Misner et al., 1973). In particular we focus on Misner's work entitled "The flatter regions of Newman, Unti, and Tamburino's generalized Schwarzschild space" (Misner, 1963). The main meaning of this work is that the author suggests to reformulate the notion of singularity itself. He proposes not to consider singularities as states with infinite physical and geometrical characteristics. Instead, he considers singularities as situations where the histories of matter particles and photons stop. In other words, they can be formulated as vanishing spacetime itself at the ends of world lines of real particles, no history behind these world points.

Let us consider the cosmological singularity in the context of Misner's idea. If we move back in time we meet the "break" of the history of the universe. Indeed, moving back in time, each world line of a real particle reaches a world point *behind* which there is no time and space. After a time-based conversation one can see that there is no spacetime *before* cosmological singularity (one has to bear in mind the singularity of, at least, the FLRW model).

Until the 1960s only the solutions of the Einstein equations had been searched for and examined. These equations are differential equations of the second order. Their solutions under these or other boundary conditions, symmetry restrictions, etc., are local functions. Then, to get a complete picture of the geometry one has to combine different local parts of the global geometry. As a result, unexpected problems can appear. Thus it is evident that for the study of global structures, global methods could be appropriate. Misner's idea of 1963 (Misner, 1963) and Penrose's work of 1965 (Penrose, 1965a) are, in fact, the origin of the development of new topological (global) methods in the framework of general relativity, see the book (Hawking&Ellis,

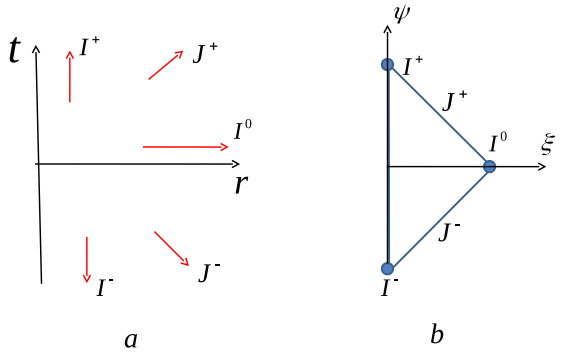


Figure 1 a) Minkowski space: I^- and I^+ denote time infinities of the past and future; I^0 denotes space infinity; J^- and J^+ denote null infinities of the past and future. b) The Penrose conformal diagram of the Minkowski space where time infinities are compressed into points I^- and I^+, space infinity is compressed into point I^0, null infinities are compressed into line segments J^- and J^+.

1973). This mathematical formalism is a very effective tool that is used in many branches of gravitation theory up to the present day.

Together with the already cited authors we note that very active participation in the development of global methods was also provided by Geroch, see for example (Geroch, 1968), Schmidt, see for example (Schmidt, 1971), and many other researches, see references in (Misner et. al., 1973; Hawking&Ellis, 1973). In the next sections, we give only the simplest elements of such studies in order to make the text comprehensible for readers who are not so close to the field. However, we hope that this will allow us to create an atmosphere of this science. We present some of theorems and lemmas, but do not prove them, although we give any geometrical and physical explanations. We discuss the singularity theorem at the end of the essay.

3 Elements of global methods

To describe objects in astronomy and astrophysics with reference to general relativity one has to model isolated systems. For such models we outline an asymptotic behaviour of metric coefficients and matter fields; moreover, we need mathematical operations with infinities. Then, it is necessary to define and classify various kinds of infinities. Of course, we have to consider *time* infinities of the past I^- and future I^+, and *space* infinities I^0 defined at spacelike cross sections. However, this is not enough. In order to describe various radiations, like gravitational waves or photons, we also have to consider *null* infinities of the past J^- and future J^+. A graphic illustration of the aforementioned infinities in the Minkowski space is given in Figure 1a.

Then, unfortunately, one has to operate with "infinities at infinities" that could be not so convenient. It would be good to consider compact geometrical objects. Penrose (Penrose, 1965b) suggested to use conformal transformations to "compress" infinities into finite objects, like line segments and points. In this case, operations

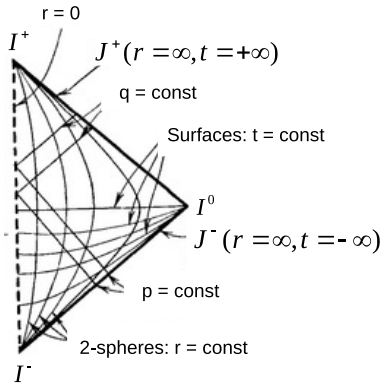


Figure 2 Conformal diagram of the Minkowski space with world lines of particles and spacelike cross sections.

with infinities are simplified significantly. Let us turn again to the Minkowski space and provide the transformations:

$$\begin{aligned} t + r &= \frac{1}{2}\tan(\psi + \xi), \\ t - r &= \frac{1}{2}\tan(\psi - \xi). \end{aligned}$$

Then the entire Minkowski spacetime plotted in Figure 1a is compressed into a triangle in Figure 1b with corresponding replacements of infinities by points and segments. In Figure 2 we rewrite the diagram in Figure 1b plotting the world lines of mass particles and photons, and spacelike sections. The lines connecting I^- and I^+ present world lines of freely falling mass particles with a constant r in the Minkowski space; the diagonal lines p symbolize ingoing photons, the diagonal lines q symbolize outgoing photons; the lines focusing at I^0 describe spacelike sections.

Analogously to the Minkowski space the Schwarzschild black hole geometry can be represented using the conformal diagrams, see Figure 3. It is not surprising that one can see the same notations for infinities. It is because asymptotically the Schwarzschild black hole goes to Minkowski space. However, in whole the structure is significantly more complicated including event horizons and singularities. Black hole solutions appear inevitably in numerous modified theories of gravity, very frequently these solutions are exotic with properties that are difficult to interpret. Conformal diagrams remain powerful tools for analyzing such models as well.

It is time to explain what is meant by the notions "horizons". Let us introduce some of the necessary definitions. Firstly, two events are causally connected if they can be connected by a timelike world line, a null geodesics or a combination of such world lines. Secondly, a spacetime could possess boundaries if two events (not connected by spacelike lines) cannot be classified as causally connected ones. Thirdly, hypersufaces confining part of spacetime, from under which the propagation of signals to the null infinity of the future J^+ is impossible, are named as horizons of the future ∂J^+ or event horizons. In Figure 4a the event horizon of the future is illustrated by a cylinder, whereas the spacelike section cuts the 2-dimensional horizon of a black hole.

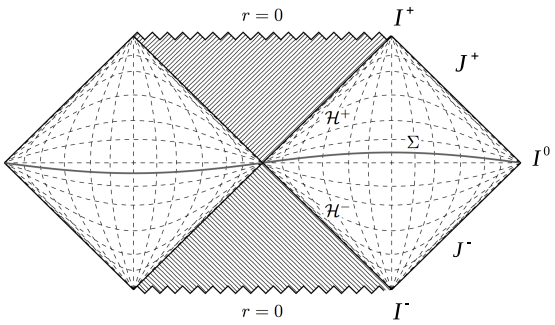


Figure 3 The Penrose conformal diagram for the Schwarzschild black hole. Infinities in the external universe coincide with infinities for the Minkowski space. In addition, there is the horizon of future null infinity (the event horizon) $\mathcal{H}^+$, and the horizon of past null infinity $\mathcal{H}^-$. Besides, there are the singularity of the past $r = 0$ and the singularity of the future $r = 0$.

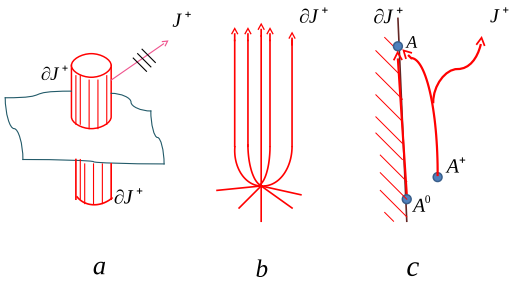


Figure 4 a) Event horizon is plotted by a cylinder, a spacelike section cuts the 2-dimensional horizon of a black hole; b) null generators of the horizon of the future with a caustic; c) A^0 is the past of A at ∂J^+; A^+ is the past of A in the area of the causal past of the null infinity of the future J^+.

Now we are in a position to turn to Penrose's theorem on the global structure of horizons (Misner et al., 1973).

THEOREM:

- *A horizon of the future is generated by null geodesics (at least, for finite values of affine parameter) lying at ∂J^+, at that time in the future there are no end points.*

We do not prove this theorem, however, we list some consequences which follow from the above theorem:

- If, following the null generators of ∂J^+, one goes to the past, one can (not necessarily) escape from ∂J^+. The event where this takes place is called the caustic of ∂J^+, see Figure 4b.
- If, proceeding to the future, a generator enters ∂J^+ through a caustic, then there is no possibility for it to escape from ∂J^+.

- A generator of ∂J^+ after entering through a caustic does not intersect with other null generators; thus, only one null generator goes through each of the events at ∂J^+.

Let us also discuss the important

LEMMA:

- *If a world line intersects ∂J^+ at any point A then, first, its past belongs either to ∂J^+ itself or to an area of the causal past of the null infinity of the future.*

This lemma is illustrated in Figure 4c. The main conclusion that follows from this lemma is that *no signals* can reach the horizon from the other side (with respect to the null infinity of the future).

Based on the aforementioned theorem and lemma in the local context of gravitational collapse Penrose proposed in 1969 the so-called *cosmic censorship conjecture* (Penrose, 1969). What does this mean? In reality, the null infinity is "asymptotically predictable". Then because singularities are "objects" with great physical uncertainties one has to confine singularities within a well-behaved event horizon surrounding a hidden space–time region. It is because of this that Wheeler coined the term "black hole", leaving a visible exterior region to be a causal past of the null infinity of the future J^+. This formulation is usually described as a "weak cosmic censorship hypothesis". Later, Penrose formulates a stronger version called the "strong censorship hypothesis". However, the censorship conjecture has the status of a hypothesis up to the present day. Combined with the Belinski–Khalatnikov–Lifshitz conjecture problems of nonlinear stability (Belinskii et al., 1971; Khalatnikov&Kamenshchik, 2008), the censorship conjecture is one of the most important outstanding problems in cosmology and general relativity till now.

4 The singularity theorem

The famous Penrose work (Penrose 1965a) which was recognised by the Nobel Committee in 2020, is very short, only two and a half pages. Unlike many papers on gravitational physics, the Penrose paper contains almost no equations and complicated formulae are absent. The main conclusions are based on the explanations of the unique diagram and we will discuss it at the end of the essay.

Firstly, collapse has been studied by Oppenheimer and Snyder in 1939 (Oppenheimer&Snyder, 1939). Penrose cites and discusses this work in which they examine a dust ball of constant density collapsing under gravitation. It turns out that such a ball irreversibly goes to its gravitational radius, after that it continues to collapse to a condition with infinitesimal volume and infinite density. In fact, this means a singularity. However, Oppenheimer and Snyder did not classify the situation so drastically, even in the sense of a hypothesis. Unlike them, Penrose classified the final collapse as a singularity in the sense of Misner's idea. In the words of Penrose, it is a "spacetime catastrophe".

Penrose restricted himself to the gravitational collapse of an isolated spherically symmetric body of a finite mass. It was not necessary to take into account any

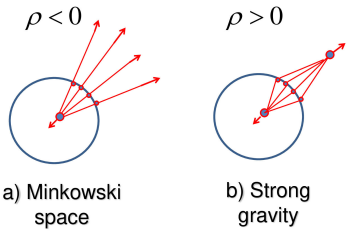


Figure 5 a) A two-sphere placed in a flat space has a negative convergence $\rho < 0$. b) A trapped surface is illustrated, for which, by definition, one has a positive convergence $\rho > 0$.

state equation, only the proof is based on the assumption that matter goes under the gravitational radius. Then, matter (world lines of particles and photons) irreversibly achieves end points of its history that is just a singularity with reference to Misner's idea. Penrose comments on this result as a necessity to develop new physics, quantum gravity, etc.

One of the main elements of the foundation of the singularity theorem is the notion of "trapped surfaces" introduced by Penrose (Penrose, 1965a). Now we explain it in more detail. They are two-dimensional surfaces. The main property of "trapped surfaces" is that null geodesics sent orthogonally to the surface are focusing. This property is related to both outgoing and ingoing rays. Mathematically this property is described by the characteristic ρ that is named as the "convergence" of a congruence of null geodesics, and for which $\rho > 0$. In Figure 5a we consider schematically a usual two-sphere in a flat space. In this case, the outgoing rays diverge because $\rho < 0$; the ingoing rays are focusing because it is the usual geometrical effect for a two-sphere embedded into a flat space. Figure 5b illustrates just a trapped surface with $\rho > 0$.

The natural question arises: where can one find trapped surfaces? The answer is simple – under the horizon of the Schwarzschild black hole. Each sphere $r, t = const$ for $r < 2m$ has the property $\rho > 0$. The boundary of a set of trapped surfaces is called the "apparent horizon", for the Schwarzschild black hole it is just the event horizon $r = 2m$ with $\rho = 0$ for the outgoing congruence of null geodesics, see Figure 6a. The correspondence between the trapped surfaces under the horizon and points under the event horizon at the conformal diagram is illustrated in Figure 6b.

The notion of the trapped surface becomes a key notion when one considers a dynamic black hole with the absence of static event horizons, see works (Ashtekar&Krishnan, 2004; Booth, 2005; Nielsen, 2010; Nielsen et al., 2011) and references therein. Therefore Penrose's theorem based on the condition of trapped horizons is more general in applications than when applied to simple static black holes.

Now, Penrose's singularity theorem of 1965 can be formulated as follows:

Spacetime manifold $\mathcal{M}^4$ with a necessity contains timelike or null geodesics, which have end points and cannot be continued (i.e., are stopped by a singularity in Misner's sense) if

- the Einstein equations hold;
- the conditions of a causality hold (there are no closed timelike curves);
- each of local observers measures a non-negative energy density;

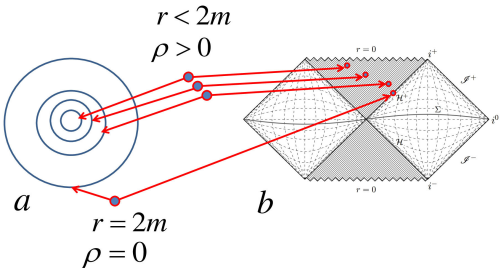


Figure 6 Trapped surfaces for the Schwarzschild geometry.

- $\mathcal{M}^4$ does not possess very high symmetries, such as anti-de Sitter space, de Sitter space, etc.;
- $\mathcal{M}^4$ contains a trapped surface, i.e., a surface for which $\rho > 0$.

Following Penrose we present his diagram in Figure 7. Let us give some explanation. The initial hypersurface for collapsing spherically symmetric matter with $r > 2m$ is denoted C^3. The state equation is not concretized. Only it is assumed that by the symmetrical collapse the matter goes under the sphere of gravitational radius $r = 2m$. An outside observer cannot fix this event. The diagram is presented with the ingoing Eddington–Finkelstein coordinates where the null coordinate V presents an advantage time, and the lines $V = \text{const}$ define ingoing rays. Two-surface S^2

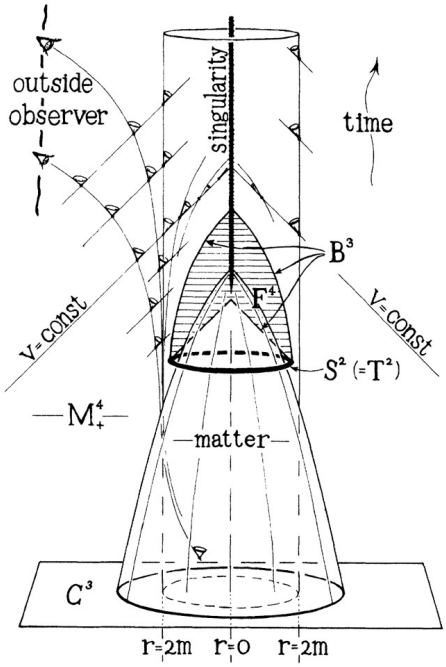


Figure 7 Diagram from the Penrose (1965a) paper.

denoted as T^2 is placed in an empty region under the gravitational radius. As we discussed above it just presents the trapped surface with all the described properties. Penrose noted that the trapped surface will still exist if the matter region has no sharp boundaries and with other more soft conditions.

All the points in F^4 are part of $\mathcal{M}^4$ and are future for the events at T^2. The hypersurface B^3 is a boundary of F^4 and is lightlike one where it is not singular. The lower cone of B^3 is a congruence of ingoing light rays, the upper surface B^3 is a congruence of outgoing light rays. Let A be an infinitesimal area of the cross sections of B^3, see horizontal lines in Figure 7. Deriving the equation for the propagation of the null geodesics from the boundary B^3, taking into account all the aforementioned conditions, one finds that A becomes zero for the finite affine parameter on each null geodesics calculated from T^2. Finally this means that the geodesics of real particles cannot be continued and, thus, inevitably reach the singularity.

Thus, the Penrose singularity theorem (Penrose, 1965) is proved.

Acknowledgments

The author was supported by the Interdisciplinary Scientific and Educational School of Moscow University "Fundamental and Applied Space Research".

References

Ashtekar A., Krishnan B. (2004) Living Rev. Rel. 7, 10
Belinskii V. A., Lifshitz E. M., Khalatnikov I. M. (1971) Sov. Phys. Usp. 13, 745
Booth I. (2005) Can. J. Phys. 83, 1073
Ellis G. F. R., Penrose R. (2010) Biograph. Mem. Fellows R. Soc. 56, 401
Encyclopaedia Britannica (2021) Roger Penrose – British mathematician, https://www.britannica.com/biography/Roger-Penrose, retr. March 28, 2021.
Geroch R. P. (1968) Ann. Phys. (USA) 48, 526
Hawking S. W., Ellis G. F. R. (1973) The Large Scale Structure of Space–Time, Cambridge: Cambridge Univ. Press
Hawking S. W., Penrose R. (1969) Proc. Roy. Soc. London A 314, 529
Khalatnikov I. M., Kamenshchik A. Yu. (2008) Phys. Usp. 51, 609
Misner C. W. (1963) J. Math. Phys. 4, 924
Misner C. W., Thorne K. S., Wheeler J. A. (1973) Gravitation, San Francisco: Freeman and Company
Nielsen A. B. (2010) Class. Quantum Grav. 27, 245016
Nielsen A. B., Jasiulek M., Krishnan B., Schnetter E. (2011) Phys. Rev. D 83, 124022
Oppenheimer J. R., Snyder H. (1939) Phys. Rev. 56, 455
Penrose R. (1955) Math. Proc. Cambridge Phil. Soc. 51, 406
Penrose R. (1965a) Phys. Rev. Lett. 14, 57
Penrose R. (1965b) Proc. Roy. Soc. London A 284, 159
Penrose R. (1969) (1969) Nuovo Cim. 1, 252
Schmidt B. G. (1971) Gen. Relat. Grav. 1, 269