Master Thesis

The *persistent* information paradox

Author: Nick Bultinck
Supervisor: Dr. Karel Van Acoleyen

A thesis submitted in fulfilment of the requirements for the degree of Master of science in de fysica en sterrenkunde

FACULTEIT WETENSCHAPPEN
Vakgroep Fysica en Sterrenkunde

June 2013
This thesis is about black holes and their turbulent but very fruitful relation with quantum mechanics. First it is explained why black holes can be seen as truly thermodynamical objects. This is done by looking at their classical properties and by doing quantum field theory around them. Just as ordinary thermodynamical objects black holes emit thermal radiation. Conservation of energy will cause them to shrink and to eventually evaporate completely. Although calculations in the semiclassical approach seem to suggest that the evaporation of a black hole evolves pure states into mixed states, we argue why and how this process should nevertheless be made unitary. This will lead us to the long-lived and well-established phenomenological description of unitary black holes called black hole complementarity. However, it is shown that a closer look at the postulates of black hole complementarity reveals a paradox. In particular, there is a conflict between unitarity, the equivalence principle and effective field theory. It seems that if one is reluctant to give up unitarity, a destructive firewall should be placed at the horizon. Jumping into a black hole may be much more dangerous than expected.

Key words General relativity, black holes, quantum field theory, information paradox, black hole complementarity, quantum gravity, firewall
Acknowledgements

I would like to thank my supervisor Dr. Karel Van Acoleyen for the freedom he gave me and for the support I could always ask. I felt great confidence from his side. Also, I am very grateful that I have been granted the opportunity to study this fascinating subject. It was fun working on this thesis.

En dan uiteraard

Mijn moeder en mijn vader,
Julie,

bedankt voor alles. Jullie zijn een grote steun voor mij.
Contents

Abstract ii
Acknowledgements iv
Preliminary x

1 Classical aspects of black holes 1
  1.1 Spacetime symmetries ........................................ 2
  1.2 Null surfaces .................................................. 5
  1.3 The Cauchy problem in General Relativity .................... 6
    1.3.1 Differential equations in curved spacetime ............. 6
    1.3.2 Global Hyperbolicity .................................. 7
  1.4 The Horizon ................................................... 8
    1.4.1 Gravitational collapse ................................ 8
    1.4.2 Outside observers and the end of time ................. 12
    1.4.3 Infalling observers and the equivalence principle .... 14
  1.5 The maximal extension of Schwarzschild spacetime .......... 16
  1.6 Killing horizons and surface gravity ........................ 19
  1.7 Penrose diagrams ............................................. 23
    1.7.1 Conformal compactification .............................. 23
    1.7.2 Gravitational collapse spacetime ....................... 28
  1.8 No hair conjecture ............................................ 30
    1.8.1 Kerr-Newman geometry and uniqueness theorems ........ 30
    1.8.2 Classical fields ....................................... 34
    1.8.3 The electric Meissner effect and equilibrium .......... 36
  1.9 Radial null geodesics in black hole spacetimes ............ 36
    1.9.1 The Schwarzschild black hole .......................... 37
    1.9.2 The Kerr black hole .................................. 39
  1.10 Energy extraction in Kerr spacetime ......................... 44
    1.10.1 The ergosphere of a Kerr black hole ................. 44
    1.10.2 The Penrose process .................................. 45
    1.10.3 Superradiance ....................................... 47
  1.11 Black hole mechanics ........................................ 49
    1.11.1 The area theorem .................................... 50
    1.11.2 Zeroth law ........................................... 52
    1.11.3 First law ............................................ 53
## 2 Quantum field theory in curved spacetime

2.1 The formulation of QFT in curved spacetime .......................... 58
  2.1.1 The canonical approach ............................................ 58
  2.1.2 A cosmological model .............................................. 63
    2.1.2.1 The set-up .................................................. 63
    2.1.2.2 Particle creation ......................................... 67
    2.1.2.3 Conclusions ............................................... 68
  2.1.3 The loss of Poincaré symmetry .................................. 69
    2.1.3.1 The particle content of the Klein-Gordon field ............ 70
    2.1.3.2 The lack of spacetime symmetries ......................... 70
    2.1.3.3 The algebraic approach .................................. 73
  2.2 The Unruh effect ....................................................... 73
    2.2.1 Rindler spacetime .............................................. 74
    2.2.2 Accelerating observers and the thermal bath .................. 77
  2.3 Particle creation by black holes .................................. 82
    2.3.1 Original derivation of the Hawking radiation ................. 83
      2.3.1.1 The Schwarzschild black hole ............................ 86
      2.3.1.2 The Kerr black hole .................................... 91
      2.3.1.3 Final remarks ......................................... 92
    2.3.2 Alternative views on the Hawking radiation .................. 94
      2.3.2.1 Static observers and the Unruh effect .................. 94
      2.3.2.2 Heuristic arguments .................................. 95
    2.3.3 Trans-Planckian physics in Hawking radiation ............... 96
  2.4 Angular momentum and gray body factors ........................... 97
  2.5 The generalized second law ....................................... 100
    2.5.1 The lowering of matter in a static black hole ............. 101
    2.5.2 A more general argument .................................... 105
  2.6 Euclidean path integral methods .................................. 107
    2.6.1 Hawking temperature derivation ............................... 108
    2.6.2 Black hole entropy derivation ............................... 109

## 3 The membrane paradigm

3.1 The stretched horizon ............................................... 115
3.2 A conducting surface ............................................... 118
3.3 Spreading of a charge ............................................... 121

## 4 Entanglement and information

4.1 Density matrices and entanglement ................................ 125
  4.1.1 Ensembles ..................................................... 126
  4.1.2 Quantum statistical mechanics ................................. 127
  4.1.3 Reduced density matrix ........................................ 130
  4.2 Unruh density matrix ............................................. 132
  4.3 Generalized second law for quasistationary semiclassical black holes ........................................ 134
  4.4 The information paradox ......................................... 137
    4.4.1 A toy model ................................................. 137
      4.4.1.1 Nice slices ............................................. 138
      4.4.1.2 Particle creation revisited ............................ 139
## Contents

4.4.1.3 Slicing the black hole geometry .................................... 140  
4.4.1.4 From pure to mixed .................................................. 143  
4.4.1.5 Mixed states and information ..................................... 146  
4.4.2 The true physical situation ........................................... 147  
4.5 Implications of non-unitary evolution ................................. 148  
4.5.1 The superscattering operator ....................................... 148  
4.5.2 A general evolution equation ...................................... 149  
4.5.3 A subclass of solutions ............................................... 152  
4.6 Possible ways to unitarity .............................................. 154  
4.6.1 Information in the Hawking radiation ............................. 154  
4.6.1.1 Backreaction and small corrections ......................... 154  
4.6.1.2 Quantum hair and fuzzballs .................................. 159  
4.6.2 Stable remnants ....................................................... 159  
4.6.3 Information release at the end ..................................... 161  
4.6.4 Baby universes ........................................................ 163  
4.6.5 Other modifications of conventional theories ................... 165  
4.7 AdS/CFT and the information paradox ............................... 166  
4.8 Euclidean gravity and unitarity ....................................... 167  
4.9 Thermodynamics of horizons .......................................... 170  
4.10 Horizon entanglement entropy ....................................... 179  
4.10.1 Entanglement entropy in flat spacetime ....................... 180  
4.10.2 Entanglement entropy of Killing horizons .................... 185  

5 Black hole complementarity .................................................. 189  
5.1 A brick wall ............................................................... 189  
5.2 Problems with information in the Hawking radiation ........... 193  
5.3 Average information in the Hawking radiation ................. 195  
5.4 The postulates ............................................................ 197  
5.5 Thought experiments ..................................................... 201  
5.5.1 Verification of the stretched horizon ......................... 201  
5.5.2 Baryon number violation ......................................... 202  
5.5.3 Entangled spins ....................................................... 204  
5.6 Old black holes as quantum mirrors ................................ 207  
5.7 Fast scrambling .......................................................... 212  
5.7.1 Scrambling in general quantum systems ..................... 212  
5.7.2 Scrambling in black holes ....................................... 215  
5.7.3 The entangled spin experiment revisited .................... 216  
5.8 Complementarity in the semiclassical framework ............... 217  

6 The Firewall ................................................................. 223  
6.1 The AMPS argument ..................................................... 223  
6.1.1 The entropy argument .............................................. 224  
6.1.2 The projection argument ......................................... 225  
6.2 The thermal zone and mining ....................................... 228  
6.3 Why complementarity is not enough ............................... 229  
6.4 Migrating singularity ................................................... 230  
6.5 Formation time of the firewall ..................................... 233
6.5.1 Generic and scrambled ............................................. 233
6.5.2 Fine grained and coarse grained ................................. 234
6.5.3 Special states and generic states ................................. 235
6.6 Non-local dynamics ..................................................... 238
6.7 The Harlow-Hayden conjecture ....................................... 240
6.7.1 The decoding process .............................................. 241
6.7.2 General unitary transformation with quantum gates ........... 245
6.7.3 Decoding is slower than black hole dynamics ................. 247
6.8 Next generation complementarity .................................... 249
6.8.1 The stretched horizon as a hologram ......................... 249
6.8.2 The transfer of (distillable) entanglement .................... 251
6.8.3 Standard complementarity (A = R_B ⊂ R) ..................... 254
6.8.4 Strong complementarity ........................................... 255
6.9 Problems with A ⊂ R .................................................. 257
6.9.1 Measurements create a firewall .................................. 257
6.9.2 State dependence .................................................. 259
6.9.3 Arbitrariness and energy considerations ...................... 260
6.10 The Hilbert space for an infalling observer ..................... 262
6.11 Static AdS black holes .............................................. 264
6.12 Conclusion ................................................................... 265
6.13 * Personal view ......................................................... 267
6.13.1 A firewall? .......................................................... 268
6.13.2 Backreaction ......................................................... 269
6.13.3 Freely falling vs. hovering ....................................... 270
6.13.4 Global vs. local .................................................... 273

A Frobenius’s theorem ....................................................... 277

B Surface gravity of a Kerr black hole .................................. 280

C The zeroth law ............................................................. 283

D The Hamiltonian formulation of general relativity ................ 287

E Scrambled entanglement ................................................... 292
   E.1 Ordered ground states .............................................. 292
   E.2 Scrambled systems .................................................. 293

Bibliography .................................................................. 297
Preliminary

General relativity. Quantum field theory. Two well established, beautiful and very successful theories. Physicists have been trying to unify these theories since a very long time now. However, this appears to be a very difficult task because at the present time there is still no satisfactory quantum theory of gravity. This thesis deals with a more modest task: reconciling black holes with unitarity. But it nevertheless is the hope that a successful unitary picture of these gravitational objects can provide us with some clues of how to think about a unified theory of general relativity and quantum mechanics.

This thesis can be seen as being composed out of two parts. The first part consists of chapters 1, 2 and 3. The main purpose of chapter 1, which deals with some classical aspects of black holes, is to give the necessary background for chapter 2. In this second chapter quantum field theory in curved spacetime is examined in order to explain particle creation by black holes. In chapter 3 a very useful and intuitive mental picture of black holes will be presented which goes under the name of the membrane paradigm. This first part is the 'positive' part. It shows how general relativity and quantum mechanics agree perfectly about the thermodynamic nature of black holes. Both theories merge beautifully in this area.

The second part consists of chapters 4, 5 and 6. These deal with the main purpose of this thesis, which is the information paradox and its ramifications. Actually, the first part could be skipped and one could simply start reading at chapter 4, just assuming some basic results of the first three chapters. However, it is my opinion that the first part provides a valuable context for the information paradox which helps to understand the depth of the matter being presented.

The second part is the 'negative' part. While quantum mechanics and general relativity matched perfectly in the first part, their combination will give rise to paradoxes and conflicts of principle in the second. This will force us to leave the conventional tracks and to go onto unknown slippery roads. While the first part was about finding the correct outcomes and interpretations of known formulas, the second part deals with the search for new formulas. Or not even that, it concentrates on the search for new principles that could possibly lead to new formulas. In this search thought experiments will prove themselves to be of great use.

Although there are many promising theories to explain quantum gravity like string theory
and loop quantum gravity, I will try to restrict to what we can learn from the experimentally
certified theories general relativity and quantum field theory. This means I will also not ex-
plicitly consider the fuzzball models, i.e. black hole models based on string theory. In some
cases I will refer to the AdS/CFT correspondence, but only as to provide additional information
to broaden the picture.

My aim was to make this thesis self-containing. If I have succeeded, then a graduate student
like myself, who has had some introductory courses on general relativity, quantum mechanics
and quantum field theory, should be able to follow everything without having to consult any of
the references.

This thesis contains 50 years of physics, from the golden age of black hole physics in the 1960’s
to the present day, with ingredients from a variety of fields.

I hope the reader may enjoy it as much as I did.

Nick Bultinck
June 2013

PS: I would also like to apologize for the typos that escaped my eye and which are undoubtedly
present in a work of this length.
Chapter 1

Classical aspects of black holes

*A luminous star, of the same density of the earth, and whose diameter should be two hundred and fifty times larger than that of the sun, would not, in consequence of its attraction, allow any of its rays to arrive at us. It is therefore possible that the largest luminous bodies in the universe may, through this cause, be invisible.*

- P.S. Laplace (1798)

Imagine a big, let’s say infinite, shallow lake. In this lake live some fish. Because of their biology these fish cannot swim faster than the speed of sound in the water. Now suppose that somewhere in this lake, there is a hole in the bottom, creating a drainhole through which water is flowing away. The water that is sucked away flows along some very sharp rocks, so everything that gets sucked in the drainhole encounters a very violent end of its existence. This situation is depicted in figure 1.1a.

![Image](figure1.png)

**Figure 1.1:** Some fish in a lake.

As in every normal drainhole, the water flows faster and faster when approaching towards the center. At some point, the water starts flowing faster than the speed of sound. The line where this happens is depicted with a circle in figure 1.1b. Now suppose that a fish, called Alice, passes through this line. The moment when she passes this line, nothing unusual happens to Alice, she feels no other than before. But there is definitely a great consequence of this passage, because for an outside fish, say Bob, Alice has ceased to exist. Alice cannot swim back to inform Bob that she is fine because she cannot swim faster than the speed of sound, she is past the point of no return. It is also impossible to let Bob hear from her existence because the soundwaves are
Chapter 1. Classical aspects of black holes

also trapped behind this point of no return. So for Bob, living on the outside, Alice is gone.
Of course, what is presented here is an analogy to a black hole (Remarkably, this analogy has
a direct and practical application to the mechanism of Hawking radiation, see section 2.3.3).
Black holes and their combination with quantum mechanics is the context in which the subject
of this thesis is situated. In this first chapter the necessary classical aspects of black holes are
presented.
Black holes are objects pushing human imagination to its utter limits. But nevertheless, they
may be quite less exotic as they would appear at first sight. Calculations show that a star of a
mass that is one and a half times the mass of our sun has a fair chance of collapsing to a black
hole at the end of its life [1]. And there are at least $10^9$ more massive stars in our galaxy alone...

1.1 Spacetime symmetries

Consider a timelike curve $C$ with endpoints $A$ and $B$. The action for a free particle of mass $m$
moving on $C$ is

$$S = -mc^2 \int_A^B d\tau = -mc \int_A^B ds,$$

(1.1)

where $\tau$ is the proper time on $C$. Since

$$ds = \sqrt{dx^\mu dx^\nu g_{\mu\nu}} = \sqrt{\dot{x}^\mu \dot{x}^\nu g_{\mu\nu}} d\lambda,$$

(1.2)

where $\lambda$ is an arbitrary affine parameter on $C$ and $\dot{x}^\mu = \frac{dx^\mu}{d\lambda}$, one has

$$S[x(\lambda)] = -mc \int_a^b d\lambda \sqrt{\dot{x}^\mu \dot{x}^\nu g_{\mu\nu}} \quad (c = 1).$$

(1.3)

The world lines of free particles or geodesics are found by demanding that

$$\frac{\delta S[x]}{\delta x(\lambda)} = 0.$$

(1.4)

This implies

$$0 = \delta S = \int_a^b d\lambda \left( \frac{\partial L}{\partial x^\sigma} - \frac{d}{d\lambda} \left( \frac{\partial L}{\partial \dot{x}^\sigma} \right) \right) \delta x^\sigma.$$  

(1.5)

Because this result should be independent of the variation one gets the Euler-Lagrange equations

$$\frac{d}{d\lambda} \left( \frac{\partial L}{\partial \dot{x}^\sigma} \right) - \frac{\partial L}{\partial x^\sigma} = 0.$$

(1.6)

Which become explicitly

$$0 = \ddot{x}^\kappa + \frac{1}{2} g^{\rho\kappa}(\partial_{\sigma} g_{\mu\rho} + \partial_{\mu} g_{\sigma\rho} - \partial_{\rho} g_{\mu\sigma}) \dot{x}^\mu \dot{x}^\sigma,$$

(1.7)

$$= \ddot{x}^\kappa + \Gamma^\kappa_{\mu\sigma} \dot{x}^\mu \dot{x}^\sigma$$

(1.8)

And these are the geodesic equations. A solution $x^\mu(\lambda)$ to these equations is called a geodesic.
Chapter 1. Classical aspects of black holes

The tangent to a general curve \( C \) with equation \( x^\mu(\lambda) \) is given by \( t^\mu = \dot{x}^\mu(\lambda) \). For a general vector field \( V^\nu(x(\lambda)) \) along \( C \) one has

\[
\begin{align*}
   t^\nu \nabla_\nu V^\mu &\equiv t^\nu \partial_\nu V^\mu + t^\nu \Gamma^\mu_{\nu\rho} V^\rho \\
   &= \frac{d}{d\lambda} V^\mu + \dot{x}^\nu \Gamma^\mu_{\nu\rho} V^\rho .
\end{align*}
\] (1.9)

Where \( \nabla_\mu \) is the covariant derivative. Since \( t \) is a tangent to the curve, a vector field \( V \) on \( C \) for which

\[
t^\nu \nabla_\nu V^\mu = f(\lambda) V^\mu
\] (1.10)

with an arbitrary function \( f \) is said to be parallely transported along the curve. From (1.8), (1.9) and by taking \( f(\lambda) \equiv 0 \) it is then clear that a geodesic is curve whose tangent is parallely transported along it.

Now consider the infinitesimal transformation

\[
x^\mu \rightarrow x'^\mu = x^\mu - \epsilon k^\mu(x),
\] (1.11)

changing the action to

\[
S'[x(\lambda)] = -mc \int_a^b d\lambda \sqrt{\dot{x}^\mu \dot{x}^\nu g_{\mu\nu}} .
\] (1.12)

Under this transformation the metric transforms like

\[
g_{\mu\nu}(x) \rightarrow g'_{\mu\nu}(x') = \frac{\partial x^\lambda}{\partial x'^\mu} \frac{\partial x^\sigma}{\partial x'^\nu} g_{\lambda\sigma}(x) .
\] (1.13)

From this it follows that

\[
g'_{\mu\nu}(x) = g_{\mu\nu}(x) + \epsilon (k^\rho \partial_\rho g_{\mu\nu} + \partial_\mu k^\rho g_{\nu\rho} + \partial_\nu k^\rho g_{\mu\rho})
\] (1.14)

\[
= g_{\mu\nu}(x) + \epsilon \mathcal{L}_k g_{\mu\nu}(x) ,
\] (1.15)

where

\[
\mathcal{L}_k g_{\mu\nu}(x) = \nabla_\mu k_\nu + \nabla_\nu k_\mu
\] (1.16)

is the Lie derivative of the metric.

The function \( x(\lambda) \) comes down to considering a line in spacetime within a certain coordinate patch. Relabeling the spacetime points by a coordinate transformation is independent of which specific line in spacetime is considered, so \( k^\mu(x) \) is independent of \( \lambda \). With all this, the transformation of the action (1.12) becomes

\[
S'[x(\lambda)] = S[x(\lambda)] - m\epsilon \int_a^b d\lambda (-\dot{x}^\mu \dot{x}^\nu g_{\mu\nu})^{-1/2} (\mathcal{L}_k g_{\mu\nu}) .
\] (1.17)

Thus the action is invariant if

\[
\mathcal{L}_k g_{\mu\nu} = 0 .
\] (1.18)
A vector field $k^\mu(x)$ satisfying this property is called a Killing vector field. It immediately follows from (1.16) that a Killing vector field satisfies

$$ \nabla_\mu k_\nu = -\nabla_\nu k_\mu, $$

which is the Killing vector lemma.

Interpreting vector fields as derivation operators on spacetime functions, a Killing vector field can be seen as the generator of a symmetry of the spacetime. To define what a spacetime symmetry is, first consider two (pseudo-) Riemannian manifolds $(M, g)$ and $(M', g')$. A diffeomorphism $f$ between $M$ and $M'$ is an isometry if $f_*g = g'$. If $M = M'$ than an isometry is called a symmetry of the spacetime. Hence, a Killing vector field is a vector field whose flow (= integral curves) consists of isometries.

Since for each Killing vector field there is a symmetry of the action, there also is a conserved charge. This charge is

$$ Q = k^\mu p_\mu $$

where $p_\mu$ is the particle’s 4-momentum

$$ p_\mu = \frac{\partial L}{\partial \dot{x}^\mu}, $$

$$ = m \frac{dx^\mu}{d\tau} g_{\mu\nu} \quad \text{when } m \neq 0. $$

To see how this charge is conserved, use the fact that the action is invariant under the above coordinate transformations

$$ 0 = \delta S = \int d\lambda \left( \frac{\partial L}{\partial x^\mu} \delta x^\mu + \frac{\partial L}{\partial \dot{x}^\mu} \delta \dot{x}^\mu \right) $$

$$ = \int d\lambda \left( \frac{d}{d\lambda} \left( \frac{\partial L}{\partial \dot{x}^\mu} \right) \delta x^\mu + \frac{\partial L}{\partial \dot{x}^\mu} \delta \dot{x}^\mu \right) $$

$$ = \int d\lambda \frac{d}{d\lambda} \left( \frac{\partial L}{\partial \dot{x}^\mu} \delta x^\mu \right), $$

where the Euler-Lagrange equations (1.6) were used in the second step. From this follows

$$ \frac{\partial L}{\partial \dot{x}^\mu} \delta x^\mu = \varepsilon \frac{\partial L}{\partial \dot{x}^\mu} k^\mu = \text{constant}. $$

As mentioned before, a vector field is a derivation operator generally expressed as

$$ k = k^\mu \partial_\mu. $$

For any vector field $k$, local coordinates can be found such that

$$ k = \frac{\partial}{\partial \xi}. $$
where $\xi$ is one of the coordinates. In such a coordinate system one has

$$L_{k}g_{\mu\nu} = \frac{\partial}{\partial \xi}g_{\mu\nu}. \quad (1.28)$$

So $k$ is a Killing vector field if $g_{\mu\nu}$ is independent of $\xi$. For example, in a static spacetime $\partial_{t}g_{\mu\nu} = 0$, so $\partial/\partial t$ is a Killing vector field. The corresponding conserved quantity is

$$mg_{00} \frac{dt}{d\tau} = m e, \quad (1.29)$$

where $e$ is the energy per unit mass.

### 1.2 Null surfaces

Let $S(x)$ be a smooth function of the spacetime coordinates $x^\mu$ and consider a family of hypersurfaces $S = \text{constant}$. The vector fields normal to the hypersurfaces are

$$l = - f(x)(g^{\mu\nu}\partial_\nu S) \frac{\partial}{\partial x^\mu}, \quad (1.30)$$

where $f(x)$ is an arbitrary non-zero spacetime function. If $l$ is a null vector, i.e. $l^2 = 0$, for a particular hypersurface $N$ in the family, then $N$ is said to be a null hypersurface.

A vector $t$ is tangent to a hypersurface if $t \cdot l = 0$. But for the null hypersurface $N$ it holds that $l \cdot l = 0$, so $l$ is itself a tangent vector:

$$l^\mu = \frac{dx^\mu}{d\lambda}, \quad (1.31)$$

for some null curve $x^\mu(\lambda)$ in $N$.

These curves appear to have a very special property. This can be seen by considering the vector $l \cdot \nabla l^\mu$. By using the definition for the normal $l^\mu$ and the fact that the covariant derivative of the metric is zero, one gets

$$l \cdot \nabla l^\mu = -l \cdot \nabla (f(x)g^{\mu\nu}\partial_\nu S) = - (l^\rho \partial_\rho f) g^{\mu\nu}\partial_\nu S - f g^{\mu\nu}l^\rho \nabla_\rho \partial_\nu S, \quad (1.32)$$

which can be rewritten by using the definition of the normal $l^\mu$ and the symmetry of the Levi-Civita connection

$$l \cdot \nabla l^\mu = (l \cdot \partial \ln f) l^\mu - f g^{\mu\nu}l^\rho \nabla_\nu \partial_\rho S. \quad (1.33)$$

Again making use of the definition of $l^\mu$ in the second term, this becomes

$$l \cdot \nabla l^\mu = (l \cdot \partial \ln f) l^\mu + l^\rho f \nabla^\mu (f^{-1}l_\rho) = (l \cdot \partial \ln f) l^\mu + l^\rho \nabla^\mu l_\rho - (\partial^\mu \ln f) l^2 = (l \cdot \partial \ln f) l^\mu + \frac{1}{2} \partial^\mu l^2 - (\partial^\mu \ln f) l^2. \quad (1.34)$$
Where it was used in the last step that \( l^2 \) is a scalar so that the covariant derivative can be replaced by a partial derivative.

When evaluating (1.34) on the null surface \( \mathcal{N} \), (1.31) and \( l^2 = 0 \) give

\[
\left. l \cdot \nabla l^\mu \right|_\mathcal{N} = \left( \frac{d}{d\lambda} \ln f \right) l^\mu |_\mathcal{N} + \frac{1}{2} \partial^\mu l^2 |_\mathcal{N}.
\] (1.35)

It does not follow that \( \partial^\mu l^2 \) is zero on \( \mathcal{N} \), unless the whole family of hypersurfaces \( S = \text{constant} \) is null. However, since \( l^2 \) is constant on \( \mathcal{N} \), it holds that \( t^\mu \partial_\mu l^2 = 0 \) for any tangent vector \( t \) of \( \mathcal{N} \). Thus it follows that

\[
\partial_\mu l^2 |_\mathcal{N} \propto l_\mu,
\] (1.36)

and therefore one gets from (1.35)

\[
\left. l \cdot \nabla l^\mu \right|_\mathcal{N} \propto l^\mu.
\] (1.37)

From (1.10) it is clear that this expresses parallel transport of the tangent vector \( l \). So \( x^\mu(\lambda) \) appears to be a geodesic. This can be made explicit by choosing the function \( f \) such that \( l \cdot \nabla l^\mu = 0 \) on \( \mathcal{N} \). Using (1.31) one gets

\[
0 = \left. l \cdot \nabla l^\mu \right|_\mathcal{N} = \dot{x}^\nu \nabla_\nu \dot{x}^\mu = \ddot{x}^\mu + \Gamma^\mu_{\nu\sigma} \dot{x}^\nu \dot{x}^\sigma,
\] (1.38)

which is the geodesic equation (1.8).

The null geodesics \( x^\mu(\lambda) \) for which the tangent vectors \( \dot{x}^\mu(\lambda) \) are normal to a null hypersurface \( \mathcal{N} \), are called the generators of \( \mathcal{N} \).

### 1.3 The Cauchy problem in General Relativity

In this thesis frequent use will be made of differential equations in a general spacetime. Here, a closer look is taken at the existence and uniqueness of solutions to such differential equations.

#### 1.3.1 Differential equations in curved spacetime

Consider the curved spacetime version of the Klein-Gordon equation

\[
(g^{\mu\nu} \partial_\mu \partial_\nu - m^2) \psi = 0,
\] (1.39)

where the Minkowski metric is replaced by a general spacetime metric according to the principle of minimal coupling, to be discussed in the next chapter. The Klein-Gordon equation will serve in this section as a representative of the entire class of second order hyperbolic equations.

One would like to obtain that, analogous to Minkowski spacetime, solutions of this Klein-Gordon are uniquely determined by some initial conditions on a spacelike hypersurface. In
other words, embedding this 3-dimensional hypersurface containing the initial data in the 4-dimensional spacetime \((M, g_{\mu\nu})\) should, together with the Klein-Gordon equation, result in a unique spacetime function \(\psi\) on \(M\).

In an arbitrary curved spacetime, the classical existence and uniqueness properties of solutions to equation (1.39) can be very different from that of Minkowski spacetime. Two examples will illustrate this point:

1) Let the spacetime be a flat 4-torus, with spatial periodicity \(L\) and time periodicity \(T\) such that \(T^2/L^2\) is irrational \((T\) and \(L\) do not have to be integers). Then in this spacetime the massless Klein-Gordon has as only solution \(\psi = \text{constant}\) [2].

2) Consider any spacetime with a timelike singularity, such as Minkowski spacetime with a timelike line removed, or the Schwarzschild solution with negative mass. Since equation (1.39) does not restrict what can emerge from a singularity, there is no possibility that uniqueness can hold for the solutions to (1.39) with given initial conditions on a spacelike hypersurface.

From these examples it is clear that there exists the necessity for a criterion to determine whether or not a spacetime permits the existence and uniqueness of solutions of second order hyperbolic equations, given some initial data on a hypersurface.

1.3.2 Global Hyperbolicity

There is a simple condition on a spacetime \((M, g_{\mu\nu})\) which guarantees that differential equations have a well posed initial value formulation. First, it is necessary to restrict the attention to spacetimes which are time orientable, which means that a continuous choice can be made throughout the spacetime of which half of each light cone constitutes the ‘future’ direction and which half constitutes the ‘past’.

Now, let \(\Sigma \subset M\) be any closed set of points which is achronal, i.e. no pair of points \(p, q \in \Sigma\) can be joined by a timelike curve. The domain of dependence of \(\Sigma\) is defined by

\[
D(\Sigma) = \{ p \in M | \text{every (past and future) inextendible causal curve through } p \text{ intersects } \Sigma \}.
\]  

(1.40)

If \(D(\Sigma) = M\), then \(\Sigma\) is said to be a Cauchy surface for the spacetime \((M, g_{\mu\nu})\). A Cauchy surface is then automatically a 3-dimensional hypersurface. If a spacetime admits a Cauchy surface it is said to be globally hyperbolic.

There is a very important theorem regarding the structure of globally hyperbolic spacetimes [3]:

If \((M, g_{\mu\nu})\) is globally hyperbolic with Cauchy surface \(\Sigma\), then \(M\) has topology \(\mathbb{R} \otimes \Sigma\). Furthermore, \(M\) can be foliated by a one-parameter family of smooth Cauchy surfaces \(\Sigma_t\), which means that a smooth ‘time coordinate’ \(t\) can be chosen on \(M\) such that each surface of constant \(t\) is a Cauchy surface.

Since every causal curve intersects a Cauchy surface in a unique point and causal curves do not intersect with other causal curves, one might expect to have a well defined deterministic
classical evolution from initial conditions given on $\Sigma$. That this is indeed the case for (1.39) is stated in the following theorem [1]:

Let $(M, g_{\mu\nu})$ be a globally hyperbolic spacetime with Cauchy surface $\Sigma$. Then the Klein-Gordon equation has a well posed initial value formulation in the following sense: Given any pair of smooth $C^\infty$-functions $(\psi_0, \dot{\psi}_0)$ on $\Sigma$, there exists a unique solution $\psi$ to (1.39), defined on all of $M$, such that on $\Sigma$ one has $\psi = \psi_0$ and $n^\mu \nabla_\mu \psi = \dot{\psi}_0$ where $n^\mu$ denotes the unit, future directed normal to $\Sigma$. Furthermore, for any closed subset $S \subset \Sigma$, the solution $\psi$ restricted to $D(S)$ depends only upon the initial data on $S$. In addition, $\psi$ is smooth and varies continuously with the initial data.

Similar results hold for a much more general class of linear wave equations and systems of linear wave equations. The above theorem also continues to hold if a smooth source term $f$ is inserted to the right hand side of (1.39). It then follows directly from the domain of dependence property of this theorem that there exists a unique advanced and retarded solution to the Klein-Gordon equation with source term. This means that in a globally hyperbolic spacetime there exist a unique advanced and retarded Green’s function for the Klein-Gordon equation.

For the remainder of this thesis, it is important to realize that a black hole spacetime is globally hyperbolic [2].

1.4 The Horizon

A basic phenomenon which underlies most of the discussions in further chapters is that of a massive object or cloud collapsing to a black hole. The necessary background on this process is given and the concept of a black hole horizon is introduced in the first part of this section. When doing quantum field theory in the gravitational collapse spacetime an intriguing phenomenon, the Hawking radiation, occurs, resulting from the special causal structure of black hole spacetimes. For that reason, the causal structure of black hole spacetimes is discussed in the second and third part of this section.

1.4.1 Gravitational collapse

Outside a static spherically symmetric object such as a star, the solution of the Einstein equations is given by the Schwarzschild metric [4]:

$$ds^2 = \left(1 - \frac{2GM}{r}\right) dt^2 - \left(1 - \frac{2GM}{r}\right)^{-1} dr^2 - r^2 d\Omega^2.$$  \hspace{1cm} (1.41)

This solution holds for $r$ greater than some $r_0$ which corresponds to the surface of the star and is joined for $r < r_0$ onto a solution which depends in detail on the radial distribution of density and pressure in the object. Now Birkhoff’s theorem states that the outside solution will still be part of the Schwarzschild solution cut off by the surface of the object even if this object is not static, provided that it remains spherically symmetric. This can be rigorously proven [1].
If the star is static then $r_0$ must be greater than the Schwarzschild radius $2GM$. This follows because the surface of a static star at $r_0$ must correspond to the orbit of a timelike Killing vector field, and in the Schwarzschild solution there is a timelike Killing vector field $k$ only where $r > 2GM$ because

$$k = \frac{\partial}{\partial t} \Rightarrow k^2 = g_{tt} = \left(1 - \frac{2GM}{r}\right),$$

(1.42)

so $k^2$ is positive only for $r > 2GM$. If $r_0$ were less than $2MG$, the surface would be expanding or contracting.

The process of black hole formation is generally known. A star’s nuclear fuel gets exhausted, it will cool and the pressure will be reduced, and so it will contract. Now suppose that this contraction cannot be halted by the pressure before the radius becomes less than the Schwarzschild radius, which seems likely for stars of greater than a certain mass [1]. Then since the solution outside the star is the Schwarzschild solution, there will be a closed trapped surface $S$ around the star. By a closed trapped surface is meant a closed spacelike two-surface such that the two families of null geodesics orthogonal to $S$ are converging at $S$. For a more formal definition, consider a two-dimensional, closed, convex, spacelike surface $S$ in a curved spacetime. Let $A$ be the surface area of $S$ (calculated using the induced metric on $S$). Define a time coordinate $t$ such that $t = 0$ on that surface. Suppose that the surface at $t = 0$ divides 3-space into two regions: an outer region $V_1$ and an inner region $V_2$. A small instant later, at time $t = \epsilon$, 3-space is divided into three regions:

1) an outer region $V_1$ that is spacelike separated from $S$,
2) an inner region $V_2$ that is also spacelike separated from $S$,
3) a region $V_3$ between $V_1$ and $V_2$ that can be reached by timelike geodesics from $S$. Its boundary can be reached with light-like geodesics from $S$.

Let $S_1$ be the boundary between $V_1$ and $V_3$ and $S_2$ be the boundary between $V_2$ and $V_3$. The surfaces $S_1$ and $S_2$ have areas $A_1$ and $A_2$. This situation is sketched on figure 1.2.

![Figure 1.2: 3-space divided in 3 parts by signals moving inwards and outwards of a surface $S$ over a time $\epsilon$.](image)
Now, define the expansion rates $\theta_1, \theta_2$ of these two surfaces as follows

$$\theta_1 = \frac{dA_1}{d\epsilon}, \quad \theta_2 = \frac{dA_2}{d\epsilon}$$  \hspace{1cm} (1.43)$$

Under non-exotic circumstances, such as in a flat spacetime, certainly the outer surface expansion rate is positive: $\theta_1 > 0$. The inner one is usually negative. However, inside a black hole, one can have a trapped surface. $S$ is called trapped if $\theta_1$ and $\theta_2$ are negative or zero. A surface is marginally trapped if both expansion rates are strictly equal to zero. For a pure Schwarzschild black hole, the surface $r = 2M$ is marginally trapped [5].

That there exists something like a closed trapped surface can be seen by (1.42). This equation implies that the Killing vector field generating time translations for distant observers becomes spacelike at $r = 2MG$, so moving forward in time for distant observers becomes moving forward in space on the inside of the closed trapped surface. One may think of $S$ as being in such a strong gravitational field that even the 'outgoing' light rays are dragged back and are, in fact, converging. In that case the singularity theorems of Hawking and Penrose [1] imply that a singularity will occur provided that causality is not violated and the appropriate energy condition holds. Of course, here, because the exterior solution is the Schwarzschild solution, it is obvious that there must be a singularity.

The two principal reasons why a star may depart from spherical symmetry are that it may be rotating or may have a magnetic field. One may get some idea of how large the rotation may be without preventing the occurrence of a trapped surface by considering the Kerr solution (see below). This solution can be thought of as representing the exterior solution for a body with mass $M$ and angular momentum $J = aGM$. If $a$ is less than $GM$ there are closed trapped surfaces, but if $a$ is greater than $GM$ they do not occur. Thus one might expect that if the angular momentum of the star were greater than the square of its mass, it would be able to halt the contraction of the star before a closed trapped surface developed. Another way of seeing this is that if $J = (GM)^2$ and angular momentum is conserved during the collapse, then the velocity of the surface of the star would be about the velocity of light when the star was at its Schwarzschild radius. Now many stars have an angular momentum greater than the square of their mass. However, it seems reasonable to expect some loss of angular momentum during the collapse because of braking by magnetic fields and because of gravitational radiation [1]. The situation is therefore that in some stars, and probably most, angular momentum would not prevent occurrence of closed trapped surfaces, and hence a singularity. It also appears that the rate of increase of the magnetic pressure is too slow to have a significant effect on the collapse.

Now consider a body of $10^8$ solar masses. If this collapsed to its Schwarzschild radius, the density would only be of the order of $10^{-4} g/cm^3$, which is less than the density of air [1]. If the matter were fairly cold initially, the temperature would not have risen sufficiently either to support the body or to ignite the nuclear fuel. This example shows that the conditions when a body passes through its Schwarzschild radius need not be in any way extreme.

There are several models of gravitational collapse where the dynamics can be calculated analytically. Examples are: the collapse of a ball of pressureless dust with uniform density and the case of spherical symmetric collapse with internal pressure [6–8]. In the first example, the
solution of the Einstein equations inside the dust ball consists of a portion of a closed Friedman-Lemaître universe which is glued to the Schwarzschild solution on the outside of the dust ball. This feature should be no surprise since the Friedman-Lemaître solution is defined to describe the inside of a homogeneous and isotropic mass density.

But the most instructive model of gravitational collapse is that of a thin contracting spherical shell of massless particles with total energy $M$. Inside the shell, before the passage of the particles, the spacetime is flat and thus Minkowski. Again by Birkhoff’s theorem, the spacetime outside the shell will be the Schwarzschild spacetime. The closed trapped surface will form at the origin of 3-space with its radius increasing in time up to the point it intersects with the contracting shell and from that point on, it remains a constant surface in 3-space [9]. This is depicted in figure 1.3. There is evidently no local quantity in the Minkowski spacetime inside the shell that will distinguish the presence of the closed trapped surface whose occurrence is due entirely to the future collapse of the shell.

![Figure 1.3: The formation of a closed trapped surface in a collapsing shell of massless particles.](image)

As another example, imagine a person hovering safely a certain distance above the closed trapped surface of a black hole of mass $M$. But then, an object of considerable mass $M'$ falls into the black hole, thereby increasing its mass to $M + M'$. As a result of this, the closed trapped surface will increase its radius, thereby trapping the at first perfectly safe person.

**Now the horizon of a black hole is defined as the boundary of the region of spacetime from which no signal can escape to infinity.** It is a surface that separates spacetime in an outer and an inner region. Any light ray which originates in the inner region can never reach any point on the outer region. Based on the two examples above, it is clear that the horizon is a truly global phenomenon since its location is determined by all future events. This will become more explicit when considering the Penrose diagram of gravitational collapse to a black hole later on.
By considering the Schwarzschild metric (1.41) it looks like the horizon is singular because of the divergence of $g_{rr}$ as $r \to 2MG$. However, this is not the case, as it will appear that this divergence is only a singularity of the Schwarzschild coordinates, just like the spherical coordinates fail to describe the north and south pole of a 2-sphere. It is not a singularity of the spacetime manifold. In the next two paragraphs we will investigate what exactly is going on in the region around the horizon, with a special emphasis on the causal structure of the spacetime, by considering an outside and an infalling observer.

1.4.2 Outside observers and the end of time

Consider two observers in Schwarzschild spacetime who are stuck at fixed spatial coordinate values $(r_1, \theta_1, \phi_1)$ and $(r_2, \theta_2, \phi_2)$. Then, the proper time of observer $i$ will be related to the coordinate time $t$ by

$$\frac{d\tau_i}{dt} = \left(1 - \frac{2GM}{r_i}\right)^{1/2}.$$  \hspace{1cm} (1.44)

Suppose that the observer $O_1$ emits a light pulse which travels to observer $O_2$, such that $O_1$ measures the time between two successive crests of the light wave to be $\Delta\tau_1$. Each crest follows the same path to $O_2$, except that they are separated by a coordinate time

$$\Delta t = \left(1 - \frac{2GM}{r_1}\right)^{-1/2} \Delta\tau_1.$$  \hspace{1cm} (1.45)

This separation in coordinate time does not change along the photon trajectories, but the second observer measures a time between successive crests given by

$$\Delta\tau_2 = \left(1 - \frac{2GM}{r_2}\right)^{1/2} \Delta t = \left(\frac{1 - 2GM/r_2}{1 - 2MG/r_1}\right)^{1/2} \Delta\tau_1.$$  \hspace{1cm} (1.46)
Since these intervals $\Delta \tau_i$ measure the proper time between two crests of an electromagnetic wave, the observed frequencies will be related by

$$\frac{\omega_2}{\omega_1} = \frac{\Delta \tau_1}{\Delta \tau_2} = \left( \frac{1 - 2GM/r_2}{1 - 2MG/r_1} \right)^{1/2}.$$  \(1.47\)

This is an exact result for the frequency shift.

Now consider radial null curves, thus those for which $\theta$ and $\phi$ are constant and $ds^2 = 0$.

$$0 = \left(1 - \frac{2GM}{r}\right) dt^2 - \left(1 - \frac{2GM}{r}\right)^{-1} dr^2,$$  \(1.48\)

from which follows that

$$\frac{dt}{dr} = \pm \left(1 - \frac{2GM}{r}\right)^{-1}.$$  \(1.49\)

This measures the slope of the light cones on a spacetime diagram of the $t - r$ plane. For large $r$ the slope is $\pm 1$ as in Minkowski spacetime, while as one approaches $r = 2MG$ one gets $dt/dr \to \pm \infty$, and the light cones 'close up'. This is depicted on figure 1.5.

![Figure 1.5: The closing up of the lightcones in Schwarschild spacetime.](image)

Thus a light ray which approaches $r = 2MG$ never seems to get there, at least in this coordinate system, instead it seems to asymptote to this radius. This is the description for a distant observer, collecting information available through his past light cone. It is now clear that such an observer never actually sees events on the horizon. In this sense the horizon must be regarded as at the end of time. It also follows from (1.47) that any particle or wave which falls through the horizon is seen by the distant observer as asymptotically approaching the horizon as it is infinitely redshifted.

Before the mid-1960s the object that is now called a 'black hole' was referred to in the English literature as a 'collapsed star' and in the Russian literature as a 'frozen star' [10]. The corresponding mental picture, based on stellar collapse as viewed in Schwarzschild coordinates as described above, was one of a collapsing star that contracts more and more rapidly as the grip of gravity gets stronger and stronger, the contraction then slowing because of a growing gravitational redshift and ultimately freezing to a halt at an 'infinite-redshift' surface at the Schwarzschild radius, there to hover for all eternity. From the work of Oppenheimer and Snyder
there was the awareness of an alternative viewpoint, that of an observer on the surface of the collapsing star who sees no freezing but instead experiences collapse to a singularity in a painfully short time. But because nothing inside the infinite-redshift surface can ever influence the external universe, that ‘comoving viewpoint’ seemed irrelevant for astrophysics. Thus astrophysical theorizing in the early 1960s was dominated by the ‘frozen-star viewpoint’. As long as this viewpoint prevailed, physicists failed to realize that black holes can be dynamical, evolving, energy-storing and energy-releasing objects (see section 1.8). Objects capable of colliding, vibrating widley and emitting huge bursts of gravitational waves [10].

1.4.3 Infalling observers and the equivalence principle

The fact that an outside observer never sees objects reach \( r = 2GM \) is a meaningful statement, but the fact that their trajectories in the \( t - r \) plane never reaches there is not. It is highly dependent on the coordinate system, and it is better to ask a more coordinate-independent question such as: do the infalling objects reach the Schwarzschild radius in a finite amount of proper time? The best way to do this is to change coordinates to a system which is better behaved at \( r = 2GM \).

As a first step, the tortoise coordinates are introduced. The change to tortoise coordinates from the convential spherical coordinates is made by a change of the radial coordinate that maps the horizon to minus infinity, so that the resulting coordinate system covers only the region \( r > 2MG \). The tortoise coordinate \( r^* \) is defined by

\[
\frac{1}{1 - \frac{2MG}{r}}dr^2 = \left(1 - \frac{2MG}{r}\right) (dr^*)^2,
\]

and is explicitely given by

\[
r^* = r + 2MG \ln \left(\frac{r - 2MG}{2MG}\right). \tag{1.51}
\]

In tortoise coordinates the Schwarzschild metric takes the form

\[
ds^2 = \left(1 - \frac{2MG}{r}\right) [dt^2 - (dr^*)^2] - r^2 [d\theta^2 + \sin^2 \theta d\phi^2]. \tag{1.52}
\]

Where \( r \) is to be thought of as a function of \( r^* \). The interesting point is that the radial-time part of the metric is now conformally flat. A space is called conformally flat if its metric can be brought to the form

\[
ds^2 = F(x)dx^\mu dx^\nu \eta_{\mu\nu}, \tag{1.53}
\]

with \( \eta_{\mu\nu} \) being the conventional Minkowski metric. Actually, any two-dimensional space is conformally flat, so a slice through Schwarschild spacetime at fixed \( \theta, \phi \) is no exception. The transition to tortoise coordinates represents some progress, since the light cones now don’t seem to close up, furthermore, none of the metric coefficients becomes infinite at \( r = 2MG \).

The next step is to define coordinates which are more naturally adapted to null geodesics.
Put
\[ u = t - r^* \]  \hspace{1cm} (1.54)
\[ v = t + r^*. \]  \hspace{1cm} (1.55)

Then outgoing radial null geodesics are characterized by \( u = \) constant and infalling ones satisfy \( v = \) constant. Now consider going back to the original radial coordinate \( r \), but replacing the timelike coordinate \( t \) with the new coordinate \( v \). These coordinates are known as Eddington-Finkelstein coordinates, in terms of which the metric becomes
\[ ds^2 = \left( 1 - \frac{2MG}{r} \right) dv^2 - 2dvdr - r^2d\Omega. \]  \hspace{1cm} (1.56)

Here, there is a first sign of progress. The determinant of the metric is
\[ g = -r^4 \sin^2 \theta, \]  \hspace{1cm} (1.57)
which is perfectly regular at \( r = 2GM \). Therefore, the metric is invertible and it is clear once and for all that \( r = 2GM \) is simply a coordinate singularity in the original system \((t,r,\theta,\phi)\). In the Eddington-Finkelstein coordinates the condition for radial null curves is solved by
\[ \frac{dv}{dr} = \begin{cases} 
0 & \text{(infalling)} \\
2 \left( 1 - \frac{2GM}{r} \right)^{-1} & \text{(outgoing)}
\end{cases} \]  \hspace{1cm} (1.58)

One can therefore see what happens: in this coordinate system the light cones remain well-behaved at \( r = 2MG \), and this surface is at finite coordinate value. There is no problem in tracing the paths of null or timelike particles past the horizon. On the other hand, something interesting is certainly going on. Although the light cones don’t close up, they do tilt over, such that for \( r < 2GM \) all future-directed paths are in the direction of decreasing \( r \). It is this aspect of the causal structure of black hole spacetimes that is so special. One could see this as if space and time 'have switched roles'. It is this effect that already was anticipated on by (1.42). This tilting of the light cone is illustrated in figure 1.6.

![Figure 1.6: Tilting of the light cone in Eddington-Finkelstein coordinates.](image.png)

The complete picture of the causal structure around a Schwarschild horizon, with the closing up and the tilting of the light cone is given in figure 1.7.
It is comforting to have found this second point of view. Because one of the cornerstones of general relativity, the equivalence principle, states that a freely falling observer locally observes Minkowski spacetime, so he should not feel anything strange when passing the horizon. It is only at the singularity that things start to go terribly wrong.

### 1.5 The maximal extension of Schwarzschild spacetime

In section 1.2.3 the original coordinate $t$ was changed to the new one $v$, which had the nice property that if one decreases $r$ along a radial null curve $v = \text{constant}$, one goes right through the event horizon without any problems. From this it is clear that the initial coordinate system didn’t do a good job of covering the entire manifold. The region $r < 2GM$ should certainly be included in the spacetime, since physical particles can easily reach there and pass through. However, there is no guarantee that now the entire spacetime is covered, perhaps there are other directions in which the manifold can be extended.

Notice that in the $(v, r)$ coordinate system the event horizon can be crossed on future-directed paths, but not on past-directed ones. This seems unreasonable, since the Schwarzschild solution is time-independent. But $u$ could have been chosen as coordinate instead of $v$, in which case the metric would have been

$$ds^2 = \left(1 - \frac{2GM}{r}\right) du^2 + (dudr + drdu) - r^2 d\Omega^2. \quad (1.59)$$
Now one can once again pass through the event horizon, but this time only along past-directed curves. This means that one can consistently follow either future-directed or past-directed curves through \( r = 2MG \) but arrive at different places. This was to be expected since from the definitions (1.54) and (1.55) it follows that if \( v \) is constant and \( r \) decreases that \( t \to +\infty \), while if \( u \) is constant and \( r \) decreases, then \( t \to -\infty \) because the tortoise coordinate goes to \(-\infty\) as \( r \to 2GM \). Therefore, the spacetime is extended in two different regions, one to the future and one to the past.

![Figure 1.8: Crossing the horizon at constant \( u \) along past directed curves.](image)

The next step would be to follow space-like geodesics and see if more regions are uncovered. But a shortcut to the process is by defining coordinates that are good all over. A first guess might be to use both \( u \) and \( v \) at once instead of \( t \) and \( r \), which leads to

\[
    ds^2 = \frac{1}{2} \left( 1 - \frac{2GM}{r} \right) (dvdu + dudv) - r^2d\Omega^2 ,
\]

with \( r \) defined implicitly in terms of \( v \) and \( u \) by

\[
    \frac{1}{2}(v - u) = r^\star.
\]

With this, the degeneracy with which we started out is re-introduced, in these coordinates \( r = 2MG \) is infinitely far away, at either \( v = -\infty \) or \( u = +\infty \). The thing to do is to change coordinates which pull these points into finite coordinate values, a good choice is

\[
    U = -\exp\left(-\frac{u}{4GM}\right),
\]

\[
    V = \exp\left(\frac{v}{4GM}\right),
\]

which in terms of the original \((t,r)\) coordinates is

\[
    U = - \left(\frac{r}{2GM} - 1\right)^{1/2} e^{(r-t)/4GM},
\]

\[
    V = \left(\frac{r}{2GM} - 1\right)^{1/2} e^{(r+t)/4GM}.
\]

In these coordinates the Schwarzschild metric is

\[
    ds^2 = \frac{32GM^3}{r} e^{-r/2GM} dUdV - r^2d\Omega^2 .
\]
Both $U$ and $V$ are null coordinates or ’radial light-like’ variables, in the sense that their corresponding vector fields $\partial/\partial U$ and $\partial/\partial V$ are null vectors. They are called the Kruskal-Szekeres coordinates in their light cone variant.

The surfaces of constant $r > 2GM$ are timelike hyperbolas in sector $I$ of figure 1.9, given by

$$UV = \text{negative constant}. \tag{1.67}$$

As $r$ tends to $2MG$ the hyperbolas become the boken straight lines $H^+$ and $H^-$ which are called the extended past and future horizons and satisfy

$$UV = 0. \tag{1.68}$$

Although the extended horizons lie at finite value of the Kruskal-Szekeres coordinates, they are located at Schwarzschild time $\pm \infty$. So a particle trajectory which crosses $H^+$ in a finite proper time, crosses $r = 2GM$ only after an infinite Schwarzschild time.

The region $r < 2MG$ is region $II$ on figure 1.9. In this region the surfaces of constant $r$ are the space-like hyperboloids given by

$$UV = \text{positive constant}. \tag{1.69}$$

The singularity at $r = 0$ occurs at

$$UV = 1. \tag{1.70}$$

Figure 1.9: The maximal extended Schwarzschild spacetime in Kruskal coordinates.
Chapter 1. Classical aspects of black holes

Now it is important to realize that equations (1.67), (1.68), (1.69) and (1.70) refer to a bigger coordinate interval than used above. The Kruskal-Szekeres coordinates should be allowed to range over every value they can take without hitting the singularity at \( r = 0 \). It is clear from figure 1.9 that up to now, only positive \( V \) are considered. By adding also the negative values for \( V \), regions \( III \) and \( IV \) are introduced on figure 1.9 and the maximal extension for the Schwarzschild geometry is obtained.

The Kruskal-Szekeres coordinates have the nice property that outgoing radial null geodesics are given by \( U = \text{constant} \) and incoming radial null geodesics by \( V = \text{constant} \). So all light rays and timelike trajectories lie within a two-dimensional light cone bounded by \( 45^\circ \) lines in figure 1.9. Also a nonradially directed light ray or time-like trajectory always lies inside the two-dimensional light cone. With this in mind, it is easy to understand the causal properties of the maximal analytic extension of the Schwarzschild geometry in figure 1.9. An observer in region \( II \) can send signals across \( H^+ \) to region \( I \) and also to infinity. All signals sent in region \( II \) must eventually hit the future singularity. From region \( III \) no signal can ever get to region \( I \), so it is also behind the horizon. On the other hand, observers in region \( IV \) can communicate with region \( I \) by sending signals across \( H^- \). Region \( I \) however, cannot communicate with region \( IV \). All of this is usually described by saying that regions \( II \) and \( III \) are behind the future horizon while regions \( III \) and \( IV \) are behind the past horizon. Region \( IV \) is the time-reverse of region \( II \), the black hole, and is a part of spacetime from which things can escape to us, while nothing can get in. This time-reversed black hole is called a white hole. There is a singularity in the past, out of which the universe appears to spring. Region \( III \) is another asymptotically flat region of spacetime, a mirror image of ours that is not able to communicate with us at region \( I \) either forward or backward in time.

### 1.6 Killing horizons and surface gravity

The concept of a Killing horizon is of vital importance for the discussion of black holes:

**Killing horizon**  A null hypersurface \( N \) is a Killing horizon of a Killing vector field \( \xi \) if on \( N \), \( \xi \) is normal to \( N \).

Now let the vector field \( l \) be the normal to \( N \). In section 1.2 it was shown that \( l \) can be chosen such that \( l \cdot \nabla l^\mu = 0 \). Since \( N \) is a Killing horizon of the Killing vector field \( \xi \), one then has

\[
\xi = \tilde{f} l
\]

for some spacetime function \( \tilde{f} \). So it follows that

\[
\xi \cdot \nabla \xi^\mu = \tilde{f} l^\nu \nabla_\sigma (\tilde{f} l^\mu)
= \tilde{f} l^\nu (\partial_\sigma \tilde{f}) l^\mu + \tilde{f} l^\nu \tilde{f} \nabla_\sigma l^\mu
= \tilde{f} l \cdot \partial \tilde{f} l^\mu
= \xi \cdot \partial \ln |\tilde{f}| \xi^\mu,
\]  

(1.72)
where $\kappa = \xi \cdot \partial \ln|\tilde{f}|$ is called the surface gravity.

This definition determines $\kappa$ up to a constant factor. Since $\xi^2 = 0$ on the null surface $N$, there is no natural normalization for $\xi$. But in an asymptotically flat spacetime there is a natural normalization at spatial infinity. For example, for a time-translation Killing vector field $k$ one can choose
\[k^2 \to 1 \quad \text{as} \quad r \to +\infty.\] (1.73)
This fixes $k$, and hence $\kappa$, up to a sign. The sign of $\kappa$ is fixed by requiring $k$ to be future-directed.

Since $\xi$ is hypersurface orthogonal at the horizon, by Frobenius’s theorem (see appendix A), one has
\[\xi_{[\mu} \nabla^{\nu} \xi_{\sigma]} = 0. \quad (\text{on the horizon})\] (1.74)
Using the Killing vector lemma $\nabla_{\nu} \xi_{\sigma} = -\nabla_{\sigma} \xi_{\nu}$, this implies
\[\xi_{\sigma} \nabla_{\nu} \xi_{\mu} = -2 \xi_{[\mu} \nabla_{\nu]} \xi_{\sigma}\] (1.75)
on the horizon. Contracting with $\nabla^{\mu} \xi^{\nu}$, one finds
\[\xi_{\sigma} (\nabla^{\mu} \xi^{\nu})(\nabla_{\mu} \xi_{\nu}) = -2 (\xi_{\mu} \nabla^{\mu} \xi^{\nu})(\nabla_{\nu} \xi_{\sigma})\]
\[= -2 \kappa \xi^{\mu} \nabla_{\nu} \xi_{\sigma}\]
\[= -2 \kappa^2 \xi_{\sigma}.\] (1.76)
Thus, one obtains a simple explicit formula for $\kappa$,
\[\kappa^2 = -\frac{1}{2} (\nabla^{\mu} \xi^{\nu})(\nabla_{\mu} \xi_{\nu}),\] (1.77)
where evaluation on the horizon is understood. This equation provides us with a physical interpretation of $\kappa$ in the following way. One has everywhere, i.e. not just on the horizon,
\[3(\xi^{[\mu} \nabla^{\nu] \xi^{\sigma]})(\xi_{[\mu} \nabla_{\nu} \xi_{\sigma]}) = \xi^{\mu} \xi_{[\mu} (\nabla^{\nu} \xi^{\sigma]})(\nabla_{\nu} \xi_{\sigma]) - 2 (\xi^{\mu} \nabla^{\nu} \xi^{\sigma]})(\xi_{\nu} \nabla_{\mu} \xi_{\sigma}).\] (1.78)
Since $\xi_{[\mu} \nabla_{\nu} \xi_{\sigma]} = 0$ on the horizon, the gradient of the left-hand side vanishes on the horizon. On the other hand, by (4.162), $\nabla_{\nu} (\xi^{\mu} \xi_{\mu})$ does not vanish when $\kappa$ is not equal to zero. Hence, by l’Hôpital’s rule, the left side of equation (4.148) divided by $\xi^{\mu} \xi_{\mu}$ must approach zero on the horizon. Thus, using (4.151), one finds
\[\kappa^2 = \lim[-(\xi^{\mu} \nabla_{\nu} \xi^{\sigma]})(\xi^{\mu} \nabla_{\mu} \xi_{\sigma})/\xi^{\mu} \xi_{\nu}],\] (1.79)
where 'lim' stands for the limit approaching the horizon. Now, a particle on a time-like orbit of a Killing vector field $\xi$ has a 4-velocity
\[u^\mu = \frac{\xi^\mu}{|\xi|}.\] (1.80)
since \( u \propto \xi \) and \( u \cdot u = 1 \). Therefore, its proper four-vector acceleration is

\[
a^\mu = u \cdot \nabla u^\mu = \frac{\xi \cdot \nabla \xi^\mu}{\xi^2} - \frac{(\xi \cdot \partial \xi^2) \xi^\mu}{2 \xi^2}.
\]

But for a Killing vector field, \( \xi \cdot \partial \xi^2 = 2 \xi^\mu \xi^\nu \nabla_\mu \xi_\nu = 0 \) by the Killing vector lemma, so

\[
a^\sigma = \frac{\xi^\nu \nabla_\nu \xi^\sigma}{\xi^\mu \xi_\mu} \tag{1.82}
\]

in (1.79) is just the proper acceleration on an orbit of \( \xi \). Thus, we find

\[
\kappa = \lim (|\xi| a), \tag{1.83}
\]

where \( a = \sqrt{a^\sigma a^\sigma} \) and \( |\xi| = \sqrt{\xi^\mu \xi_\mu} \). In the case of a static black hole, one has \( \xi = k \). Then \( |\xi| \) is just the redshift factor, and \( |\xi| a \) is the force that must be exerted at infinity to hold a unit test mass in place [11]. Thus, \( \kappa \) is the limiting value of this force at the horizon, which explains the name surface gravity. (Of course, the locally exerted force \( a \) becomes infinite at the horizon.) As we will see later, for a rotating black hole, a test mass cannot be held stationary with respect to infinity near the black hole, but one continues to refer to \( \kappa \) as the surface gravity.

It can be shown for a general Killing horizon that \( \kappa \) is constant on orbits of \( \xi \) [12]. Now suppose that \( \kappa \neq 0 \) on one orbit of \( \xi \) in \( \mathcal{N} \). Then this orbit coincides with only a part of a null generator of \( \mathcal{N} \). To see this, one can choose coordinates on \( \mathcal{N} \) such that

\[
\xi = \frac{\partial}{\partial \alpha} \tag{1.84}
\]

at all points where \( \xi \neq 0 \). This means that the group parameter \( \alpha \) is one of the coordinates. Then if \( \alpha = \alpha(\lambda) \) on an orbit of \( \xi \) with an affine parameter \( \lambda \), one has

\[
\xi|_{\text{orbit}} = \frac{d\lambda}{d\alpha} \frac{d}{d\lambda} = \tilde{f} l. \tag{1.85}
\]

So

\[
f = \frac{d\lambda}{d\alpha} \quad \text{and} \quad l = \frac{d}{d\lambda} = \frac{dx^\mu(\lambda)}{d\lambda} \partial_\mu.
\]

Then, the surface gravity \( \kappa \) is by definition

\[
\kappa = \frac{\partial}{\partial \alpha} \ln |\tilde{f}|, \tag{1.87}
\]

where \( \kappa \) is constant on an orbit on \( \mathcal{N} \). Thus, for such orbits, \( f = f_0 e^{\kappa \alpha} \) for an arbitrary constant \( f_0 \). Because of the freedom to shift \( \alpha \) by a constant, one can choose \( f_0 = \pm \kappa \) without loss of generality. This choice implies

\[
\frac{d\lambda}{d\alpha} = \pm \kappa e^{\kappa \alpha} \Rightarrow \lambda = \pm e^{\kappa \alpha}, \tag{1.88}
\]

where the integration constant was chosen to be zero. As \( \alpha \) ranges from \(-\infty \) to \(+\infty \), one covers either the portion \( \lambda > 0 \) of the generator of \( \mathcal{N} \), or the portion \( \lambda < 0 \), depending on the sign
choise in (1.88). The bifurcation point \( \lambda = 0 \) is a fixed point of \( \xi \), which can be shown to be a 2-sphere, called the bifurcation 2-sphere. A Killing horizon satisfying these properties is called a bifurcate Killing horizon. The general structure of a bifurcate Killing horizon is given on figure 1.10.

![Figure 1.10: A bifurcate Killing horizon.](image)

Now consider the Schwarzschild metric in ingoing Eddington-Finkelstein coordinates (1.56) as derived in section 1.4.3. The vector field normal to the family of surfaces \( S = r = \text{constant} \) according to (1.30) is given by

\[
l = f(r) \left[ \left( 1 - \frac{2MG}{r} \right) \frac{\partial S}{\partial r} + \frac{\partial S}{\partial v} + \frac{\partial S}{\partial v} \frac{\partial}{\partial r} \right] = f(r) \left[ \left( 1 - \frac{2MG}{r} \right) \frac{\partial}{\partial r} + \frac{\partial}{\partial v} \right]. \tag{1.89} \]

From which follows

\[
l^2 = g^{\mu\nu} \partial_{\mu} S \partial_{\nu} S f^2 = g^{rr} f^2 = \left( 1 - \frac{2GM}{r} \right) f^2. \tag{1.90} \]

So the horizon \( r = 2MG \) is a null surface, with normal

\[
l_{|r=2MG} = f \frac{\partial}{\partial v}. \tag{1.91} \]

In Kruskal-Szekeres coordinates the horizon is given by \( U = 0 \). The vector field normal to the family of surfaces \( U = \text{constant} \) is

\[
\hat{l} = \frac{fr}{32M^3} v^{r/2M} \frac{\partial}{\partial V}, \tag{1.92} \]
where (1.66) was used. So at the horizon this becomes

\[ \hat{l}|_N = \frac{f e}{16M^2} \frac{\partial}{\partial V}. \] (1.93)

Because \( g^{VV} \) is zero in the Kruskal-Szekeres metric, \( \hat{l}^2 \) is identically zero, so \( U = \) constant is a null surface for any constant. It also follows that \( \partial^\mu \hat{l}_\mu \) is zero, so (1.35) implies that \( \hat{l} \cdot \nabla \hat{l} = 0 \) if \( f \) is constant. By choosing \( f = 16M^2e^{-1} \) the normal to the horizon \( U = 0 \) becomes

\[ \hat{l} = \frac{\partial}{\partial V}. \] (1.94)

Now take \( \xi = k \), with \( k \) the time-translation Killing vector field of the stationary black hole spacetime as normalized above. Making use of (1.65), \( k \) can be expressed in terms of the Kruskal-Szekeres coordinates as

\[ k = \frac{\partial}{\partial t} = \frac{\partial U}{\partial V} \frac{\partial}{\partial U} + \frac{\partial V}{\partial t} \frac{\partial}{\partial V} = -\frac{1}{4MG} U \frac{\partial}{\partial U} + \frac{1}{4MG} V \frac{\partial}{\partial V}, \] (1.95)

Where only region \( I \) of the Kruskal-Szekeres spacetime is considered (see figure 1.9). So on the future horizon \( U = 0 \), \( k \) is given by

\[ k = \frac{1}{4MG} V \frac{\partial}{\partial V} = \tilde{f} \hat{l}, \] (1.96)

where it follows from (1.94) that

\[ \tilde{f} = \frac{1}{4MG} V. \] (1.97)

So the horizon \( U = 0 \) is a Killing horizon of \( k \). The surface gravity is

\[ \kappa = k \cdot \partial \ln |\tilde{f}| = \frac{1}{4MG} V \frac{\partial}{\partial V} \ln |\frac{1}{4MG} V| = \frac{1}{4MG}. \] (1.98)

Reinstalling factors of \( c \), the surface gravity of a Schwarzschild black hole is \( \kappa = c^3/4GM \).

1.7 Penrose diagrams

When considering problems involving black holes frequent use is made of the so-called Penrose diagrams, which are a very convenient schematic way to represent spacetimes. In this section the main features of Penrose diagrams are presented, with an emphasis on the specific examples to be used later in this thesis.

1.7.1 Conformal compactification

A Penrose diagram is basically obtained by performing two subsequent transformations (each one possibly in multiple steps). The first one, which is not always necessary, is a coordinate
transformation ensuring that radial null geodesics lie at $\pm 45^\circ$. The second is a conformal transformation which respects angles but changes distances. The purpose of this second transformation is to bring infinity at finite distance so that a compact representation of the spacetime can be made. To be more precise, all the points at infinity in the original metric should be at a finite affine parameter value in the new metric. The recipe can’t be made more specific because it varies for different spacetimes. This will be illustrated by the following two examples.

**Minkowski spacetime**

Take the Minkowski metric in spherical coordinates

$$ds^2 = dt^2 - dr^2 - r^2d\Omega^2.$$  \hfill (1.99)

Radial light rays propagate on the light cone $dt \pm dr = 0$. This means that the first coordinate transformation putting radial null geodesics at $\pm 45^\circ$ is not necessary here. Under the transformation

$$u' = t - r$$
$$v' = t + r$$  \hfill (1.100)

the Minkowski metric becomes

$$ds^2 = du'dv' + \frac{1}{4}(u' - v')^2d\Omega^2.$$  \hfill (1.101)

Now set

$$u' = \tan \omega \quad - \frac{\pi}{2} < \omega < \frac{\pi}{2}$$
$$v' = \tan \eta \quad - \frac{\pi}{2} < \eta < \frac{\pi}{2},$$  \hfill (1.102)

where $\eta \geq \omega$ since $r \geq 0$. In these coordinates the metric is

$$ds^2 = (2 \cos \omega \cos \eta)^{-2}[4d\omega d\eta - \sin^2(\eta - \omega)d\Omega^2].$$  \hfill (1.103)

To approach infinity in this metric one must take $|\omega| \to \pi/2$ or $|\eta| \to \pi/2$, so by taking

$$\Lambda = 2 \cos \omega \cos \eta$$  \hfill (1.104)

these points are brought to finite affine parameter in the new metric obtained by the conformal transformation

$$ds^2 = \Lambda^2 ds^2 = 4d\omega d\eta - \sin^2(\eta - \omega)d\Omega^2.$$  \hfill (1.105)

Now the points at infinity can be added. Taking the restriction $\eta \geq \omega$ into account, these are

1) \begin{align*}
\omega &= -\pi/2 \\
\eta &= \pi/2
\end{align*} \quad \Leftrightarrow \quad \begin{align*}
u' &\to -\infty \\
v' &\to +\infty
\end{align*} \quad \Leftrightarrow \quad \begin{align*}
r &\to +\infty \\
t &\text{is finite} \quad \Rightarrow \text{spatial infinity } i^0
\end{align*}  \hfill (1.106)
2) \[
\begin{align*}
\omega &= \pm \pi/2 \\
\eta &= \pm \pi/2 \\
u' &= \pm \infty \\
v' &= \pm \infty \\
r &\text{ is finite} \\
t &\rightarrow \pm \infty \\
\Rightarrow & \text{ future/past timelike infinity } i^\pm
\end{align*}
\] (1.107)

3) \[
\begin{align*}
\omega &= -\pi/2 \\
\eta &\neq \pi/2 \\
u' &\rightarrow -\infty \\
v' &\text{ is finite} \\
r &\rightarrow +\infty \\
t &\rightarrow -\infty \\
\Rightarrow & \text{ past null infinity } I^- 
\end{align*}
\] (1.108)

4) \[
\begin{align*}
\omega &\neq +\pi/2 \\
\eta &= +\pi/2 \\
u' &\text{ is finite} \\
v' &\rightarrow +\infty \\
r &\rightarrow +\infty \\
t &\rightarrow +\infty \\
\Rightarrow & \text{ future null infinity } I^+ 
\end{align*}
\] (1.109)

These points together form conformal infinity. They are not part of the original spacetime. Minkowski spacetime is now conformally embedded in the new spacetime described by the metric \( d\tilde{s}^2 \) with boundary at \( \Lambda = 0 \).

Introducing the new time and space coordinates \( \tau, \chi \) by
\[
\begin{align*}
\tau &= \eta + \omega \\
\chi &= \eta - \omega
\end{align*}
\] (1.110)
the metric becomes
\[
d\tilde{s}^2 = \Lambda^2 ds^2 = d\tau^2 - d\chi^2 - \sin^2 \chi d\Omega^2,
\] (1.111)
with \( \Lambda = \cos \tau + \cos \chi \).

The original coordinates \( (t,r) \) are related to \( (\tau,\chi) \) by
\[
\begin{align*}
2t &= \tan \left( \frac{1}{2}(\tau + \chi) \right) + \tan \left( \frac{1}{2}(\tau - \chi) \right) \\
2r &= \tan \left( \frac{1}{2}(\tau + \chi) \right) - \tan \left( \frac{1}{2}(\tau - \chi) \right)
\end{align*}
\] (1.112)
\( \chi \) is an angular variable which must be identified modulo \( 2\pi \). If no other restriction is placed on the ranges of \( \tau \) and \( \chi \), this metric \( d\tilde{s}^2 \) is that of the Einstein static universe, which has the topology \( \mathbb{R}(\text{time}) \otimes S^3(\text{space}) \). The 2-spheres of constant \( \chi \neq 0 \), have radius \( |\sin \chi| \) (the points \( \chi = 0, \pi \) are the poles of the spherical coordinate system describing a 3-sphere). The entire manifold \( \mathbb{R} \otimes S^3 \) can be drawn as a cylinder, in which each circle is a 3-sphere. Each point \( (\tau, \chi) \) on the cylinder represents the half of a 2-sphere, where the other half is the point \( (\tau, -\chi) \). This is represented in figure 1.11. The shaded region represents the sector corresponding to \(-\pi \leq \tau + \chi \leq \pi \) and \(-\pi \leq \tau - \chi \leq \pi \) as follows from (1.114) and (1.110).
The restriction $r \geq 0$ now becomes $\chi \geq 0$. This should not lead to the wrong conclusion, it is indeed the entire shaded region that represents the compactified Minkowski spacetime, and not only half of it. The way the topology $\mathbb{R} \otimes S^3$ is depicted is invariant under $\chi \rightarrow -\chi$. Restricting $\chi$ to be positive just picks one of these two possible representations. It are indeed the entire 2-spheres, represented by opposite points on the circles that make up Minkowski spacetime. So actually, the restriction is fulfilled quite naturally.

Identifying opposite points as one two-sphere of surface $4\pi \sin^2 \chi$ and representing them by a single point results in a triangle representing Minkowski spacetime, as seen in figure 1.12. This is the Penrose diagram of Minkowski spacetime. As a result of this construction every point represents a 2-sphere, except $i^0$, $i^+$, $i^-$ and the $r = 0$ line.

Spatial sections of the compactified spacetime are topologically $S^3$ because of the addition of the point $i^0$. Thus, they are compact but have no boundary. This is not true for the whole spacetime. Asymptotically it is possible to identify points on the boundary of the compactified spacetime to obtain a compact manifold without boundary. In general, this is not possible because $i^+$ and $i^-$ can be singular points which cannot be added. This will be the case in the next example.

**Schwarzschild spacetime**

Take the metric of Schwarzschild spacetime in the form (1.60) as derived in section 1.3:

$$ds^2 = \frac{1}{2} \left( 1 - \frac{2GM}{r} \right) (dvdu + dudv) - r^2d\Omega^2.$$  \hspace{1cm} (1.113)

Again, this metric already has the property that radial null geodesics lie at $\pm 45^\circ$, so only a conformal transformation has to be applied taking infinity at finite affine parameter value.
Essentially the same transformation as in Minkowski spacetime can be used

\[ u = \tan \omega \quad -\frac{\pi}{2} < \omega < \frac{\pi}{2} \]
\[ v = \tan \eta \quad -\frac{\pi}{2} < \eta < \frac{\pi}{2} \].

(1.114)

With this transformation the metric becomes

\[ ds^2 = (2 \cos \omega \cos \eta)^{-2} \left[ 4 \left(1 - \frac{2GM}{r}\right) d\omega d\eta - r^2 \cos^2 \omega \cos^2 \eta d\Omega^2 \right]. \]

(1.115)

Using the fact that

\[ r^* = \frac{1}{2}(v - u) = \frac{\sin(\eta - \omega)}{2 \cos \omega \cos \eta} \]

(1.116)

one has

\[ ds^2 = \Lambda^2 ds^2 = 4 \left(1 - \frac{2GM}{r}\right) d\omega d\eta - \left(\frac{r}{r^*}\right)^2 \sin^2(\eta - \omega) d\Omega^2, \]

(1.117)

where

\[ \Lambda = 2 \cos \omega \cos \eta. \]

(1.118)

In this metric the asymptotic flatness of Schwarzschild spacetime is manifest. It approaches the metric of compactified Minkowski spacetime (1.105) as \( r \to \infty \), with or without fixing \( t \). This means that \( i^0 \) and \( I^\pm \) can be added as before. All \( r = \) constant hypersurfaces meet at \( i^+ \), including the \( r = 0 \) hypersurface, which is singular, so \( i^+ \) is a singular point. Similarly for \( i^- \), so these points cannot be added. Near \( r = 2GM \) one can introduce Kruskal-Szekeres type coordinates to pass through the horizon. In this way the Penrose diagram for the maximal...
extended Schwarzschild spacetime can be constructed. This is shown on figure 1.13. Sometimes it is convenient to adjust the function $\Lambda$ so that $r = 0$ is a vertical line.

\begin{figure}[h]
\centering
\includegraphics[width=0.6\textwidth]{penrose_diagram}
\caption{The Penrose diagram the maximal extended Schwarzschild spacetime.}
\end{figure}

1.7.2 Gravitational collapse spacetime

From the two examples discussed above the Penrose diagram of the important spacetime of gravitational collapse can be constructed. This spacetime has much more physical relevance since the static Schwarzschild spacetime is an idealization.

Consider again the case of a spherical symmetric contracting shell of massless particles. This shell can be represented by an incoming light-like line in the Penrose diagram of Minkowski space (figure 1.12). The infalling shell divides the Penrose diagram into two regions, an inner and an outer region. The inner region is the interior of the shell throughout time and represents the initial flat spacetime before the shell passes. Since we are only interested in the inner part here, one could really perform the mental cutting and removing of the outer region of the Penrose diagram since this part has to be modified due to the gravitational field exterior to the mass $M$.

As already mentioned in section 1.2.1, Birkhoff’s theorem states that the spacetime outside the collapsing shell will be that portion of Schwarzschild spacetime cut off by the surface of the shell. Now the same procedure can be done in the Penrose diagram of Schwarzschild spacetime (figure 1.13), but this time the outer region is of interest and the inner region can be removed.

The inner region of Minkowski spacetime can be matched onto the outer region of Schwarzschild spacetime to form the Penrose diagram of a gravitational collapse spacetime, as illustrated in figure 1.14. This matching procedure must be done so that the radius of the local two sphere represented by the angular coordinates $(\theta, \phi)$ is continuous. In other words, the mathematical identification of the boundaries of the two regions must respect the continuity of the variable $r$. Since in both cases $r$ varies smoothly from $r = \infty$ at $I^-$ to $r = 0$, the identification is
always possible. One should not worry if the two sides needed to be matched don’t have the
same length because we are working with conformal pictures and the appropriate stretching or
contracting can be performed without changing the physics of these diagrams. That is to say,
this deformation will not disturb the form of the light cones.

The final Penrose diagram of a gravitational collapse spacetime makes explicitely clear what
was referred to in section 1.4.1, namely that the horizon already forms in the Minkowski space-
time inside the contracting shell. It is readily seen that any light ray or timelike trajectory
originating behind the dotted line $H$ on figure 1.14 cannot escape to infinity but only end up at
the singularity. This defines $H$ as the horizon. In the outer region, $H$ is identical to the surface
$H^+$ considered above, and therefore it is found at $r = 2MG$. In the inner region, the value of $r$
on the horizon grows from an initial value $r = 0$ to the value $r = 2MG$ when crossing the shell.
Notice that $H$ is given by a line at $45\degree$, implying that it is a null surface, as seen in section 1.6.

Although this model might seem like an unrealistic idealization, it contains all the necessary
features to understand a general gravitational collapse spacetime. In figure 1.15 the Penrose
diagram for a general gravitational collapse is shown.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure14.png}
\caption{The construction of the spacetime of a gravitational collapse to a black hole.}
\end{figure}
1.8 No hair conjecture

In the previous sections we made use of Birkhoff’s theorem, which ensures the uniqueness of the Schwarzschild solution on the outside of a spherical symmetric object. In other words, it states that the part of the spacetime on the outside of an object that remains spherical symmetric is necessarily static. It will appear that this idea of uniqueness of a spacetime solution to the Einstein equations can be extended to more general black hole spacetimes.

1.8.1 Kerr-Newman geometry and uniqueness theorems

In order to make a first step towards generalizing the idea of uniqueness we introduce the Einstein-Maxwell action. This is the generalization of the Einstein-Hilbert action which takes into account the electromagnetic field. It is given by

$$ S = \frac{1}{16\pi G} \int d^4x (-g)^{1/2} [R - F_{\mu\nu}F^{\mu\nu}] $$

(1.119)

where $R$ is the scalar curvature and $F_{\mu\nu}$ is the electromagnetic tensor. The unusual normalization of the Maxwell term means that the magnitude of the Coulomb force between two point charges $Q_1$, $Q_2$ at large separation $r$ in flat space is

$$ \frac{G|Q_1Q_2|}{r^2} $$

(1.120)

which implies the use of ‘geometrized’ units of charge. The source-free Einstein-Maxwell equations derived from (1.119) are

$$ G_{\mu\nu} = 2 \left( F_{\mu\lambda}F^\lambda_\nu - \frac{1}{4} g_{\mu\nu} F_{\rho\sigma}F^{\rho\sigma} \right) $$

(1.121)

$$ \nabla_\mu F^{\mu\nu} = 0. $$

(1.122)
The right hand side of (1.121) is the energy-momentum tensor of the electromagnetic field. The source-free Einstein-Maxwell equations have the spherically symmetric static Reissner-Nordström solution which generalizes the Schwarzschild solution

$$\begin{align*}
\sigma^2 &= \left(1 - \frac{2GM}{r} + \frac{GQ^2}{r^2}\right) dt^2 - \frac{dr^2}{\left(1 - \frac{2GM}{r} + \frac{GQ^2}{r^2}\right)} - r^2 d\Omega^2 \\
A &= \frac{Q}{r} dt,
\end{align*}$$

(1.123)

(1.124)

Where $A$ is the Maxwell 1-form, i.e. the generalization of the electromagnetic potential to a general manifold, and $Q$ is clearly the electric charge. The time component of the metric changes sign when $r$ equals $r = G(M \pm \sqrt{M^2 - Q^2})$. So the Reissner-Nordström solution only has a horizon when $M \geq Q$. In the other case, the electrostatic repulsion would halt the gravitational collapse before a black hole is formed.

Now Birkhoff’s theorem can be generalized straightforwardly to the Reissner-Nordström case, stating that the outside of a spherically symmetric charged object is described by the Reissner-Nordström metric. Thus, also in the case of a charged object, spherical symmetry implies time-independence of the metric on the outside of the object.

So far, nothing surprising has happened. Even in Newtonian gravity, the gravitational field on the outside of a spherical symmetric mass distribution is the same as if the entire mass would be located in the centre of the distribution. So in a sense, Birkhoff’s theorem could be seen as a relativistic generalization of this feature of Newtonian gravity. But there is more, much more, going on about the idea of uniqueness in the context of black holes. To be able to formulate the theorems below in their simplest form, i.e. without having to use statements involving topology, a theorem by Hawking is used\[13, 14\]:

A stationary black hole must have a horizon with spherical topology and it must either be static (zero angular momentum), axially symmetric, or both.

A asymptotically flat spacetime is stationary if and only if there exists a Killing vector field $k$ that is timelike near infinity. This means, outside a possible horizon, $k = \partial/\partial t$, where $t$ is a time coordinate. A stationary spacetime is static at least near infinity if it is also invariant under time-reversal. This requires $g_{00} = 0$. An asymptotically flat spacetime is axisymmetric if there exists an axial Killing vector field $m$ that is spacelike near infinity and for which all orbits are closed. Coordinates can be chosen such that $m = \partial/\partial \phi$, where $\phi$ is a coordinate identified modulo $2\pi$, such that $m^2/r^2 \to 1$ as $r \to +\infty$. Thus, as for $k$, there is a natural choice of normalization for an axial Killing vector field in an asymptotically flat spacetime.

With Hawking’s theorem, the first uniqueness theorem by Israel \[15, 16\] reads:

Any static black hole has external fields determined uniquely by its mass $M$ and charge $Q$. Moreover, those external fields are the Schwarzschild solution if $Q = 0$ and the Reissner-Nordström solution if $Q \neq 0$. 

Thus, for a black hole Birkhoff’s theorem also works in the reversed direction, i.e. any static black hole has to be spherical symmetric. In general, this is not the case for an object like a star. So black holes really have a special nature compared to ‘normal’ objects. Still, Israel’s theorem appears to be only the tip of the iceberg. To proceed, a bigger class of black hole solutions to the Einstein equations needs to be introduced.

The Kerr-Newman three-parameter family in Boyer-Lindquist coordinates, which are the generalization of the Schwarzschild coordinates, is described by the metric

$$ds^2 = \frac{\Delta - a^2 \sin^2 \theta}{\rho^2} dt^2 + 2a \sin^2 \theta \frac{r^2 + a^2 - \Delta}{\rho^2} dt d\phi - \left( \frac{r^2 + a^2}{\rho^2} - \Delta \sin^2 \theta \right) \sin^2 \theta d\phi^2 - \frac{\rho^2}{\Delta} dr^2 - \rho^2 d\theta^2,$$

(1.125)

where

$$\rho^2 = r^2 + a^2 \cos^2 \theta \quad \text{(1.126)}$$
$$\Delta = r^2 - 2GMr + a^2 + Ge^2 \quad \text{(1.127)}$$

The three parameters are $M, a$ and $e$. The meaning of $a$ is given by

$$a = \frac{J}{GM},$$

(1.128)

with $J$ the total angular momentum of the black hole. For $e$ one has the expression

$$e = \sqrt{Q^2 + P^2}$$

(1.129)

where $Q$ is the electric charge and $P$ a hypothetical magnetic monopole charge. The Maxwell 1-form of the Kerr-Newman solution is

$$A = \frac{Qr(dt - a \sin^2 \theta d\phi) - P \cos \theta[adt - (r^2 + a^2)d\phi]}{\rho^2}.$$  

(1.130)

The metric coefficients are independent of $t$ and $\phi$, so the Kerr-Newman family represents a time-independent and axially symmetric solution to the Einstein equations. When $a = 0$, the Kerr-Newman solution reduces to the Reissner-Nordström solution. Hence, with $a = 0$ and $Q = 0$ one again finds the Schwarzschild solution. The two-parameter family which follows from (1.125) by putting $Q = 0$ is called the Kerr solution. Taking $\phi \to -\phi$ effectively changes the sign of $a$, so one may choose $a \geq 0$ without loss of generality. The metric also has the discrete symmetry

$$t \to -t, \quad \phi \to -\phi.$$  

(1.131)

A Kerr-Newman geometry has a horizon, and therefore describes a black hole, if and only if $G^2 M^2 \geq G^2 Q^2 + a^2$. It seems likely that in any collapsing body which violates this contraint, centrifugal forces and/or electrostatic repulsion will halt the collapse before it reaches a size $\sim GM$. When $G^2 M^2 = G^2 Q^2 + a^2$, the solution is called an extreme Kerr-Newman geometry. The horizon is located at $r_+ = GM + \sqrt{G^2 M^2 - G^2 Q^2 - a^2})$. By looking at the electric and magnetic fields surrounding the Kerr-Newman black hole, one sees that it has an electric charge
$Q$ and a magnetic dipole moment given by $M \equiv Qa$. This means that the gyromagnetic ratio $\gamma = Q/M$, just as for an electron.

In 1970, Carter proved another uniqueness theorem concerning uncharged stationary black holes [17]:

An uncharged, stationary black hole is a member of the two-parameter Kerr family. The parameters are the mass $M$ and the angular momentum $J$. In other words, the external gravitational field of the black hole is uniquely determined by its mass and its angular momentum.

The key difference of this theorem with Birkhoff’s theorem is that it only applies to black holes. The Kerr metric is important astrophysically because it is a good approximation to the metric of a rotating star at large distances where all multipole moments except $l = 0$ and $l = 1$ are unimportant. The only known solution of Einstein’s equations for which Kerr is exact for $r > r_+$ is when $T^{\mu\nu} = 0$, i.e. the Kerr black hole itself. So it has not been matched to any known non-vacuum solution that could represent the interior of a star, in contrast to the Schwarzschild solution which is guaranteed by Birkhoff’s theorem to be the exact exterior spacetime that matches on to the interior solution for any spherically symmetric star. That there is a thing as Carter’s theorem for black holes but not for stars again emphasises the special nature of black holes.

The importance of these uniqueness theorems must not be underestimated. Since stationarity is equivalent with equilibrium, the final state of gravitational collapse is expected to be a stationary spacetime. Carter’s theorem says that if the collapse is to an uncharged black hole then this spacetime is uniquely determined by its mass and angular momentum. This gives extra information about the nature of gravitational collapse to a black hole. Because in contrast to the case with spherical symmetry, where the geometry outside a gravitational collapse is Schwarzschild in character at all stages of the collapse, in the Kerr-Newman case the geometry outside the collapsing body initially departs from Kerr-Newman character. Only well after the collapse occurs, i.e. in the asymptotic future, and in the region at and outside the horizon, is the Kerr-Newman geometry a faithful description of a black hole. Thus, all multipole moments of the gravitational field are radiated away during the collapse to a black hole, except the monopole and dipole moments which can’t be radiated away because the graviton has spin 2.

In other words, the uniqueness theorems come very close to proving that the external gravitational and electromagnetic fields of a black hole that has settled down to its final state are determined uniquely by the hole’s mass $M$, charge $Q$ and intrinsic angular momentum $J$. Thus, a collapse ends with a Kerr-Newman black hole. Other ‘quantum numbers’ of the particles that went in the black hole like baryon number, lepton number, strangeness, etc. have no place in the external observer’s description of a black hole. This is what is meant by the expression that a black hole ‘has no hair’.

A natural question to be raised now is, what is so special about mass, angular momentum and electric charge? The answer is that they are all conserved quantities subject to a Gauss type law. Thus, one can determine these properties of a black hole by measurements from afar. Obviously, this reasoning has to be completed by including magnetic monopole charge as
a fourth parameter because it also is conserved in Einstein-Maxwell theory, it also submits to a Gauss type law. In the updated no-hair conjecture, the forbidden 'hair' is any field not of gravitational or electromagnetic nature associated with a black hole.

1.8.2 Classical fields

The above statements of the theorems are all somewhat heuristic. Each theorem and its proof are based on the techniques of global geometry [1, 6] and make several highly technical assumptions about the global properties of spacetime. These assumption seem physically reasonable and innocuous, but they might not be, resulting in a no-hair conjecture rather then a no-hair theorem. Nevertheless, there are several exact results which all support the no-hair conjecture, some of which will be discussed here.

The no-scalar hair theorem can be proven exactly in a rather short and simple manner [18]. Consider the action for a classical, real, massive and minimally coupled (see chapter 2) scalar field

$$S = \frac{1}{2} \int d^4x \left( -g^{\mu\nu} \partial_\mu \psi \partial_\nu \psi + V(\psi^2) \right),$$

resulting in the field equation

$$\nabla^\mu \partial_\mu \psi - \psi V(\psi^2) = 0.$$  \hspace{1cm} (1.133)

Because we are looking for a solution in a stationary black hole exterior, assume that the configuration is asymptotically flat and stationary: $\partial \psi / \partial x^0 = 0$, where $x^0$ is a time-like variable in the black hole exterior. Multiplying the field equation (1.133) with $\psi$ and integrating over the black hole exterior $V$ at a given $x^0$ gives after integration by parts

$$- \int_V d^3x \left( -g^{ab} \partial_a \psi \partial_b \psi + \psi^2 V'(\psi^2) \right) + \oint_{\partial V} \partial V d\Sigma_{\mu} \psi \partial_\mu \psi,$$

where $d\Sigma_{\mu}$ is the 2D element of the boundary hypersurface $\partial V$ and $V' = \partial V / \partial \psi^2$. The indices $a$ and $b$ run over the space coordinates only, so the restricted metric $g^{ab}$ is positive definite in the black hole exterior. Now suppose that the boundary $\partial V$ is taken as a large sphere at infinity, which has topology $\mathbb{R} \otimes S^2$, together with a surface close to the horizon $H$, also with topology $\mathbb{R} \otimes S^2$. Then so long as $\psi$ decays as $1/r$ or faster at large distances, which will be true for static solutions of (1.133), infinity’s contribution to the boundary integral vanishes. At the inner boundary, Schwarz’s inequality can be used to state that at every point

$$|\psi \partial^\mu \psi | d\Sigma_{|\mu|} \leq (\psi^2 \partial^\mu \psi \partial_\mu d\Sigma_{\nu} d\Sigma_{\mu})^{1/2}.$$  \hspace{1cm} (1.135)

As the boundary is pushed to the horizon, a null surface, $d\Sigma_{\nu} d\Sigma_{\mu}$ must necessarily tend to zero. Thus the inner boundary term will also vanish unless $\psi^2 \partial^\mu \psi \partial_\mu$ blows up at $H$. But this is unacceptable for a black hole since nothing strange should happen at the horizon if one wants to preserve the equivalence principle. To be more precise, with $T^{\mu\nu}$ the energy-momentum tensor of the field, the physically relevant scalars $T_{\mu\nu}T^{\mu\nu}$ and $T^\mu_\mu$ should remain bounded at $H$. This implies that $\partial^\mu \psi \partial_\mu \psi$ and $V$ should remain bounded. If $V$ diverges for large arguments, then $\psi$ has to remain finite. This implies that $\psi^2 \partial^\mu \psi \partial_\mu$ is bounded on $H$. But even if $V$ does not diverge at large arguments, so that $\psi$ is allowed to diverge, this will almost certainly cause
\[ \partial_\mu \psi \partial_\mu \] to diverge. So one can conclude that the boundary term vanishes.

Thus for a generic \( V \) if follows that the generalized 4D integral in (1.134) must itself vanish. In the case that \( V'(\psi^2) \) is non-negative everywhere and vanishes only at some discrete values \( \psi_j \), then it is clear that the field must be constant everywhere outside the black hole, taking on one of the values \( \{0, \psi_j\} \). The scalar field is thus trivial, either vanishing or taking a constant value as dictated by spontaneous symmetry breaking without the black hole. In particular, the theorem works for the Klein-Gordon field for which \( V'(\psi^2) = m^2 \), where \( m \) is the field’s mass. In that case \( \psi = 0 \) outside the black hole. Obviously, this result supports the no-hair conjecture by ruling out that black holes could have parameters associated with a scalar field.

It is remarkable that nowhere is made use of the gravitational field equations. The statement that there are no black holes with scalar hair is thus just as true in other metric theories of gravity. Another advantage is that this method can be easily extended to exclude massive vector field hair. A shortcoming is that it does not rule out hair in the form of a Higgs field with a Mexican hat potential because then \( V' \) becomes negative for certain \( \psi \) values. But it has been shown by other techniques that a black hole also has no Higgs-hair [19].

When the massive vector field’s mass is removed, gauge invariance sets in and the fundamental field, which is basically the covariant time component of the vector field \( A_\mu \), cannot be required to be bounded at the horizon because it is not gauge invariant [18]. No no-hair theorem can be proven: a black hole can have an electromagnetic field as present in the Kerr-Newman family. By the same logic, one can conclude that the gauge invariance of the non-abelian gauge theories should likewise allow one or more of the gauge field components generated by sources in the black hole to escape from it. Thus gauge fields around a black hole may be possible in every gauge theory. Explicit solutions were found [20, 21]. But these solutions should not be taken into account because they are highly unstable, in fact, almost all known hairy black hole solutions in 3+1 general relativity are known to be unstable [22–25]. The only one which is certified to be stable [26], at least in linearized theory, is the Skyrmion hair black hole [27]. It differs from the Schwarzschild one in that it involves a parameter with properties of a topological winding number. This is not an additive quantity among several black holes, so that the Skyrmion black hole may not represent a true exception to the no-hair theorem.

Based upon the no-hair conjecture, it was already known in the 1970’s what the consequences are of the the loss of baryon and lepton number down a black hole. Hartle and Teitelboim [28–31] have shown that a black hole cannot exert any weak-interaction forces caused by the leptons which have gone down it. A similar analysis to show the absence of strong-interaction forces from baryons that have gone down the hole has been done by Bekenstein and Teitelboim [32–34]. This implies the non-electromagnetic force between two baryons or leptons resulting from exchange of various force carriers would vanish if one of the particles was allowed to approach a black hole horizon.
1.8.3 The electric Meissner effect and equilibrium

As mentioned above, an outside observer sees no object ever really arriving at the horizon. So naively, this could lead one to think that when throwing a charged particle in a black hole we will be able to tell forever where the particle entered because we could deduce its position by measuring the electromagnetic field. This appears not to be true. In an article by Cohen and Wald [35] the electrostatic field of a point charge at rest in Schwarzschild space is derived. The solution is then used to study the problem of a point charge being slowly lowered into a nonrotating black hole. It is found that the electric field of the charge remains well behaved as the charge passes the horizon and that all the multipole moments except the monopole fade away. So a Reissner-Nordstrom black hole is produced. This apparent paradox is resolved by the fact that the curvature of spacetime deforms the electric and magnetic fields produced by the charge. In [36] it is shown how there appears to be some sort of 'electric Meissner effect', bending the field lines of the charge around the black hole as shown on figure 1.16. We will return to this feature in chapter 3 in the context of the membrane paradigm.

![Figure 1.16: Electric field lines of a point charge bending around a black hole.](image)

From all this, one can conclude that at the classical level, black holes effectively destroy information. All information about the initial state of the matter which collapsed to a black hole is lost once the final stationary Kerr-Newman state is reached. This makes the process of black hole formation highly irreversible. The uniqueness of a final stationary state has a strong resemblance with the concept of thermal equilibrium in statistical physics. In both cases different initial states of the system evolve towards the same stationary final state characterized by a limited amount of macroscopic quantities. This analogy was the basis for the introduction of a notion of entropy for black holes. It had to account for this irreversibility and give meaning to the laws of black hole mechanics which will be discussed in the final section of this chapter.

1.9 Radial null geodesics in black hole spacetimes

For the derivation of the Hawking spectrum in the next chapter it is necessary to go into further detail about radial null geodesics. In particular, a formula relating the constant value of the
null coordinate $u$ for an outgoing geodesic to the constant value of the null coordinate $v$ for the corresponding ingoing geodesic needs to be derived, and this for geodesics close to the last geodesic that can escape to infinity.

### 1.9.1 The Schwarzschild black hole

Consider the geodesics in the plane taken at $\theta = \pi/2$, which can be done without loss of generality. The Schwarzschild metric is independent of $t$ and of $\phi$ and, as seen in section 1.1, this implies the existence of two Killing vector fields given by $k = \partial/\partial t$ and $m = \partial/\partial \phi$ and two corresponding conserved quantities

$$E = k \cdot p = g_{\mu \nu} k^\mu p^\nu = g_{tt} \frac{dt}{d\lambda} = \left(1 - \frac{2MG}{r}\right) \frac{dt}{d\lambda}$$  \hspace{1cm} (1.136)

$$L = m \cdot p = g_{\mu \nu} m^\mu p^\nu = g_{\phi \phi} \frac{d\phi}{d\lambda} = r^2 \frac{d\phi}{d\lambda},$$  \hspace{1cm} (1.137)

where (1.20) and (1.21) were used with a rescaling of the conserved quantities such that the mass can be left out of (1.21). Instead of the proper time $\tau$ a general affine parameter $\lambda$ is used along the geodesic.

Requiring the geodesic to be null implies

$$0 = \frac{dx^\mu}{d\lambda} \frac{dx_\mu}{d\lambda} = g_{\mu \nu} \frac{dx^\mu}{d\lambda} \frac{dx^\nu}{d\lambda} = \left(1 - \frac{2MG}{r}\right) \left(\frac{dr}{d\lambda}\right)^2 - \left(1 - \frac{2MG}{r}\right)^{-1} \left(\frac{dr}{d\lambda}\right)^2 - r^2 \left(\frac{d\phi}{d\lambda}\right)^2.$$  \hspace{1cm} (1.138)

So by means of (1.136) and (1.137) it follows that along a null geodesic

$$E^2 = \left(\frac{dr}{d\lambda}\right)^2 + \frac{L^2}{r^2} \left(1 - \frac{2MG}{r}\right).$$  \hspace{1cm} (1.139)

The radial geodesics are those with $L = 0$, so

$$\frac{dr}{d\lambda} = \pm E.$$  \hspace{1cm} (1.140)

The upper sign corresponds to outgoing geodesics (for $r > 2MG$), and the lower sign to incoming geodesics. From (1.140) and (1.136) one has

$$0 = \frac{dt}{d\lambda} \mp \left(1 - \frac{2MG}{r}\right) \frac{dr}{d\lambda} = \frac{d}{d\lambda} \left(t \mp r^*\right)$$  \hspace{1cm} (1.141)

where $r^*$ is again the tortoise coordinate defined by (1.51). This implies again that the null coordinate $u = t - r^*$ is constant along outgoing radial null geodesics and $v = t + r^*$ is constant along incoming radial null geodesics.
Now let \( C \) be an incoming radial null geodesic defined by \( v = v_1 \) for some \( v_1 \) that passes through the event horizon of the Schwarzschild black hole. Let \( \lambda \) be an affine parameter along this geodesic. The null coordinate \( u \) is given along \( C \) by some function \( u(\lambda) \). It is the form of this function just outside the event horizon that will determine the spectrum of the particles created by the black hole (see chapter 2). Along the null geodesic \( C \) we have

\[
\frac{du}{d\lambda} = \frac{dt}{d\lambda} - \frac{dr^*}{d\lambda},
\]

(1.142)

where \( dt/d\lambda \) is given by (1.136). It follows from (1.140), with the minus sign, that

\[
\frac{dr^*}{d\lambda} = \frac{dr^*}{dr} \frac{dr}{d\lambda} = - \left(1 - \frac{2GM}{r}\right)^{-1} E
\]

(1.143)

and thus

\[
\frac{du}{d\lambda} = 2 \left(1 - \frac{2GM}{r}\right)^{-1} E.
\]

(1.144)

Integrating (1.140) along \( C \) gives

\[
r - 2GM = -E\lambda,
\]

(1.145)

where the integration constant is chosen such that \( \lambda \) is zero at the event horizon. For \( r > 2GM \), the affine parameter \( \lambda \) is negative. With this, we can write

\[
\left(1 - \frac{2GM}{r}\right)^{-1} = 1 - \frac{2GM}{E\lambda},
\]

(1.146)

and

\[
\frac{du}{d\lambda} = 2E - \frac{4GM}{\lambda}.
\]

(1.147)

Therefore, along the incoming null geodesic \( C \),

\[
u = 2E\lambda - 4GM \ln(\lambda/K_1),
\]

(1.148)

where \( K_1 \) is a negative constant. Far from the event horizon, \( u \approx 2E\lambda \), while near the event horizon

\[
u \approx -4GM \ln(\lambda/K_1).
\]

(1.149)

The null coordinate \( u \) is \(-\infty\) at past null infinity \( \mathcal{I}^- \) and \(+\infty\) at the event horizon.

Now consider the situation as depicted on figure 1.17. The null ray with constant incoming null coordinate \( v \) originates on \( \mathcal{I}^- \), passes though the center of the collapsing body and becomes the null ray having constant outgoing null coordinate \( u = u(v) \). The incoming ray with \( v = v_0 \) is the last one that passes through the center of the body and reaches \( \mathcal{I}^+ \). Incoming null rays with \( v > v_0 \) enter the black hole and run into the singularity.

The affine parameter \( \lambda \) along all radially incoming null geodesics that pass through the horizon can chosen such that (1.149) relates \( u \) to \( \lambda \) near the event horizon. Then the affine parameter distance between the outgoing rays \( u(v_0) \) and \( u(v) \) is constant along the entire length of the geodesics, as measured by the change of the affine parameter \( \lambda \) along any incoming null
ray intersecting the two outgoing null rays. Moving backwards along these outgoing geodesics through the collapsing body, they become the incoming geodesics that originate on $\mathcal{I}^{-}$ at $v_0$ and $v$ respectively. The affine separation along the null direction between these two geodesics can be chosen to remain constant along their entire length, as they go from $\mathcal{I}^{-}$ to $\mathcal{I}^{+}$. Therefore, the affine separation between $v$ and $v_0$ at $\mathcal{I}^{-}$ is the same as that between $u(v)$ and $u(v_0)$ at $\mathcal{I}^{+}$. Because $\lambda = 0$ at the horizon, the affine separation between $u(v)$ and $u(v_0)$ at $\mathcal{I}^{+}$ has the value $\lambda$ that satisfies (1.149) with $u$ having the value $u(v)$.

Because the coordinate $v$ is itself an affine parameter along $\mathcal{I}^{-}$, $v - v_0$ must be related to the affine separation $\lambda$ between $u(v)$ and $u(v_0)$ on $\mathcal{I}^{+}$ by

$$v_0 - v = K_2 \lambda,$$

(1.150)

where $K_2$ is a negative constant. Hence

$$u(v) = -4MG \ln(\lambda/K_1)$$
$$= -4MG \ln \left( \frac{v_0 - v}{K_1K_2} \right)$$
$$= -4MG \ln \left( \frac{v_0 - v}{K} \right),$$

(1.151)

with $K$ a positive constant. This is the relation that will determine the spectrum of the created particles when considering quantum fields in a black hole spacetime in chapter 2.

1.9.2 The Kerr black hole

In this section the Kerr metric, which is the Kerr-Newman metric (1.125) with $Q \equiv 0$, is considered for $GM > a$. The Kerr metric is singular at $r = r_{\pm}$, the zeros of $\Delta$. In section 1.8.1 it
was already mentioned that the horizon is located at $r = r_+ = GM + \sqrt{G^2M^2 - a^2}$. The singularities at $r = r_\pm$ are coordinate singularities, they are an insufficiency of the Boyer-Lindquist coordinates. To see this one can introduce null coordinates in analogy to the Schwarzschild case

$$u = t - r^*$$

$$v = t + r^*,$$  \hspace{1cm}(1.152)

where the tortoise coordinate is now defined by

$$\frac{dr^*}{dr} = \frac{r^2 + a^2}{\Delta},$$  \hspace{1cm}(1.154)

which can be solved explicitly as

$$r^* = r + 2GM \frac{r_+}{r_+ - r_-} \ln|r - r_+| - 2GM \frac{r_-}{r_+ - r_-} \ln|r - r_-|$$  \hspace{1cm}(1.155)

$$= r + GM \left( 1 + \frac{GM}{\sqrt{G^2M^2 - a^2}} \right) \ln|r - r_+|$$

$$- GM \left( 1 - \frac{GM}{\sqrt{G^2M^2 - a^2}} \right) \ln|r - r_-|.$$  \hspace{1cm}(1.156)

But in the Kerr case, also a new angular coordinate $\chi$ has to be defined

$$d\chi = d\phi - \frac{a}{\Delta} dr.$$  \hspace{1cm}(1.157)

Making the transformation from Boyer-Lindquist to Kerr coordinates $(v, r, \theta, \chi)$, the analogs of the Eddington-Finkelstein coordinates for a Schwarzschild black hole, the Kerr metric transforms into

$$ds^2 = \left( \frac{\Delta - a^2 \sin^2 \theta}{\rho^2} \right) dv^2 - 2dv dr + 2a \sin^2 \theta \left( \frac{r^2 + a^2 - \Delta}{\rho^2} \right) dv d\chi$$

$$- 2a \sin^2 \theta d\chi dr - \left( \frac{(r^2 + a^2)^2 - \Delta a^2 \sin^2 \theta}{\rho^2} \right) \sin^2 \theta d\chi^2 - \rho d\theta^2,$$  \hspace{1cm}(1.158)

which is well-behaved at $\Delta = 0$.

As mentioned in section 1.8.1, the Kerr metric describes a rotating black hole with angular momentum $J = aGM$. This can be seen explicitly as follows. Let $\mathcal{N}_\pm$ be the hypersurfaces at $r = r_\pm$. The vector fields normal to $\mathcal{N}_\pm$ are given by (1.30):

$$l_\pm = -f_\pm g^{\mu \nu} \partial_\nu |\mathcal{N}_\pm \partial_\mu$$

$$= - \left( \frac{r_\pm^2 + a^2}{r_\pm^2 + a^2 \cos^2 \theta} \right) f_\pm \left( \frac{\partial}{\partial v} + \frac{a}{r_\pm^2 + a^2} \frac{\partial}{\partial \chi} \right).$$  \hspace{1cm}(1.159)

They have the property that

$$l_\pm^2 \propto \left( g_{vv} + \frac{2a}{r_\pm^2 + a^2} g_{v\chi} + \frac{a^2}{(r_\pm^2 + a^2)^2} g_{\chi\chi} \right) |\mathcal{N}_\pm = 0.$$  \hspace{1cm}(1.161)
So \( N_\pm \) are null hypersurfaces. It then follows from (1.160) that they are Killing horizons of the Killing vector fields
\[
\xi_\pm = \frac{\partial}{\partial v} + \left( \frac{a}{r_\pm^2 + a^2} \right) \frac{\partial}{\partial \chi},
\]
(1.162)
because the metric coefficients are independent of \( v \) and \( \chi \). Using (1.153), (1.157) and (1.154) this can be written in the original Boyer-Lindquist coordinates as
\[
\xi_\pm \mid_{N_\pm} = \left( \frac{\partial}{\partial t} + \frac{1}{r_\pm^2 + a^2} \frac{\partial}{\partial \phi} \right) + \frac{a r_\pm^2 + a^2 + a^2}{r_\pm^2 + a^2} \frac{\partial}{\partial t} = k + \left( \frac{a}{r_\pm^2 + a^2} \right) m.
\]
(1.163)
As explained in section 1.6 one can find the surface gravities \( \kappa_\pm \) by computing \( \xi_\nu \nabla_\nu \xi_\mu \). A lengthy calculation in appendix A gives
\[
\kappa_\pm = \frac{r_\pm - r_\mp}{2(r_\pm^2 + a^2)}.
\]
(1.164)
Thus, we have found that the event horizon \( r = r_+ \) of a Kerr black hole is a Killing horizon of \( \xi = k + \Omega_H m \), with
\[
\Omega_H = \frac{a}{r_+^2 + a^2} = \frac{J}{2GM(G^2M^2 + \sqrt{G^4M^4 - J^2})}
\]
(1.165)
and surface gravity \( \kappa = \kappa_+ \).
In coordinates for which \( k = \partial/\partial t \) and \( m = \partial/\partial \phi \), the definition of \( \xi \) implies
\[
\xi^\mu \partial_\mu (\phi - \Omega_H t) = 0,
\]
(1.166)
so \( \phi = \Omega_H t + \) constant on orbits of \( \xi \), whereas \( \phi \) is constant on orbits of \( k \). Note that \( k \) is unique. Consider
\[
(k + am)^2 = g_{tt} + 2\alpha g_{t\phi} + \alpha^2 g_{\phi\phi},
\]
(1.167)
as long as \( g_{t\phi} \) is finite and \( g_{\phi\phi} \sim r^2 \) as \( r \to +\infty \), one has \( (k + am)^2 \sim \alpha^2 r^2 > 0 \) as \( r \to +\infty \). So there can be only one Killing vector \( k \) that is time-like at infinity and normalized, meaning \( k^2 \to +1 \) as \( r \to +\infty \).

We can conclude that objects on orbits of \( \xi \) rotate with angular velocity \( \Omega_H \) relative to static particles, which are those on orbits of \( k \), and hence relative to stationary observers at infinity. Since the null geodesic generators of the horizon follow orbits of \( \xi \), the black hole is rotating with angular velocity \( \Omega_H \).

We now return to Boyer-Lindquist coordinates to find the radial null geodesics. Just like in the Schwarzschild case, the two Killing vector fields \( \partial/\partial t \) and \( \partial/\partial \phi \) imply the existence of two
conserved quantities

\[ E = \left( 1 - \frac{2GMr}{\rho^2} \right) \frac{dt}{d\lambda} + \frac{2aGMr \sin^2 \theta}{\rho^2} \frac{d\phi}{d\lambda} \]  
\[ L = -\frac{2aGMr \sin^2 \theta}{\rho^2} \frac{dt}{d\lambda} + \left( r^2 + a^2 + \frac{2a^2GMr}{\rho^2} \sin^2 \theta \right) \sin^2 \theta \frac{d\phi}{d\lambda}. \]  

(1.168)

(1.169)

But for the Kerr metric, there is an additional conserved quantity, the Carter constant \( K_C \), which is given for null geodesics by [37]

\[ K_C = \frac{1}{\Delta} \left[ \Delta \frac{dt}{d\lambda} - (a \Delta \sin^2 \theta) \frac{d\phi}{d\lambda} \right]^2 - \frac{\rho^4}{\Delta} \left( \frac{dr}{d\lambda} \right)^2 \]  
\[ K_C = \left[ a \sin \theta \frac{dt}{d\lambda} - (r^2 + a^2) \sin \theta \frac{d\phi}{d\lambda} \right]^2 + \rho^4 \left( \frac{d\theta}{d\lambda} \right)^2 \]  

(1.170)

(1.171)

The conserved quantities given by (1.168), (1.169), (1.170) and (1.171), together with the demand that the geodesics under consideration are null, leads to the following equations chandrasekhar

\[ \rho^4 \left( \frac{dr}{d\lambda} \right)^2 = [(r^2 + a^2)E - aL]^2 - K_C \Delta \]  
\[ \rho^4 \left( \frac{d\theta}{d\lambda} \right)^2 = - \left( aE \sin \theta - \frac{L}{\sin \theta} \right)^2 + K_C \]  
\[ \rho^2 \frac{dt}{d\lambda} = \frac{1}{\Delta} \left\{ [(r^2 + a^2)^2 - \Delta a^2 \sin^2 \theta]E - 2aGMrL \right\} \]  
\[ \rho^2 \frac{d\phi}{d\lambda} = \frac{1}{\Delta} \left[ 2aGMrE + (r^2 - 2GMr) \frac{L}{\sin^2 \theta} \right] \]  

(1.172)

(1.173)

(1.174)

(1.175)

Radial null geodesics move in planes of constant \( \theta \). From (1.173) it is clear that a solution with constant \( \theta = \theta_0 \) is only possible if

\[ K_C = 0 \]  
\[ L = aE \sin^2 \theta_0. \]

(1.176)

(1.177)

With this (1.172), (1.174) and (1.175) become

\[ \frac{dr}{d\lambda} = \pm E \]  
\[ \frac{dt}{d\lambda} = \frac{r^2 + a^2}{\Delta} E \]  
\[ \frac{d\phi}{d\lambda} = \frac{aE}{\Delta}. \]  

(1.178)

(1.179)

(1.180)

In simplifying (1.174), the identity

\[ (r^2 + a^2)^2 - \Delta a^2 \sin^2 \theta = \rho^2 (r^2 + a^2) + 2a^2GMr \sin^2 \theta \]  

(1.181)

is used. For \( a = 0 \), these equations of motion reduce to the ones for radial null geodesics in Schwarzschild spacetime, as found in the previous section. The geodesics specified by (1.178), (1.179) and (1.180) are called the principal null congruence in the Kerr spacetime. A congruence
is a family of curves such that precisely one curve of the family passes through each spacetime point. As \( r \) approaches its horizon value \( r_+ \) it follows that \( t \to \infty \) and \( \phi \to \infty \). So in Boyer-Lindquist coordinates an incoming object takes an infinite coordinate time to reach the event horizon, but it also winds around the \( z \)-axis an infinite amount of times.

The spectrum of the particles created by a Kerr black hole will be determined by the function \( u(\lambda) \), with \( u \) the null coordinate as introduced in (1.154) and \( \lambda \) an affine parameter along an incoming geodesic \( C \) of the principal null congruence. Along such a geodesic

\[
\frac{du}{d\lambda} = \frac{dt}{d\lambda} - \frac{dr^*}{dr} \frac{dr}{d\lambda} = \frac{2E r^2 + a^2}{\Delta},
\]

where (1.178), (1.179) and (1.154) were used. (1.178) can be integrated to give

\[
r - r_+ = -E\lambda,
\]

where the integration constant was chosen such that \( \lambda = 0 \) at \( r = r_+ \). The affine parameter \( \lambda \) is negative for \( r > r_+ \). Combining (1.182) and (1.183) one gets

\[
\frac{du}{d(E\lambda)} = \frac{2(\lambda_+ - E\lambda)^2 + a^2}{E\lambda[E\lambda - (r_+ - r_-)]},
\]

which can be solved to give along the incoming null geodesic \( C \)

\[
u = 2E\lambda - \frac{1}{\kappa_+} \ln \left( \frac{E\lambda}{K_1} \right) + \frac{1}{\kappa_-} \ln \left( \frac{E\lambda - (r_+ - r_-)}{K_1'} \right),
\]

where \( \kappa_+ \) and \( \kappa_- \) are given by (1.164). When \( a \to 0 \), \( r_- \to 0 \) and \( r_+ \to 2GM \). Then \( \kappa_+ \to 1/4GM \) and \( \kappa_- \to +\infty \), so the results of the Schwarzschild spacetime are recovered.

Far outside the event horizon of the Kerr black hole, \( u \approx 2E\lambda \) with \( u \) approaching \( -\infty \) at \( I^- \). As \( r \to r_+ \), then \( u \to +\infty \) with

\[
u \approx -\frac{1}{\kappa_+} \ln \left( \frac{E\lambda}{K_1} \right).
\]

In the spacetime of a rotating body that undergoes gravitational collapse, consider two outgoing null geodesics \( C_1 \) and \( C_2 \) that at late times belong to the principal null congruence. Along these geodesics, \( u \) is constant. Let \( C_\infty \) and \( C_2 \) intersect the incoming null geodesic \( C \) where its affine parameter \( \lambda \) has values \( \lambda_1 \) and \( \lambda_2 \) respectively. Let these values of \( \lambda \) be negative and of sufficiently small magnitude that (1.186) is valid. The null geodesics \( C_\infty \) and \( C_2 \) originate at \( I^- \) as incoming null geodiscs at \( v = v_1 \) and \( v = v_2 \), respectively. As \( \lambda_1 \to 0^- \), we have \( v_1 \to v_0 \), where \( v_0 \) is the value of \( v \) on the last incoming null geodesic that starts from \( I^- \), passes through the rotating collapsing body and reaches \( I^+ \) on the outgoing null geodesic \( u = u(v) \). Similarly \( C_2 \) originates at a value \( v_2 < v_0 \) and reaches \( I^+ \) at \( u = u(v_2) \). The affine separation along null geodesics between \( C_\infty \) and \( C_2 \) is given by (1.186) with \( u = u(v_2) \). At \( I^- \), \( v \) is an affine parameter. As in
the Schwarzschild spacetime, it holds that
\[ v_0 - v = K_2 \lambda, \]  
(1.187)
where \( K_2 \) is a negative constant, and the subscript on \( v_2 \) is dropped. It follows from (1.186) that
\[ u(v) \approx -\frac{1}{\kappa} \ln \left( \frac{v_0 - v}{K} \right), \]  
(1.188)
where \( K \) is a positive constant. This expression will determine the spectrum of the outgoing particles created by the Kerr black hole.

1.10 Energy extraction in Kerr spacetime

One of the defining properties of a black hole is that nothing, not even light, can escape from behind its horizon. Therefore, it is very surprising that one can nevertheless extract energy out of a rotating black hole. This phenomenon is presented here because it can be seen as some sort of classical analogon to the quantum mechanical process of particle creation by black holes.

1.10.1 The ergosphere of a Kerr black hole

The spacetime of a rotating black hole is asymptotically flat, which means that the metric at spatial infinity is the Minkowski metric. Therefore, the Killing vector field describing time translations for observers at large distances has the simple form
\[ k = \frac{\partial}{\partial t}. \]  
(1.189)
The norm of this vector field is given by
\[ k^2 = g_{tt} = \left( \Delta - \frac{a^2 \sin^2 \theta}{\rho^2} \right) = \left( 1 - \frac{2GMr}{r^2 + a^2 \cos^2 \theta} \right). \]  
(1.190)
Where (1.125) was used. So one sees that however \( k^\mu \) is timelike at infinity, it does not have to be timelike everywhere. In particular, it follows that \( k^\mu \) is timelike provided that
\[ r^2 + a^2 \cos^2 \theta - 2MGr > 0. \]  
(1.191)
For \( M^2 \gg a^2 \) this implies that
\[ r > GM + \sqrt{G^2M^2 - a^2 \cos^2 \theta} \]  
(1.192)
(or \( r < GM - \sqrt{G^2M^2 - a^2 \cos^2 \theta} \), but this is physically not relevant).

The boundary of this region, i.e. the hypersurface given by
\[ r = GM + \sqrt{G^2M^2 - a^2 \cos^2 \theta}, \]  
(1.193)
is called the ergosphere. The ergosphere intersects the event horizon at $\theta = 0, \pi$, but it lies outside the horizon for other values of $\theta$. Thus, $k^\mu$ can become space-like in a region outside the event horizon. This region is called the ergoregion. Because $k^\mu$ is spacelike, an observer in the ergoregion cannot 'stand still' relative to a stationary observer at infinity. In fact, an observer in the ergoregion must rotate relative to infinity in the same direction as the black hole. This is an extreme example of the 'dragging of inertial frames'.

\begin{figure}[h]
\centering
\includegraphics[width=0.7\textwidth]{ergosphere.png}
\caption{The ergosphere of a Kerr black hole}
\end{figure}

1.10.2 The Penrose process

Suppose that a particle approaches a Kerr black hole along a geodesic. If $p$ is its 4-momentum one can identify the constant of motion

$$E = p \cdot k$$

as its energy since $E = p^0$ at infinity. Now suppose that the particle decays into two other particles, one of which falls behind the horizon while the other escapes to infinity.

\begin{figure}[h]
\centering
\includegraphics[width=0.7\textwidth]{penrose.png}
\caption{The Penrose process}
\end{figure}

By conservation of energy one has

$$E_2 = E - E_1.$$ 

(1.195)
Normally $E_1 > 0$ so $E_2 < E$, but in this case

$$E_1 = p_1 \cdot k$$

(1.196)

is not necessarily positive in the ergoregion since $k$ may be space-like there. This means that is is possible for classical particles to have a total negative energy (including rest mass energy) relative to infinity. Thus, if the decay takes place in the ergoregion one may have $E_2 > E$, so energy has been extracted from the black hole.

The event horizon was shown to be a Killing horizon of the Killing vector field $\xi = k + \Omega_H m$, so for particles passing through the horizon at $r = r_+$ one has

$$p \cdot \xi \geq 0$$

(1.197)

because $\xi$ is future-directed null on the horizon (the horizon is a null surface) and $p$ is future directed time-like or null. It follows that

$$E - \Omega_H L \geq 0$$

(1.198)

where $L = -p \cdot m$ is the component of the particle’s angular momentum in the direction defined by $m$ (only this component is a constant of motion). Thus

$$L \leq \frac{E}{\Omega_H}.$$ \hspace{1cm} (1.199)

If $E$ is negative, as it is for particle 1 in the Penrose process then $L$ is also negative, so the black hole’s angular momentum is reduced. In the end, one has a black hole of mass $M + \delta M$ and angular momentum $J + \delta J$, where $\delta M = E$ and $\delta J = L$, so

$$\delta J \leq \frac{\delta M}{\Omega_H} = \frac{2GM(G^2M^2 + \sqrt{G^4M^4 - J^2})}{J} \delta M,$$

(1.200)

where (1.165) was used. A little algebra then gives

$$0 \leq \left(2G^3M^3 + 2GM \sqrt{G^4M^4 - J^2} - J J \frac{\delta J}{\delta M} \delta M \right) \delta M$$

(1.201)

and this is equivalent with

$$\delta \left(G^2M^2 + \sqrt{G^4M^4 - J^2} \right) \geq 0.$$ \hspace{1cm} (1.202)

The area of the event horizon is

$$A = \int_{r=r_+} d\theta d\phi \sqrt{g_{\theta\theta}g_{\phi\phi}}$$

$$= 4\pi(r_+^2 + a^2)$$

$$= 8\pi GM(GM + \sqrt{G^2M^2 - a^2})$$ \hspace{1cm} (1.203)

This means that energy contraction by the Penrose process is limited by the requirement that $\delta A \geq 0$. In the next section this will be shown to be a special case of the second law of black hole mechanics, which has a striking resemblance with the second law of thermodynamics.
1.10.3 Superradiance

The Penrose process has a close analogue in the scattering of radiation by a Kerr black hole. For simplicity, consider a massless scalar field $\psi$. Its energy-momentum tensor is

$$ T_{\mu\nu} = \partial_\mu \psi \partial_\nu \psi - \frac{1}{2} g_{\mu\nu} (\partial \psi)^2. $$

(1.204)

Since $\nabla_\mu T^\mu_\nu = 0$ one has

$$ \nabla_\mu (T^\mu_\nu k^\nu) = T^\mu_\nu D_\mu k^\nu = 0, $$

(1.205)

so one can consider

$$ j^\mu = -T^\mu_\nu k^\nu = -\partial^\mu \psi k \cdot \partial \psi + \frac{1}{2} k^\mu (\partial \psi)^2 $$

(1.206)

as the future directed ($k \cdot J > 0$) energy-momentum flux 4-vector of $\psi$. Now consider the following region $S$ of spacetime, which has the null hypersurface $\mathcal{N} \subset \mathcal{H}^+$ as one boundary.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{superradiance_region.png}
\caption{A region of spacetime for superradiance}
\end{figure}

Assume that $\partial \psi = 0$ at $i_0^+$ Since $\nabla_\mu j^\mu = 0$ one has

$$ 0 = \int_S d^4x (-g)^{1/2} \nabla_\mu j^\mu = \int_{\partial S} dS_\mu j^\mu $$

$$ = \int_{\Sigma_2} dS_\mu j^\mu - \int_{\Sigma_1} dS_\mu j^\mu - \int_{\mathcal{N}} dS_\mu j^\mu $$

$$ = E_2 - E_1 - \int_{\mathcal{N}} dS_\mu j^\mu $$

(1.207)

where $E_i$ is the energy of the scalar field on the spacelike hypersurface $\Sigma_i$. The energy going through the horizon is therefore

$$ \Delta E = E_1 - E_2 = -\int_{\mathcal{N}} dS_\mu j^\mu $$

$$ = -\int dA dv \xi_\mu j^\mu, $$

(1.208)
where $v$ is the Kerr time coordinate, defined by (1.153). The energy flux lost per unit of Kerr time is therefore

$$P = -\int dA \xi^\mu j^\mu = \int dA (\xi \cdot \partial \psi)(k \cdot \nabla \psi)$$

where (1.206) was used, together with the fact that $\xi \cdot k = 0$ on a Killing horizon $\mathcal{N}$ of $\xi$. This can easily be seen by

$$\xi \cdot k|_\mathcal{N} = \xi^2|_\mathcal{N} - \Omega_H \xi \cdot m|_\mathcal{N}$$

because $\mathcal{N}$ is a null surface and therefore $\xi$, as its Killing vector field, is a null vector on $\mathcal{N}$. Now, $\mathcal{N}$ is a fixed point set of $m$, since $m$ is a Killing vector field (Choose coordinates such that $m = \partial/\partial \phi$. The metric (1.125) is independent of $\phi$, so the position of the horizon is independent of $\phi$.) So $m$ must be tangent to $\mathcal{N}$ or $l \cdot m = 0$ where $l$ is normal to $\mathcal{N}$. But $\xi \propto l$ on $\mathcal{N}$, so $\xi \cdot m|_\mathcal{N} = 0$.

So, explicitly

$$P = \int dA \left( \frac{\partial}{\partial v} \psi + \Omega_H \frac{\partial}{\partial \phi} \psi \right) \left( \frac{\partial \psi}{\partial v} \right).$$

For a wave-mode of angular frequency $\omega$

$$\psi = \psi_0 \cos(\omega v - \mu \phi), \quad \mu \in \mathbb{Z}$$

where $\mu$ is the angular momentum quantum number. The time average power lost across the horizon is

$$P = \frac{1}{2} \psi_0 A \omega (\omega - \mu \Omega)$$

where $A$ is the area of the horizon. $P$ is positive for most values of $\omega$, but for $\omega$ in the range

$$0 < \omega < \mu \Omega_H$$

it is negative, so a wave mode with $\omega, \mu$ satisfying this inequality is amplified by scattering off the rotating black hole. $\mu$ cannot be zero because the amplified field must also take away angular momentum from the black hole. The backreaction on the metric because of the energy loss of the black hole has been neglected in this derivation. Strictly speaking, superradiance is incompatible with stationary black hole spacetimes, but when the superradiant process is sufficiently slow, it is a good approximation.

It is this superradiant phenomenon that made people search for a quantum mechanical process of particle emission by black holes because it has a close resemblance to stimulated emission in atomic physics. And quantum mechanics predicts that where there is stimulated emission, there also is spontaneous emission. That this 'rule' also applies to the black hole context will become clear in the next chapter. The spontaneous emission will have major consequences for black hole mechanics, which we will discuss in the next section.
1.11 Black hole mechanics

In the sections about the no-hair conjecture and superradiance, some first signs appeared that there exists an analogy between black holes and thermodynamics. Here, this analogy will be discussed. The remarkable thing is that the results in this section do not refer in any way to the Kerr-Newman family, although they are expected to represent any stationary black hole. The laws of black hole mechanics are derived in a very general way based on some simple physical assumptions about the spacetime concerning causality and asymptotic behaviour.

First of all, the spacetimes considered are supposed to be asymptotically flat at null infinity. This means that one can conformally map the spacetime \((M, g_{\mu\nu})\) into another spacetime \((M', g'_{\mu\nu})\) such that the image of \(M\) has null boundaries \(I^-\) and \(I^+\). When using really exotic spacetimes, one should check if these null boundaries satisfy the properties to be found in [11]. But for generic spacetimes, and certainly the ones to be considered in this thesis, these properties are fulfilled automatically.

The second assumption concerns spacetimes containing horizons and states that the part of the spacetime on the outside of the future event horizon should be a regular predictable spacetime. The notion of a predictable spacetime will be explained below. First, we note that the assumption forbids the existence of 'naked singularities', i.e. singularities that are visible to and have influence upon observers at large distances. It is widely believed that no such naked singularity can occur in a physically realistic gravitational collapse. The conjecture that no naked singularities occur is known as the cosmic censorship hypothesis and can be formulated in a relatively precise manner as follows:

**Cosmic censorship hypothesis**  
Consider asymptotically flat initial data which are physically achievable on a spacelike hypersurface for a solution of Einstein’s equations with 'suitable' matter. Then the maximal Cauchy development of these data (i.e. the largest spacetime uniquely determined by these data and Einstein’s equations) is asymptotically flat at null infinity.

The term 'suitable' appearing in this formulation of the cosmic censorship hypothesis requires some further explanation. Two necessary conditions on matter for it to be 'suitable' are that it be governed by deterministic (i.e. hyperbolic) differential equations and that it have locally positive energy density. The latter meaning that its energy-momentum tensor \(T^\mu\nu\) satisfies the dominant energy condition, i.e. \(T^{\mu\nu}k_\mu k_\nu \geq 0\) for any future directed time-like vector field \(k\), and for every future directed causal vector field \(l\) (time-like or null), the vector field \(-T^{\mu\nu}l_\nu\) is also a future directed causal vector field, implying that matter and energy can not be observed to flow faster than the speed of light. An additional requirement on matter fields for them to be 'suitable' is that when their differential equations are evolved on a fixed, non-singular and globally hyperbolic spacetime (e.g. Minkowski spacetime), one always obtains globally non-singular solutions. Consequently, any singularities occurring in the Einstein-matter system necessarily would be attributable to gravitational effects.

Now let us come back to the notion of predictability. Let \((M, g_{\mu\nu})\) be an asymptotically flat solution of Einstein’s equations with 'suitable' matter, which contains an asymptotically flat slice
\(\Sigma\) with compact interior region, and is such that \(M\) contains the maximal Cauchy development of data on \(\Sigma\). Then, by means of the cosmic censorship hypothesis, \(M\) will be asymptotically flat at \(I^+\), and the domain of dependence \(D(\Sigma)\) of \(\Sigma\) in \(M\) will include all events to the future of \(\Sigma\) which are visible from infinity, i.e. it will include \(I^+(\Sigma) \cap I^-(I^+)\), where \(I^\pm(A)\) stands for the future or past development of data on \(A\). This means that phenomena in the region exterior to any black hole that may form is predictable from \(\Sigma\). It does not follow, even with the above formulation of the cosmic censorship hypothesis, that any events in the black hole need to be contained in \(D(\Sigma)\). However, if the future event horizon \(H^+\) is also contained in \(D(\Sigma)\), then the black hole is said to be predictable.

1.11.1 The area theorem

Consider a null surface \(N\) and let \(\lambda\) be an affine parameter of the null geodesic generators of \(N\) and denote their tangents with respect to this parametrization by \(k^\mu\). Take \(\alpha\) to be such a null geodesic generator and let \(p \in \alpha\). The expansion \(\theta\) of the null geodesic generators of \(N\) at a point \(p\) is defined by \(\theta = \nabla_\mu k^\mu\). It is the trace of the quantity \(B^{\mu\nu} = \nabla_\mu k^\nu\), which measures geodesic deviation. Now consider an infinitesimal cross-sectional area element of area \(A\) of a bundle of geodesics at \(p\) and Lie transport this area element along the null geodesic generators of \(N\). A detailed analysis based upon geodesic congruences (see [12]) shows that

\[
\frac{dA}{d\lambda} = \theta A, \tag{1.215}
\]

so \(\theta\) measures the local rate of change of cross-sectional area as one moves up the geodesics. The geodesic deviation equation governs the rate of change of \(\theta\). Using this, and the fact that we are working with generators of a null surface \(N\), one obtains the Raychaudhuri equation [11]

\[
\frac{d\theta}{d\lambda} = -\frac{1}{2} \theta^2 - \sigma_{\mu\nu} \sigma^{\mu\nu} - R_{\mu\nu} k^\mu k^\nu, \tag{1.216}
\]

where \(\sigma\) denotes the shear of the geodesics and \(R_{\mu\nu}\) is the Ricci tensor of the spacetime. Now, by means of the Einstein equations with the energy-momentum tensor satisfying the null energy condition \(T_{\mu\nu} k^\mu k^\nu \geq 0\) it follows that \(R_{\mu\nu} k^\mu k^\nu \geq 0\), so

\[
\frac{d\theta}{d\lambda} \leq -\frac{1}{2} \theta^2. \tag{1.217}
\]

It now immediately follows that

\[
\frac{1}{\theta(\lambda)} \geq \frac{1}{\theta_0} + \frac{1}{2} \lambda, \tag{1.218}
\]

where \(\theta_0\) denotes the initial value of \(\theta\). Thus, if the geodesics initially are converging, i.e. \(\theta_0 < 0\), it follows that \(\theta(\lambda_1) = -\infty\), implying infinite convergence, at some \(\lambda_1 \leq 2/|\theta_0|\), provided that the geodesic \(\alpha\) can be extended that far. Infinite convergence means that the null geodesic generators form a caustic, which is shown on figure 1.21. The structure of the caustic causes the points \(p\) and \(q\) on the figure to be contained within the local light cone, so they are time-like separated. Hence, if the family of null geodesic generators is complete, implying that every two points on \(N\) are light-like separated, this is a contradiction.
With this background, Hawking’s area theorem can be stated and proven \[13\]:

**Area theorem**  \textit{For a predictable black hole satisfying the null energy condition, the area of spatial cross sections of the future event horizon never decreases with time.}

It has been shown explicitly for a Schwarzschild black hole in section 1.6, but from its very definition it is clear that a horizon is a null surface since nothing can go faster than the speed of light. If one assumes that the family of null generators of the horizon is complete, then it follows from the reasoning above that \( \theta \) must be greater than zero at every point on the event horizon. But it appears that the condition of completeness is not strictly necessary for \( \theta \) to be bigger than zero everywhere on the horizon. For the proof of this, see \[2\].

Because \( \theta \geq 0 \), the cross-sectional area of the horizon locally increases as one moves up the generators of \( H^+ \). Nevertheless, one has to worry about the possibility that these null geodesic generators might not reach a sufficiently late time slice \( \Sigma \), e.g. they might terminate on a singularity on the horizon, thus causing the area of \( H^+ \cap \Sigma \) to be smaller than the initial area. However, this possibility cannot occur for a predictable black hole, wherein the event horizon as well as the exterior region is contained in a globally hyperbolic region \( \mathcal{O} \) of \( M \). Namely, if \( \Sigma \) is a Cauchy surface for this region, then every null geodesic in \( \mathcal{O} \) must intersect \( \Sigma \). Thus, if \( \Sigma_1 \) and \( \Sigma_2 \) are Cauchy surfaces with \( \Sigma_2 \subset I^+(\Sigma_1) \), every generator of \( H^+ \) at \( \Sigma_1 \) must reach \( \Sigma_2 \). Thus the area of \( H^+ \cap \Sigma_2 \) must be at least as large as the area of \( H^+ \cap \Sigma_1 \), as was desired to show.

Bekenstein pointed out that there is a close analogy between this result and the second law of thermodynamics in the way that both results assert that a certain quantity never decreases with time, and used it together with thermodynamic considerations to argue that black holes should be assigned an entropy proportional to the area of the event horizon \[38\]. In the next section it is shown that other laws of black hole mechanics bear a striking mathematical similarity to the laws of thermodynamics.
1.11.2 Zeroth law

The area theorem is the only of the classical results which truly concerns the dynamics of black hole event horizons. The zeroth and first laws of black hole mechanics are concerned with equilibrium or quasi-equilibrium processes. That is, they involve stationary black holes, or adiabatic changes from one stationary black hole to another. A key ingredient for the zeroth and first law is given by [1] Hawking and Ellis

Let \((M, g_{\mu\nu})\) be an asymptotically flat spacetime which is stationary. Suppose further that \((M, g_{\mu\nu})\) is a solution of the Einstein equations with matter satisfying suitable hyperbolic equations, and that the metric and matter fields are analytic. Then the event horizon \(H^+\) of any black hole in \((M, g_{\mu\nu})\) is a Killing horizon.

Since \(H^+\) must be invariant under the Killing vector field \(k\) which is time-like at infinity as implied by stationarity, it is obvious that this Killing vector field must be tangent to \(H^+\). The above theorem states that if \(k\) fails to be normal to \(H^+\), then there exists an additional Killing field \(\xi\) which is. In the case where \(k \neq \xi\), it can be shown that a linear combination \(m\) of these two Killing fields can be chosen so that the orbits of \(m\) are closed. Thus, if \(k \neq \xi\), then \((M, g_{\mu\nu})\) is axisymmetric as well as stationary. Note the close relation to the uniqueness theorems of section 1.6. For a stationary black hole, the angular velocity of the horizon \(\Omega_H\) is defined by \(\xi = k + \Omega_H m\), just as in section 1.9.2. \(m\) is normalized such that closed orbits have period \(2\pi\) and \(k\) is normalized by requiring \(k^2 \to 1\) at infinity.

It was mentioned in section 1.6 that \(\kappa\) is constant on a bifurcate Killing horizon. Here, we will make a more general statement:

**Zeroth law** If \(T_{\mu\nu}\) obeys the dominant energy condition then the surface gravity \(\kappa\) is constant on the future event horizon of a stationary black hole.

To see this one uses again Raychaudhuri’s equation (1.216) to show that for a Killing horizon \(\mathcal{N}\) of the Killing vector field \(\xi\) it holds that

\[
R_{\mu\nu\xi\xi}\big|_{\mathcal{N}} = 0. \tag{1.219}
\]

Using this and the fact that \(\xi^2 = 0\) on \(H^+\) and the above theorem that every horizon of a stationary black hole is a Killing horizon, Einstein’s equations imply

\[
0 = -T_{\mu\nu}\xi^\mu\xi^\nu\big|_{H^+} \equiv J_\mu \xi^\mu\big|_{H^+}, \tag{1.220}
\]

so that \(J = (-T_{\mu\nu}\xi^\nu)\partial_\mu\) is tangent to \(H^+\). It follows that \(J\) can be expanded on a basis of tangent vectors to \(H^+\)

\[
J = a\xi + b_1 \eta^{(1)} + b_2 \eta^{(2)} \quad \text{on } H^+. \tag{1.221}
\]

Now \(J^2 = b_1^2 \eta^{(1)} \cdot \eta^{(1)} + b_2^2 \eta^{(2)} \cdot \eta^{(2)}\) since \(\xi \cdot \eta^{(i)} = \xi^2 = 0\). \(\xi\) is the tangent of the generators of the null surface, so it is lies in the time-direction. This implies that \(\eta^{(i)}\) are space-like vectors. So \(J\) is space-like or null (in the case that \(b_1 = b_2 = 0\)). But the dominant energy condition states that it should be time-like or null because of (1.219). So it follows that \(J \propto \xi\) which
implies
\[ \xi_{[\sigma} J_{\rho]} |_{H^+} = \frac{1}{2} (\xi_{\sigma} J_{\rho} - \xi_{\rho} J_{\sigma}) |_{H^+} = 0. \] (1.222)

Using the definition of \( J \) and again Einstein’s equations one gets
\[
0 = \xi_{[\sigma} T_{\rho]}^{\lambda} \xi_{\lambda} |_{H^+} = \xi_{[\sigma} R_{\rho]}^{\lambda} \xi_{\lambda} |_{H^+} \] (1.223)

A lengthy calculation, given in appendix B, then yields
\[
\xi_{[\rho} \partial_{\sigma] \kappa} |_{H^+} = 0 , \] (1.224)

from which one can conclude that \( \partial_{\sigma} \kappa \propto \xi_{\sigma} \). So it follows that
\[
t \cdot \partial \kappa = 0 \] (1.225)

for any tangent vector \( t \) to \( H^+ \). This shows that \( \kappa \) is constant on \( H^+ \). It provides a first indication that the surface gravity is an analogue of the temperature. This may seem a weak analogy since there are presumably many constant quantities in a stationary black hole solution. Nonetheless, it is a non-trivial statement. The link between surface gravity and temperature becomes more explicit when considering the first law of black hole mechanics.

1.11.3 First law

For simplicity, consider a vacuum black hole which is altered by dumping in a small amount of matter represented by the energy-momentum tensor \( \Delta T_{\mu \nu} \). Then, to first order in \( \Delta T_{\mu \nu} \), the change in in the black hole geometry can be neglected when computing the resulting changes in mass and angular momentum of the black hole.

First, we note that the equation
\[
\xi^\nu \nabla_\nu \xi^\mu = \kappa \xi^\mu \] (1.226)

is just the geodesic equation in the nonaffinely parametrized form. A Killing parameter \( v \) on a Killing horizon is defined by \( \xi^\mu \nabla_\mu v = 1 \). Now suppose that (1.226) holds for some parameter \( \tau \) along the null generators of the Killing horizon, so that \( \xi^\mu = dx^\mu/d\tau \). Then, if one makes the transition to another parameter \( \lambda = \lambda(\tau) \), it follows that
\[
\xi^\nu \nabla_\nu \left( \frac{dx^\nu}{d\lambda} \right) = \left( \kappa - \xi^\nu \partial_\nu \ln \left( \frac{d\lambda}{d\tau} \right) \right) \frac{dx^\nu}{d\lambda} . \] (1.227)

So, if we want to maintain the form of (1.226), it should hold that
\[
\kappa - \xi^\nu \partial_\nu \ln \left( \frac{d\lambda}{d\tau} \right) = \kappa \Rightarrow \xi^\nu \nabla_\nu \left( \frac{d\lambda}{d\tau} \right) = 0 . \] (1.228)

And because \( \xi^\nu \nabla_\nu \) is independent of \( \tau \), this can be integrated to give
\[
\xi^\nu \nabla_\nu \lambda = \text{constant} . \] (1.229)
This allows us to conclude that the geodesic equation of the null generators of a Killing horizon parametrized by a Killing parameter takes the form (1.226).

For the Kerr black hole, \( v \) reduces to the previously defined incoming null coordinate, as can be seen by using (1.178), (1.179), (1.180) and (1.154) to rewrite \( \xi \) along an outgoing radial null geodesic:

\[
\begin{align*}
\xi &= \frac{\partial}{\partial t} + \frac{a}{r^2 + a^2} \frac{\partial}{\partial \phi} \\
&= \frac{\partial}{\partial t} + \frac{a}{r^2 + a^2} \frac{\partial \lambda}{\partial \lambda} \frac{\partial r}{\partial r} \frac{\partial r}{\partial r} \\
&= \frac{\partial}{\partial t} + \frac{\partial}{\partial r^*},
\end{align*}
\]  

(1.230)

so \( v = (1/2)(t + r^*) \).

Now introduce a new parameter

\[ V = e^{\kappa v}, \]  

(1.231)

which is a generalization of the ingoing Kruskal-Szekeres coordinate. It can be shown that \( V \) is an affine parameter along the null geodesics tangent to \( \xi \) which generate the horizon. First, use (1.231) to obtain

\[
\xi^\mu = \frac{dx^\mu}{dV} dV = \kappa e^{\kappa v} \frac{dx^\mu}{dV}.
\]  

(1.232)

With this one gets

\[
\xi^\nu \nabla_\nu \xi^\mu = \xi^\nu \nabla_\nu \left( \kappa e^{\kappa v} \frac{dx^\mu}{dV} \right) = \kappa \xi^\mu.
\]  

(1.233)

Now using the zeroth law, implying that \( \kappa \) is constant, and the fact that \( v \) is a Killing parameter, this becomes

\[
\frac{dx^\mu}{dV} \nabla_\nu \left( \frac{dx^\mu}{dV} \right) = 0,
\]  

(1.234)

from which it follows that \( V \) is an affine parameter.

The changes in the mass and angular momentum of the black hole by dumping in a small amount of matter can be written as [2]

\[
\begin{align*}
\Delta M &= \int_{t^0}^{t^1} dV \int d^2S \Delta T_{\mu\nu} k^\mu t^\nu, \\
\Delta J &= -\int_{t^0}^{t^1} dV \int d^2S \Delta T_{\mu\nu} m^\mu t^\nu,
\end{align*}
\]  

(1.235)  

(1.236)

where \( k \) and \( m \) are the same vector fields as before and \( t \) is the tangent vector to the null geodesic generators of the horizon with \( V \) as an affine parameter. The second integral is over the cross-section \( S \) of the horizon corresponding to ‘time’ \( V \). Note that the product \( k^\mu dV \) is parametrization independent.

On the other hand, the change in area is governed by the Raychaudhuri equation (1.216) applied to the exact horizon. To first order in \( \Delta T_{\mu\nu} \), the quadratic terms \( \theta^2 \) and \( \sigma_{\mu\nu} \sigma^{\mu\nu} \) can be
neglected. Hence, by means of the Einstein equations, one obtains

$$\frac{d\theta}{dV} = -8\pi G \Delta T_{\mu\nu} t^\mu t^\nu.$$ (1.237)

When integrating the right hand side of this equation over the horizon, the change in the black hole geometry may be neglected. Thus, the tangent vector $t$ on the right hand side can be written in terms of the affine parameter $V$ as

$$t^\mu = \left( \frac{\partial}{\partial V} \right)^\mu,$$ (1.238)

which by means of (1.231) becomes

$$t^\mu = \frac{1}{\kappa V} \left( \frac{\partial}{\partial v} \right)^\mu.$$ (1.239)

Because $v$ is a Killing parameter it satisfies

$$1 = \xi^\mu \nabla_\nu v = \xi^\mu \partial_\nu v = \left( \frac{\partial}{\partial t} + \Omega_H \frac{\partial}{\partial \phi} \right) v,$$ (1.240)

from which follows that

$$v = \frac{1}{2} \left( t + \frac{1}{\Omega_H} \phi \right).$$ (1.241)

This allows one to write

$$\frac{\partial}{\partial v} = \frac{1}{2} \left( \frac{\partial}{\partial t} + \Omega_H \frac{\partial}{\partial \phi} \right),$$ (1.242)

which after the proper normalization leads to the identification $\partial/\partial v = \xi$. So (1.239) becomes

$$t^\mu = \frac{1}{\kappa V} \xi^\mu = \frac{1}{\kappa V} (k^\mu + \Omega_H m^\mu).$$ (1.243)

Multiplying both sides of (1.237) by $\kappa V$ and integrating over the horizon, one obtains [39]

$$\kappa \int_0^{+\infty} dV \int d^2 S V \frac{d\theta}{dV} = -8\pi G \int_0^{+\infty} dV \int d^2 S \Delta T_{\mu\nu} (k^\mu + \Omega_H m^\mu) t^\nu = -8\pi G (\Delta M - \Omega_H \Delta J),$$ (1.244)

by means of (1.235) and (1.236). The left side of this equation can be evaluated by integration by parts

$$\int d^2 S \int_0^{+\infty} dV \left( V \frac{d\theta}{dV} \right) = \int d^2 S \left[ \theta V \right]_0^{+\infty} - \int_0^{+\infty} dV \theta.$$ (1.245)

By equation (2.220), the second term on the right is just minus the change in area of the black hole. On the other hand, the first term vanishes since $V = 0$ at the lower limit and $\theta$ must vanish faster than $1/V$ as $V \to +\infty$ is the black hole is to settle down to a stationary final state with finite area. Thus, one obtains

$$\frac{\kappa}{8\pi G} \Delta A = \Delta M - \Omega_H \Delta J,$$ (1.246)
which is the first law of black hole mechanics.

Now the mathematical analogy between black hole mechanics and thermodynamics is complete. The first law of black hole mechanics has the same form as the first law of thermodynamics

\[ T\Delta S = \Delta E + P\Delta V, \quad (1.247) \]

Identification of the two laws leads to the conclusion that the black hole mass plays the role of energy in the ordinary first law and \(-\Omega_H \Delta J\) is a work term. Because if one considers a rotating body in thermodynamics, one obtains precisely such a \(-\Omega \Delta J\) term in the ordinary first law. This leads to the identification of the black hole horizon cross-section area with the entropy

\[ S = \frac{1}{4G} A, \quad (1.248) \]

implying that \(\kappa/2\pi\) should take the role of the temperature, an idea which is enforced by the zeroth law because in ordinary thermodynamics the temperature of a body in thermal equilibrium must be uniform over the body. As mentioned above, the area theorem can be viewed as an analogue of the second law, stating that the entropy cannot decrease. However, it might appear that the nature of the area theorem and the ordinary second law could hardly be more different. The area theorem is a rigorous theorem in differential geometry applicable to predictable black holes satisfying \(R_{\mu\nu}k^\mu k^\nu \geq 0\). The time asymmetry in the area theorem arises from the fact that one is dealing with a future horizon \(H^+\) rather than a past horizon. On the other hand, the ordinary second law is not believed to be a rigorous law but rather one which holds with overwhelmingly high probability. The time asymmetry of this law arises from a choice of a highly improbable initial state. Nevertheless, there are very few laws of physics which involve time asymmetric behavior of a quantity, so the analogy between these laws should not be dismissed. The third law of ordinary thermodynamics states that it is impossible to achieve absolute zero temperature in a finite series of processes. It appears that also the third law in this formulation has an analogue in black hole physics [16].

Taken all together, the mathematical analogy is very strong. Furthermore, there even is a hint that the analogy may have some physical content as well: the quantity in the laws of black hole physics which plays the role mathematically analogous to the total energy is the mass \(M\) of the black hole, which, in general relativity, physically is the total energy of the black hole. However, at this stage, here the physical analogy ends. In classical black hole physics, \(\kappa\) has nothing whatsoever to do with the physical temperature of a black hole, which is absolute zero by any reasonable criterion. We shall see in chapter 2 that this situation changes drastically when a quantum field is placed in the black hole spacetime.
Chapter 2

Quantum field theory in curved spacetime

"It is wrong to think that the task of physics is to find out how nature is. Physics concerns what we can say about nature...”
- N. Bohr

Quantum field theory in curved spacetime is a theory wherein gravity is treated in the classical, general relativistic way but matter is treated fully according to the laws of quantum field theory. It is known that the combination of these theories is not an exact description of nature, albeit there is a lot of convincing experimental evidence for both of them separately. At the present time many research is going on in finding the true theory for quantum gravity, with string theory and loop quantum gravity as two of the major candidates. But still it can be rewarding to look how quantum fields behave in a curved background. Knowing that this is only an approximation to the real situation, it is necessary to check if one is working within the range of applicability of the approximation. One should at all times stay away from the Planck-scale where quantum gravity dictates the behaviour of matter, space and time. A rule of thumb is that the radius of curvature should be bigger than the Compton wavelength of the field. This is consistent with the fact that there are no problems with putting massless fields on a curved background.

In the end, the extension of quantum field theory to a curved background turns out to be richer in consequences than one could have anticipated. It gives rise to the important processes of particle creation in cosmological and black hole spacetimes and it describes inflationary expansion which explains the primordial fluctuations that are now observed in the cosmic microwave background radiation.

In this chapter the main focus will be on the way quantum field theory is formulated in a curved background and the implications for black holes.
Chapter 2. *Quantum field theory in curved spacetime*

2.1 The formulation of QFT in curved spacetime

Quantum field theory is well-established in Minkowski spacetime, with a consistent mathematical framework, clear physical interpretations and loads of experimental confirmation. In this section it is shown how the necessary concepts can be extended to a general curved spacetime. The first results of quantum field theory in curved spacetime, obtained in the mid-sixties of the past century, were based upon the canonical formulation and contained for example the creation of particles in an expanding universe. Very soon after that it became clear that the loss of Poincaré symmetry had major impacts on the theory. Many concepts of Minkowski field theory became spoiled and the theory needed a new formulation. This resulted in the algebraic approach which, despite the successes so far, is at the present time still under development.

2.1.1 The canonical approach

First, the pioneering canonical treatment of quantum fields in a curved spacetime is presented. This is done in close analogy to [40].

Consider a curved spacetime with line element

\[ ds^2 = g_{\mu\nu}(x)dx^\mu dx^\nu. \]  

(2.1)

The metric will be treated as a given unquantized field. The spacetime is assumed to have a well-defined causal structure and a set of Cauchy hypersurfaces. Let \( n \) denote the dimension of the spacetime, with \( x^0 \) being the time coordinate and \( x^1, x^2, ..., x^{n-1} \) being the spatial coordinates. The matter fields to be quantized are denoted by \( \phi_a(x) \).

The action \( S \) is constructed from the field \( \phi_a \), so that it is invariant under general coordinate transformations (diffeomorphisms):

\[ S[\phi'(x'), \nabla' \phi'(x'), g'_{\mu\nu}(x') \]  

\[ = S[\phi(x), \nabla \phi(x), g_{\mu\nu}(x)]. \]  

(2.2)

The most easy way to do this is to start from the known Minkowski spacetime action and replace partial derivatives by covariant derivatives, the Minkowski metric by a general metric, and introduce the invariant volume element. This is called the minimal coupling description, and is consistent with the equivalence principle:

\[ \partial_\mu \rightarrow \nabla_\mu, \quad \eta_{\mu\nu} \rightarrow g_{\mu\nu}, \quad d^n x \rightarrow d^n x \left| g \right|^{1/2}, \quad g = \det(g_{\mu\nu}) \]

For the Minkowski metric the mostly minus convention (+−−−) is used. Occasionally, a term which does not vanish at the origin of a locally inertial frame can be added to the Lagrangian to increase the symmetry.
The requirement that variations of the action
\[ S = \int d^n x \, L(\phi, \nabla \phi, g_{\mu\nu}) \] (2.3)
vanish with respect to variations of the fields \( \phi_a \), which are zero on the boundary of integration, yields the equations of motion
\[ \partial_{\mu} \left( \frac{\partial L}{\partial(\partial_{\mu} \phi_a)} \right) - \frac{\partial L}{\partial \phi_a} = 0. \] (2.4)
The general covariance of the equations of motion is insured by the invariance of the action. Because \( L \) is a scalar density it transforms like \( |g|^{1/2} \).

Variation of the action with respect to the field \( g_{\mu\nu} \) generally does not vanish. However, because of the invariance of \( S \) under general coordinate transformations, \( \delta S \) will be zero under the change in \( g_{\mu\nu} \) induced by an infinitesimal coordinate transformation
\[ x^\mu \to x'^\mu = x^\mu - \epsilon^\mu(x), \] (2.5)
where \( x \) and \( x' \) refer to the same event in spacetime. Under this transformation the metric transforms like
\[ g_{\mu\nu}(x) \to g'_{\mu\nu}(x') = \frac{\partial x^\lambda}{\partial x'^\mu} \frac{\partial x^\sigma}{\partial x'^\nu} g_{\lambda\sigma}(x). \] (2.6)
And as shown in section 1.1, this leads to the following variation
\[ \delta g_{\mu\nu}(x) = g'_{\mu\nu}(x) - g_{\mu\nu}(x) = \mathcal{L}_\epsilon g_{\mu\nu}, \] (2.7)
where \( \mathcal{L}_\epsilon g_{\mu\nu} \) is the Lie derivative of the metric
\[ \mathcal{L}_\epsilon g_{\mu\nu} = \nabla_\mu \epsilon_\nu + \nabla_\nu \epsilon_\mu. \] (2.8)
Let’s assume that \( \epsilon^\mu(x) \) and \( \partial_\lambda \epsilon^\mu(x) \) are zero on the boundary of the region of integration defining the action \( S \) (when integrating over an intire infinite spacetime these quantities should drop off to zero sufficiently fast when going to infinity). In that case the variation of the action under the above infinitesimal coordinate transformation becomes
\[ \delta S = \int d^n x \, \frac{\delta L}{\delta g_{\mu\nu}(x)} \delta g_{\mu\nu}, \] (2.9)
because variations in \( S \) produced by the changes in the dynamical fields \( \phi_a \) vanish as a consequence of the equations of motion and the boundary conditions on \( \epsilon_\mu \). The invariance of \( S \) under coordinate transformations now requires \( \delta S \) to be zero. Hence, with \( dv_x = d^n x |g|^{1/2}, \)
\[ \delta S = - \int dv_x \, T^{\mu\nu} \nabla_\mu \epsilon_\nu = 0, \] (2.10)
where the energy-momentum tensor is introduced
\[ T^{\mu\nu} = -2 |g|^{-1/2} \frac{\delta S}{\delta g_{\mu\nu}(x)}, \] (2.11)
and its symmetry under interchange of indices is used. Because $T^\mu\nu\epsilon_\nu$ is a vector one gets
\[
\nabla_\mu (T^\mu\nu\epsilon_\nu) = |g|^{-1/2}\partial_\mu(|g|^{1/2}T^\mu\nu\epsilon_\nu) = (\nabla_\mu T^\mu\nu)\epsilon_\nu + T^{\mu\nu}\nabla_\mu \epsilon_\nu. \tag{2.12}
\]
So by integrating and equating the two right hand sides of this expression and using (2.10), one gets
\[
\int dv_x (\nabla_\mu T^{\mu\nu})\epsilon_\nu = 0. \tag{2.13}
\]
And because $\epsilon_\nu$ is any arbitrary infinitesimal vector the final result is
\[
\nabla_\mu T^{\mu\nu} = 0. \tag{2.14}
\]
This expresses the conservation of energy and momentum in a general curved spacetime. The advantage of this construction is that the energy-momentum tensor defined by it is symmetric, which is not true in general for the canonical energy-momentum tensor
\[
\Theta^{\mu\nu} = \frac{\partial \mathcal{L}}{\partial (\partial_\nu \phi_a)} \partial_\mu \phi_a - \delta^{\mu}_{\nu} \mathcal{L}. \tag{2.15}
\]
From $\delta (g^{\mu\nu} g_{\mu\nu}) = 0$, it follows that
\[
g_{\mu\lambda}g_{\nu\sigma} \frac{\delta}{\delta g^{\lambda\sigma}} = - \frac{\delta}{\delta g^{\mu\nu}}. \tag{2.16}
\]
And with this one can rewrite (2.11) as
\[
T^{\mu\nu} = 2|g|^{-1/2} \frac{\delta S}{\delta g^{\mu\nu}(x)}. \tag{2.17}
\]
The sign convention in the definition of $T^{\mu\nu}$ is chosen such that $T^{00}$ is positive for the classical electromagnetic field with the used mostly minus sign convention for the metric.

One can calculate the symmetric energy-momentum tensor $T^{\mu\nu}$ in curved spacetime and then go to the flat spacetime limit, thereby obtaining a symmetric energy-momentum tensor satisfying $\partial_\mu T^{\mu\nu} = 0$ in Minkowski spacetime. For any isolated system in Minkowski spacetime, both $\Theta^{\mu\nu}$ and $T^{\mu\nu}$ yield a unique conserved energy-momentum vector $p_\mu$. In curved spacetime it is the symmetric energy-momentum tensor $T^{\mu\nu}$ which describes the matter and radiation and couples to the gravitational field through the Einstein field equations.

The Schwinger operator action principle [41] continues to hold in curved spacetime for an arbitrary infinitesimal transformation of the form
\[
x^\mu \rightarrow x'^\mu = x^\mu + \delta x^\mu
\]
\[
\phi(x) \rightarrow \phi'(x) = \phi(x) + \delta \phi(x), \tag{2.18}
\]
provided that under this transformation $\delta g_{\mu\nu}(x) = 0$, or
\[
g'_{\mu\nu}(x) = g_{\mu\nu}(x). \tag{2.19}
\]
Because in that case the derivation as done in Minkowskian space can be extended to a general spacetime without $g_{\mu\nu}$ having to satisfy the Euler-Lagrange equations. The Schwinger operator action principle then gives an expression for the variation of the action:

$$\delta S = G(t_2) - G(t_1),$$

with

$$G(t) = \int d^{n-1}x \left[ \pi_a \delta \phi_a - \Theta^0 \delta x_{\nu} \right].$$

The integration is on a constant time hypersurface, and

$$\pi_a = \frac{\partial L}{\partial (\partial_0 \phi_a)}.$$  

$G$ is the generator of the transformation, satisfying

$$i \delta F = [F, G],$$

where $F$ is a functional of the $\phi_a$ and $\pi_a$. Quantization of the theory in the canonical way can be done in complete analogy to Minkowski spacetime, with the relations

$$[\phi_a(\vec{x}, t), \phi_b(\vec{x}', t)] = 0,$$

$$[\pi_a(\vec{x}, t), \pi_b(\vec{x}', t)] = 0,$$

$$[\phi_a(\vec{x}, t), \pi_b(\vec{x}', t)] = i \delta_{ab} \delta(\vec{x} - \vec{x}')$$

for bosons, and

$$\{\phi_a(\vec{x}, t), \phi_b(\vec{x}', t)\} = 0,$$

$$\{\pi_a(\vec{x}, t), \pi_b(\vec{x}', t)\} = 0,$$

$$\{\phi_a(\vec{x}, t), \pi_b(\vec{x}', t)\} = i \delta_{ab} \delta(\vec{x} - \vec{x}')$$

for fermions. Here $\delta(\vec{x} - \vec{x}')$ is the Dirac delta-function satisfying $\int d^{n-1}x \delta(\vec{x} - \vec{x}')f(\vec{x}) = f(\vec{x}')$ with the integral being performed over the spacelike hypersurface $t = \text{constant}$. One can show that $\delta(\vec{x} - \vec{x}')$ and $\pi(\vec{x}', t)$ transform as spatial scalar densities under transformations of the spatial coordinates on the constant-$t$ hypersurface. Hence, the above commutation and anti-commutation relations are covariant under transformations of the spatial coordinates on the hypersurface, and are therefore the spatially covariant generalization of the corresponding relations that hold in flat spacetime. Also, they are consistent with the equations of motion of the fields, in the sense that if they hold on one constant-$t$ spatial hypersurface, then they also will hold on the other constant-$t$ hypersurfaces.

There are several cases of interest when $G$ is conserved. The simplest is when $\delta x^\mu = 0$ and $\delta \phi_a$ is a symmetry of $L$. Then of course $\delta g_{\mu\nu} = 0$ and the symmetry of $L$ implies that $\delta S = 0$ so that $G$ is independent of time. One also has in that case $\partial_{\mu} J^\mu = 0$, with

$$J^\mu = \frac{\partial L}{\partial (\partial_{\mu} \phi_a)} \delta \phi_a.$$  

This follows because $J^\mu$ is a vector density, so that we have $\nabla_{\mu} (|g|^{-1/2} J^\mu) = 0$. Thus, in a curved spacetime the charges and generators of internal symmetries continue to be conserved.
The generator $G$ is also conserved when $\delta x^\mu \neq 0$, but is generated by a Killing vector field so that $\delta g_{\mu\nu} = 0$ holds. Then, invariance of the action under coordinate transformations implies that $\delta S = 0$ and it follows from the Schwinger operator action principle that $G$ is constant. For example, if it is possible to choose a coordinate system in which a particular coordinate, say $x^\lambda$, does not appear in $g_{\mu\nu}$, then under a translation in the $x^\lambda$ direction one has $\delta g_{\mu\nu} = 0$ and $\delta \phi(x) = 0$. Then it follows from the definition of $G$ that

$$p_\lambda = \int d^{n-1}x \, \Theta^0_\lambda$$  \hspace{1cm} (2.27)

is constant. Since $g^{\lambda\mu}$ and the other components of $p_\mu$ may not be constants, it does not follow that $p^\lambda$ is constant.

A similar expression involving the symmetric energy-momentum tensor $T_{\mu\nu}$ also holds. Consider again an infinitesimal transformation of the spacetime coordinates, this time slightly rewritten as

$$x^\mu \rightarrow x'^\mu = x^\mu - \epsilon \xi^\mu(x),$$  \hspace{1cm} (2.28)

where $\epsilon$ is an infinitesimal parameter. Now take $\xi$ to be a Killing vector field, i.e. $\delta g_{\mu\nu} = \mathcal{L}_\xi g_{\mu\nu} = 0$. Because of the conservation of energy in curved spacetime (2.14) one has

$$\nabla_\mu (T_{\mu\nu} \xi_\nu) = T_{\mu\nu} \nabla_\mu \xi_\nu.$$  \hspace{1cm} (2.29)

Since $\xi$ is a Killing vector field, the Killing vector lemma implies that $\nabla_\mu \xi_\nu$ is anti-symmetric. So because of the symmetry of $T_{\mu\nu}$ one gets

$$\nabla_\mu (T_{\mu\nu} \xi_\nu) = 0.$$  \hspace{1cm} (2.30)

But because $T_{\mu\nu} \xi_\nu$ is a vector, it holds that

$$\nabla_\mu (T_{\mu\nu} \xi_\nu) = |g|^{-1/2} \partial_\mu (|g|^{1/2} T_{\mu\nu} \xi_\nu),$$  \hspace{1cm} (2.31)

and therefore (2.30) becomes

$$\partial_\mu (|g|^{1/2} T_{\mu\nu} \xi_\nu) = 0.$$  \hspace{1cm} (2.32)

So one gets for the conserved quantity

$$P_\xi \equiv \int dv_x T^\nu_{\nu}(x) \xi^\nu(x).$$  \hspace{1cm} (2.33)

In the case where the coordinates are such that $g_{\mu\nu}$ is independent of a particular coordinate $x^\lambda$ then $\xi^\nu = \delta^\nu_\lambda$ is a Killing vector field, and the conserved quantity reduces to $P_\lambda = \int dv_x T^\lambda_\lambda$. In such case, (2.27) should give the same result up to an additive constant independent of the field configuration.

So far, the action, field equations, symmetric energy-momentum tensor, generators of field transformations, commutation or anti-commutation relations and conservation laws were taken under consideration. Up to this point everything was a straightforward extension from flat to curved spacetime. But from this point on, conceptual difficulties will arise and we will be forced to take our conceptions about quantum field theory to a deeper level. The promised richer then
expected features of quantum field theory in curved spacetime will begin to reveal themselves in the next section.

2.1.2 A cosmological model

In this section the canonical treatment of a quantum field in an expanding universe is presented [42]. This is done because it gives a first intuitive notion of the difficulties of quantum fields in general spacetimes. So, however this thesis does not concern itself with cosmological purposes, it is still very instructive (albeit in an indirect way) to consider this model in order to develop a deeper understanding of the necessary concepts to be used later in this thesis, such as: the creation of particles by a changing metric, the role of Bogoliubov transformations and the ambiguity of the vacuum state in curved spacetime.

2.1.2.1 The set-up

The spacetime under consideration is described by the spatially flat isotropically changing metric

\[ ds^2 = dt^2 - a^2(t)(dx^2 + dy^2 + dz^2). \]  

(2.34)

The scale factor can have an arbitrary time dependence but the (highly unrealistic) assumption is made that it must asymptotically approach constant values at early and late cosmic time. This cosmic time \( t \) is the proper time of a set of clocks on geodesic worldlines that remain at constant values of the spatial coordinates \( (x, y, z) \). So, following behaviour is assumed

\[ a(t) \sim \begin{cases} 
  a_1 & \text{as } t \to -\infty \\
  a_2 & \text{as } t \to +\infty 
\end{cases} \]  

(2.35)

Also, \( a(t) \) has to be sufficiently smooth and approach the constant values sufficiently fast.

As quantum field, a massless scalar is used according to the rules of the minimal coupling description. Because covariant derivatives are the same as partial derivatives for scalar quantities, the action and Lagrangian density become

\[ S = \int d^n x \mathcal{L}, \]

\[ \mathcal{L} = \frac{1}{2} |g|^{1/2} g^{\mu \nu} \partial_\mu \phi \partial_\nu \phi. \]  

(2.36)

This Lagrangian density gives rise to the following field equation

\[ \Box \phi = 0. \]  

(2.37)
Chapter 2. *Quantum field theory in curved spacetime*

Before continuing the treatment of a scalar field in an expanding universe a scalar product between two spacetime functions, which provides the spacetime with a natural symplectic structure, is introduced:

\[
(f_1, f_2) \equiv i \int d^{n-1}x \, |g|^{1/2} g^{0\nu} (f_1^*(\vec{x}, t) \partial_\nu f_2(\vec{x}, t) - \partial_\nu f_1^*(\vec{x}, t) f_2(\vec{x}, t))
\]

Where the integration is taken over a constant \(t\) hypersurface. If \(f_1\) and \(f_2\) are solutions of the field equation which vanish at spatial infinity, then their scalar product is conserved since

\[
\frac{d}{dt}(f_1, f_2) = i \int d^{n-1}x \, \partial_0 (|g|^{1/2} g^{0\nu} f_1^* \overset{\leftrightarrow}{\partial}_\nu f_2)
\]

\[
= i \int d^{n-1}x \, |g|^{1/2} \nabla_\mu (g^{\mu\nu} f_1^* \overset{\leftrightarrow}{\partial}_\nu f_2) - i \int d^{n-1}x \, \partial_i (|g|^{1/2} g^{i\nu} f_1^* \overset{\leftrightarrow}{\partial}_\nu f_2)
\]

\[
= 0
\]

(2.38)

Where the basic identity \(\nabla_\mu V^\mu = |g|^{-1/2} \partial_\mu (|g|^{1/2} V^\mu)\), valid for any vector field \(V^\mu\), was used.

In terms of a general spacelike hypersurface \(\sigma\) with future-directed unit normal \(n^\mu\) the scalar product is

\[
(f_1, f_2) = i \int_\sigma d\sigma \, |g|^{1/2} n^\nu f_1^* \overset{\leftrightarrow}{\partial}_\nu f_2.
\]

(2.40)

This scalar product is conserved under deformations of \(\sigma\). Suppose \(\sigma \to \sigma'\) such that \(\sigma\) and \(\sigma'\) form the spacelike boundaries of a volume \(v\) (there may also be timelike boundaries of \(v\) at spatial infinity). Then by the Gauss divergence theorem

\[
(f_1, f_2)_{\sigma'} - (f_1, f_2)_{\sigma} = i \int_v d^n x \, \partial_\mu (|g|^{1/2} f_1^* \overset{\leftrightarrow}{\partial}_\mu f_2)
\]

\[
= i \int_v d^n x \, |g|^{1/2} \nabla_\mu (f_1^* \overset{\leftrightarrow}{\nabla}_\mu f_2)
\]

\[
= 0
\]

(2.41)

as a consequence of the field equation.

Now all the necessary ingredients are acquired to start the first discussion of a quantum field in a specific non-Minkowski spacetime. First, the field equation can be written down explicitly as

\[
a^{-3} \partial_t (a^3 \partial_t \phi) - a^{-2} \sum_i \partial_i^2 \phi = 0.
\]

(2.42)

It is convenient to impose periodic boundary conditions in a cube having sides of coordinate length \(L\). Just as in Minkowski spacetime, this is a mathematical trick, with \(L\) taken to infinity after physical quantities are calculated. The field operator can now be expandend in the form

\[
\phi = \sum_k \{ A_k f_k(x) + A_k^* f_k^*(x) \},
\]

(2.43)
where

\[ f_{\vec{k}} = V^{-1/2} e^{i\vec{k} \cdot \vec{x}} \psi_k(\tau) \quad , \quad k^i = \frac{2\pi n^i}{L} \quad (n^i \in \mathbb{Z} \quad , \quad k = |\vec{k}|) \]  

(2.44)

And a \( \tau \) is defined by

\[ \tau = \int t a^{-3}(t') dt' . \]  

(2.45)

It follows from the field equation that

\[ \frac{d^2 \psi_k}{d\tau^2} + k^2 a^4 \psi_k = 0 . \]  

(2.46)

Now the initial condition is imposed that in the flat spacetime with \( a = a_1 \), which is approached at early times, we have the Minkowski spacetime field expansion for \( \psi \), but with the constant scale factor \( a_1 \) taken into account. This implies that for \( t \to -\infty \), we get the asymptotic behaviour

\[ f_{\vec{k}} \sim (V a_1^3)^{-1/2} (2\omega_{1k})^{-1/2} \exp[i(\vec{k} \cdot \vec{x} - \omega_{1k} t)] \]  

(2.47)

with \( \omega_{1k} = k/a_1 \). In the spacetime at early times the coordinates can be rescaled like \( x^i \to x'^i = a_1 x^i \), so that we have the usual Minkowski metric and \( x'^i \) is the physical or measured distance. The appropriate rescaled physical momentum is then \( k'^i = k^i / a_1 \), and the physical energy of a particle is \( |\vec{k}'| = k/a_1 = \omega_{1k} \).

From (2.44) it can be seen that the above asymptotic behaviour for \( f_{\vec{k}} \) comes down to putting the following asymptotic condition on \( \psi_k \) as \( t \to -\infty \)

\[ \psi_k(\tau) \sim (2a_1^3 \omega_{1k})^{-1/2} \exp(-i\omega_{1k} a_1^3 \tau) , \]  

(2.48)

where (2.45) was used to replace \( t \) by \( a_1^3 \tau + \) constant, and the constant phase factor was absorbed into the definition of \( A_{\vec{k}} \) (or into the choice of the time origin).

With (2.47), the scalar product (2.38) becomes

\[ (f_{\vec{k}}, f_{\vec{k}'}^\dagger) = \delta_{\vec{k}, \vec{k}'} , \quad (f_{\vec{k}}, f_{\vec{k}'}^\dagger) = 0 . \]  

(2.49)

And because the scalar product is conserved, these relations must hold at all times. Similarly, the Minkowski spacetime quantization in the initial flat spacetime implies that the operators \( A_{\vec{k}} \) satisfy

\[ [A_{\vec{k}}, A_{\vec{k}'}^\dagger] = \delta_{\vec{k}, \vec{k}'} , \quad [A_{\vec{k}}, A_{\vec{k}'}^\dagger] = 0 , \]  

(2.50)

and that \( A_{\vec{k}} \) annihilates particles with momentum \( \vec{k}/a_1 \) and energy \( \omega_{1k} \) in the initial Minkowski spacetime. The operators \( A_{\vec{k}} \) are time-independent so (2.50) is valid at all times.

From (2.51) and (2.50), it follows that the canonical commutation relations hold,

\[ [\phi(\vec{x}, t), \phi(\vec{x}', t)] = 0 \quad , \quad [\pi(\vec{x}, t), \pi(\vec{x}', t)] = 0 \]  

\[ [\phi(\vec{x}, t), \pi(\vec{x}', t)] = i\delta(\vec{x} - \vec{x}') \]  

(2.51)

where \( \pi \) is given by (2.22) as

\[ \pi = a^3 \partial_t \phi = \partial_x \phi . \]  

(2.52)
The proof for the canonical commutation relations (2.51) is as follows. From the field operator expansion and (2.50) one finds
\[ [\phi(\vec{x}, t), \pi(\vec{x}', t)] = a^3(t) \sum_k \{ f_k^*(\vec{k}, t) \partial_t f_k(\vec{x}', t) - f_k^*(\vec{k}, t) \partial_t f_k^*(\vec{x}, t) \}. \] (2.53)

An arbitrary solution of \( \Box h = 0 \) can be expanded in the form
\[ h(\vec{x}, t) = \sum_k \{ f_k^*(\vec{k}, t)(f_k, h) - f_k^*(\vec{k}, t)(f_k^*, h) \} \]
\[ = -i \int d^3x' a^3(t) \sum_k \{ f_k^*(\vec{k}, t) \partial_t f_k^*(\vec{x}', t) - f_k^*(\vec{k}, t) \partial_t f_k(\vec{x}', t) \} h(\vec{x}', t) \]
\[ + i \int d^3x' \sum_k \{ f_k^*(\vec{k}, t)f_k^*(\vec{x}', t) - f_k^*(\vec{k}, t)f_k(\vec{x}, t) \} \partial_t h(\vec{x}', t), \] (2.54)

where the \( a^3(t) \) comes from the \(|g|^{1/2} \) in the scalar product (2.38). From the above identity one can conclude
\[ a^3(t) \sum_k \{ f_k^*(\vec{k}, t) \partial_t f_k^*(\vec{x}', t) - f_k^*(\vec{k}, t) \partial_t f_k^*(\vec{x}', t) \} = i\delta(\vec{x} - \vec{x}'), \] (2.55)
\[ \sum_k \{ f_k^*(\vec{k}, t)f_k^*(\vec{x}', t) - f_k^*(\vec{k}, t)f_k(\vec{x}', t) \} = 0. \] (2.56)

The canonical commutation relation of \( \phi \) with \( \pi \) follows from (2.55), and that of \( \phi \) with \( \phi \) from (2.56).

The equal time commutator of \( \pi \) with \( \pi \) is found by taking the time derivative of the previous expansion of \( h \) and recalling that the scalar products are conserved. Then
\[ \partial_t h(\vec{x}, t) = \sum_k \{ \partial_t f_k^*(\vec{k}, t)(f_k, h) - \partial_t f_k^*(\vec{k}, t)(f_k^*, h) \}, \] (2.57)

where the scalar products can be expanded as before, and using (2.55) one obtains
\[ \sum_k \{ \left( \partial_t f_k^*(\vec{x}', t) \right) \partial_t f_k^*(\vec{x}', t) - \left( \partial_t f_k^*(\vec{x}', t) \right) \partial_t f_k^*(\vec{x}, t) \} = 0, \] (2.58)
from which \([\pi(\vec{x}, t), \pi(\vec{x}', t)] = 0 \) follows.

For the results of this model both asymptotically flat regions were not required. It would be sufficient, for example, to suppose that at some time in the distant future the universe becomes asymptotically flat, although it may have never been flat at earlier times. Then the canonical commutation relations would have to hold when \( a(t) \) is changing rapidly if they are to hold the far future. Causality then implies that the canonical commutation relations must hold even if the universe never becomes asymptotically flat. Thus, one sees that the canonical commutation relations hold in this curved spacetime with any \( a(t) \) as a consequence of their holding in Minkowski spacetime.
2.1.2.2 Particle creation

Because of the asymptotic behaviour at early times, in the initial Minkowski spacetime, \( f_\vec{k} \) is a positive frequency solution of the field equation (2.37) and \( A_\vec{k} \) is a particle annihilator. Suppose now that the state vector describing the system in the Heisenberg picture is such that no particles are present at early times. Denoting this state vector by \( |0\rangle \), this implies

\[
A_\vec{k} |0\rangle = 0, \quad \forall \vec{k}.
\]  

(2.59)

The time evolution of \( \psi_k(\tau) \) is governed by the ordinary second-order differential equation (2.46). This equation has two linearly independent solutions \( \psi_k^{(\pm)}(\tau) \) with asymptotic behaviour at late times (\( t \to +\infty \))

\[
\psi_k^{(\pm)} \sim (2a_2^3 \omega_{2k})^{-1/2} \exp(\mp i \omega_{2k} \tau),
\]  

(2.60)

where \( \omega_{2k} \equiv k/a_2 \). Therefore, the solution of the differential equation (2.46) can be written in its most general form by

\[
\psi_k(\tau) = \alpha_k \psi_k^{(+)}(\tau) + \beta_k \psi_k^{(-)}(\tau)
\]  

where \( \alpha_k \) and \( \beta_k \) are two complex constants depending on the form of \( a(t) \). So the solution of interest here, with the early time behaviour as imposed by (2.48), must have late time behaviour as \( t \to +\infty \)

\[
\psi_k(\tau) \sim (2a_2^3 \omega_{2k})^{-1/2} [\alpha_k e^{-i \omega_{2k} \tau} + \beta_k e^{i \omega_{2k} \tau}]
\]  

(2.61)

The Wronskian of the differential equation for \( \psi_k \) (2.46) gives the conserved quantity

\[
\psi_k \partial_\tau \psi_k^* - \psi_k^* \partial_\tau \psi_k = i
\]  

(2.62)

where the right hand side is determined by the imposed early time asymptotic form of \( \psi_k \) (2.48). Filling in the late time asymptotic form of \( \psi_k \) then requires that

\[
|\alpha_k|^2 - |\beta_k|^2 = 1.
\]  

(2.63)

From (2.44) and the late time asymptotic form of \( \psi_k \) (2.61), one finds the following late time behaviour for \( f_\vec{k} \)

\[
f_\vec{k} \sim (Va_2^3)^{-1/2} (2\omega_{2k})^{-1/2} e^{i \vec{k} \cdot \vec{x}} [\alpha_k e^{-i \omega_{2k} \tau} + \beta_k e^{i \omega_{2k} \tau}],
\]  

(2.64)

where it is used that \( a_2^3 \tau \sim t + \text{constant} \) at late \( t \) and the constant phase factors are absorbed in \( \alpha_k \) and \( \beta_k \). At this point, the asymptotic form at late times of \( \phi \) can be written down by regrouping the early time expansion (2.43) according to late time positive and negative frequency parts:

\[
\phi(x) = \sum_k \{a_k^* g_k^*(x) + a_k^\dagger g_k(x)\},
\]  

(2.65)

with \( g_k \) being a solution of the field equation which is positive the positive frequency part at late times,

\[
g_k(x) \sim (Va_2^3)^{-1/2} (2\omega_{2k})^{-1/2} \exp[i(\vec{k} \cdot \vec{x} - \omega_{2k} t)],
\]  

(2.66)

and with

\[
a_k \equiv \alpha_k A_k + \beta_k^* A_k^\dagger.
\]  

(2.67)
Chapter 2. *Quantum field theory in curved spacetime*

68

The $a_k$ can be interpreted as annihilation operators for particles of momentum $\vec{k}/a_2$ at late times. This interpretation is consistent, since we have

$$[a_{\vec{k}}, a^\dagger_{\vec{k}'}] = \delta_{\vec{k},\vec{k}'}(|\alpha_{\vec{k}}|^2 - |\beta_{\vec{k}}|^2) = \delta_{\vec{k},\vec{k}'}$$

(2.68)

The transformation of annihilation and creation operators such as that in (2.67) are known as Bogoliubov transformations.

And now happens the final 'magic' of this model. Using the $a_k$ and the state vector $|0\rangle$ one can calculate the expectation value of the number of particles present at late times in mode $\vec{k}$:

$$\langle N_{\vec{k}} \rangle_{t \to +\infty} = \langle 0 | a^\dagger_{\vec{k}} a_{\vec{k}} | 0 \rangle = |\beta_{\vec{k}}|^2 .$$

(2.69)

And on the other hand, at early times

$$\langle N_{\vec{k}} \rangle_{t \to -\infty} = \langle 0 | A^\dagger_{\vec{k}} A_{\vec{k}} | 0 \rangle = 0 .$$

(2.70)

Thus, if $a(t)$ is such that $|\beta_{\vec{k}}|^2$ is non-zero, as is generally the case, *particles are created by the changing scale factor of the universe*. The above results can readily be extended to the massive case. All results remain the same, but now with the particle energies at early times given by $\omega_{1k} = \sqrt{(k/a_1)^2 + m^2}$ and at late times by $\omega_{2k} = \sqrt{(k/a_2)^2 + m^2}$.

2.1.2.3 Conclusions

In this model the important and very subtle role of boundary conditions (at Cauchy surfaces in spacetime) appeared for the first time. It is one of the key features of quantum field theory in curved spacetime because of the absence of the uniformity of spacetime. It will also play a vital role in the derivation of particle creation by black holes.

Particles are created, rather than annihilated, regardless of the relation between $a_1$ and $a_2$. This occurs, despite the time reversal invariance of the field equation, because we have chosen the state vector such that no particles are present at early times. In the time-reversed situation, in which particles are annihilated so that none are present at late times, we would have to take the state vector to be one in which initially there are correlated pairs of particles present. Such an initial state unnatural in a physical context because of the correlations required.

During a rapid change of $a(t)$ in which particle creation is occurring, the particle number is not operationally well defined. Suppose one tries to measure the particle number in a comoving volume (one bounded by geodesics of the spacetime), and that the measurement process takes place in a time interval $\Delta t$. If $\Delta t$ is very small, a significant number of particles will be created by the measurement process because of the time-energy uncertainty relation. But if $\Delta t$ is large, then a significant number of particles will be created by the change of $a(t)$ during the time of the measurement. There is no value of $\Delta t$ for which the minimum uncertainty in the measured particle number is 0. This irreducible imprecision in the measured particle number will become large during a process of rapid particle creation. The uncertainty is reflected in the theory by
the absence of an unambiguous or unique definition of a positive frequency solution corresponding to physical particles during a period when $a(t)$ is changing. This ambiguity of the particle interpretation of quantum field theory naturally carries over to more general non-static curved spacetimes as well as to spacetimes with event horizons.

The lack of a unique particle interpretation means that in a general curved spacetime, in contrast to Minkowski spacetime, there is no physically unambiguous unique Heisenberg state vector which can be identified as the vacuum state. This is explicitly the case for the cosmological model above, where although there were no non-gravitational interactions present, the state vector containing no particles at early times was different from the state vector containing no particles at late times. In the context of this model it is also shown that the early time and late time vacuum states are orthogonal to one another, thus giving unitary inequivalent representations of the commutation relations in curved spacetime [43].

The very intimate relation between spin and statistics appears very naturally when putting a quantum field in a general spacetime background. For the scalar field as treated above only Bose-Einstein statistics seems to be consistent with the dynamics of the field [40]. Otherwise, the particles at early times would obey different statistics than the particles at late times, clearly something which is physically not acceptable. This curved-spacetime derivation of the spin-statistics theorem has been extended to higher spin fields [44] and to ghost fields [45].

2.1.3 The loss of Poincaré symmetry

The above canonical treatment of quantum field theory in curved spacetime exposed some conceptual difficulties, like the non-unique definition of annihilation operators, creation operators and the vacuum state. In this section the aim is to take a closer look at exactly what difficulties come up and why they come up. This is done to point out how subtle and highly non-straightforward it is to place a quantum field in a curved background. Because as shown below, quantum field theory as usually formulated contains many elements that are very special to Minkowski spacetime.

It is relatively simple to generalize classical field theory from flat to curved spacetime. That is because there is a clean separation between the field equations and the solutions. The field equations can be easily generalized to curved spacetime in an entirely local and covariant manner.

In quantum field theory, ‘states’ are the analogs of ‘solutions’ in classical field theory. However, properties of these states are deeply embedded in the usual formulations of quantum field theory in Minkowski spacetime. One particular and important example is the Poincaré invariance of the vacuum state.
2.1.3.1 The particle content of the Klein-Gordon field

A simple and concrete example that illustrates the key features is given in [46, 47]. Consider a free, real Klein-Gordon field $\psi$ in flat spacetime

$$(\partial^2 - m^2)\psi = 0. \tag{2.71}$$

The usual route towards formulating a quantum theory of $\psi$ is to decompose it into a series of modes, and then treat each mode by the rules of ordinary quantum mechanics. The field is put in a cubic box of side $L$ with periodic boundary conditions. The field can then be decomposed as a Fourier series in terms of the modes

$$\psi_k \equiv \frac{1}{L^{3/2}} \int d^3x \, e^{-ik\cdot x} \psi(t, \vec{x}) \; , \; \vec{k} = (2\pi/L)(n_1, n_2, n_3). \tag{2.72}$$

The Hamiltonian of the system is given by

$$H = \sum_k \frac{1}{2} \left( |\psi_k|^2 + \omega_k^2 |\psi_k|^2 \right) \; , \; \omega_k^2 = |\vec{k}|^2 + m^2. \tag{2.73}$$

So it follows that the free Klein-Gordon field in flat spacetime is equivalent to an infinite collection of decoupled harmonic oscillators. Going to normal modes and quantizing the field by means of the usual commutation relations then gives

$$\psi(t, \vec{x}) = \frac{1}{L^{3/2}} \sum_k \frac{1}{2\omega_k} \left( e^{ik\cdot (\vec{x} - \omega_k t)} a_k + e^{-ik\cdot (\vec{x} + \omega_k t)} a_k^\dagger \right). \tag{2.74}$$

States of the free Klein-Gordon field are given the following interpretation: the state denoted by $|0\rangle$ in which all of the oscillators comprising the Klein-Gordon field are in their ground state is interpreted as 'the vacuum'. States of the form $(a_k^\dagger)^n|0\rangle$ are interpreted as ones containing $n$ 'particles'. In an interacting theory, the evolution of the field may be such that it behaves like a free field at early and at late times. In that case, one has a particle interpretation at those early and late times. The relationship between the early and late time particle description of a state is given by the S-matrix and contains a great deal of information about the interacting theory.

The cornerstone of the definition and interpretation of the 'vacuum' and 'particles' in the discussion above is the ability to decompose the field into its positive and negative frequency parts as can be seen in (2.74). The ability to define this decomposition makes crucial use of the presence of a time translation symmetry in the background Minkowski spacetime. In a generic curved spacetime without symmetries, there is no natural notion of 'positive frequency solutions' and, consequently, no natural notion of a 'vacuum state' or 'particles'.

2.1.3.2 The lack of spacetime symmetries

To examine what properties of Minkowski spacetime are used in an essential way in the usual formulation of quantum field theory, the Wightman axioms [48] are considered because they
abstract the key features of quantum field theory in Minkowski spacetime in a mathematically clear way. The **Wightman axioms** are the following:

1. The states of the theory are unit rays in a Hilbert space $\mathcal{H}$ that carries a unitary representation of the Poincaré group.

2. The four-momentum that is defined by the action of the Poincaré group on the Hilbert space is positive which means its spectrum is contained within the closed future light cone. (= spectrum condition)

3. There exists a unique Poincaré invariant state, the 'vacuum'.

4. The quantum fields are operator-valued distributions defined on a dense domain $\mathcal{D} \subset \mathcal{H}$ that is both Poincaré invariant and invariant under the action of the fields and their adjoints.

5. The fields transform in a covariant manner under the action of Poincaré transformations.

6. At spacelike separations, the quantum fields either commute or anti-commute.

It is clear that the Wightman axioms rely strongly on Poincaré symmetry, except for the last one. So only this sixth and last axiom can be readily extended to a general spacetime. Since a generic curved spacetime will not possess any symmetries at all, one can certainly not require Poincaré invariance/covariance or invariance under any other type of spacetime symmetry. In the following, the implications for axioms 2 and 3, and the perturbation and renormalization prescriptions for a quantum field theory are discussed.

**Axiom 2**

The energy-momentum tensor $T_{\mu\nu}$ of a classical field in curved spacetime is well defined and it satisfies local energy-momentum conservation in the sense that $\nabla^\mu T_{\mu\nu} = 0$. If $t^\mu$ is a vector field on the spacetime that represents time translations and $\Sigma$ is a Cauchy surface, one can define the total energy $E$ of the field at 'time' $\Sigma$ by

$$E = \int_\Sigma d\Sigma T_{\mu\nu} t^\mu n^\nu.$$  \hfill (2.75)

Classically, the energy-momentum tensor satisfies the dominant energy condition which means $T_{\mu\nu} t^\mu n^\nu \geq 0$ [1]. Thus, classically, one has $E \geq 0$. However, unless $t^\mu$ is a Killing vector field, which means that the spacetime would be stationary, $E$ will not be conserved, i.e. independent of the choice of Cauchy surface $\Sigma$.

In quantum field theory, it is expected that the energy-momentum operator will be well defined as an operator-valued distribution (see below), and it is expected to be conserved, $\nabla^\mu T_{\mu\nu}$. However, this definition requires spacetime smearing (see (2.76) below). In Minkowski spacetime one can do 'time smearing' without changing the value of $E$, since $E$ is conserved, and there is a unique and well defined notion of total energy. However, in the absence of time translation symmetry, one cannot expect $E$ to be well defined at a sharp moment of time. More importantly, it is well known that $T_{\mu\nu}$ cannot satisfy the dominant energy condition in quantum field
theory, even when it holds for the corresponding classical theory, so locally energy densities can be arbitrarily negative [47]. It is nevertheless true in Minkowski spacetime that the total energy is positive for physically reasonable states. However, in a curved spacetime without symmetries there is no reason to expect any ‘time smeared’ version of $E$ to be positive.

Furthermore, there are simple examples with time translation symmetry, such as a two-dimensional massless Klein-Gordon field in an $S^1 \otimes \mathbb{R}$ background, where $E$ can be computed explicitly and is found to be negative [49]. Or, as another example, in de Sitter spacetime there is no globally timelike Killing field and therefore no global notion of energy that is positive [47]. Thus, it appears hopeless to generalize the spectrum condition to curved spacetime in terms of the positivity of a quantity representing the ‘total energy’.

Axiom 3
As already noted above, for a free field in Minkowski spacetime, the notion of ‘particles’ and ‘vacuum’ is intimately related to the notion of ‘positive frequency solutions’ which in turn relies on the existence of a time translation symmetry. These notions of a unique ‘vacuum state’ and ‘particles’ can be straightforwardly generalized to globally stationary curved spacetimes. However, there is no natural notion of ‘positive frequency solutions’ in a general, non-stationary curved spacetime.

Nevertheless for a free field on a general spacetime, a notion of ‘vacuum state’ can be defined as follows. A state is said to be quasi-free if all of its n-point functions $\langle \psi(x_1) \ldots \psi(x_n) \rangle$ can be expressed in terms of its 2-point function by the same formula as holds for the ordinary vacuum state in Minkowski spacetime. A state is said to be Hadamard if the singularity structure of its 2-point function $\langle \psi(x_1)\psi(x_2) \rangle$ in the coincidence limit $x_1 \to x_2$ is the natural generalization to curved spacetime of the singularity structure of $\langle 0|\psi(x_1)\psi(x_2)|0 \rangle$ in Minkowski spacetime. Thus, in a general curved spacetime, the notion of a quasi-free Hadamard state provides a notion of a ‘vacuum state’, associated to which is a corresponding notion of ‘particles’.

The problem is that this notion of a vacuum state is highly non-unique. For spacetimes with a non-compact Cauchy surface, different choises of quasi-free Hadamard states give rise, in general, to unitarily inequivalent Hilbert space constructions of the theory, so in this case it is not even clear what the correct Hilbert space of states should be. In the absence of symmetries or other special properties of a spacetime, there does not appear to be any preferred choise of quasi-free Hadamard state.

Perturbation and renormalization prescriptions
The loss of Poincaré symmetry also has some major consequences for the perturbation rules and the regularization and renormalization prescriptions of a quantum field theory. To begin with, Wick’s theorem becomes ambiguous because it requires normal ordering which relies on the existence of a preferred vacuum state with respect to which the normal ordering is carried out. Furthermore, renormalization prescriptions used to define time-ordered products in Minkowski spacetime make use of momentum-space methods and/or Euclidean methods. The momentum-space methods are based on global Fourier transforms of quantities, but a global Fourier transform is a spoiled concept in curved spacetime. The Euclidean methods are based upon analytic continuation and require the ability to ‘Euclideanize’ Minkowski spacetime by the
transformation $t \to it$, something which clearly is impossible in a general spacetime. Albeit these difficulties might seem insurmountable, it has been showed that quantum field theories which are renormalizable in Minkowski spacetime are also renormalizable in a general spacetime, by using the algebraic framework [50].

### 2.1.3.3 The algebraic approach

One could see the quest for a preferred vacuum state in quantum field theory in curved spacetime like the quest for a preferred coordinate system in classical general relativity. They appear both to be equally meaningless. In general relativity this is manifestly present by formulating the theory in a geometrical way, wherein one does not have to specify a choice of coordinate system. This inspired people to search a formulation of quantum field theory that did not require to specify a choice of state (or representation) to define the theory. This lead to the algebraic approach to quantum field theory in curved spacetime [2, 46, 47] which states that the fundamental observables in quantum field theory are the local fields themselves. The algebraic approach is intimately related to axiomatic quantum field theory.

The algebraic approach makes use of the observation that the Fourier decomposition of the field (2.74) does not make sense as a definition of $\psi$ as an operator at each point $(t, \vec{x})$. In essence, the contributions from the modes at large $|\vec{k}|$ do not diminish rapidly enough with $|\vec{k}|$ for the sum to converge. However, these contributions are rapidly varying in spacetime so if we average the right hand side of (2.74) in an appropriate manner over a spacetime region, the sum will converge. This is mathematically translated by the fact that (2.74) defines $\psi$ as an 'operator valued distribution', i.e. for any smooth test function $f$ with compact support the quantity

$$\psi(f) = \int d^4x f(t, \vec{x})\psi(t, \vec{x})$$

(2.76)

is well defined by (2.74) if the integration is done prior to the summation. The algebraic approach considers particles to have no fundamental meaning in quantum field theory. It derives most results directly from $n$-point correlation functions. In calculating these correlation functions or results following from them, crucial use is made from the 'spacetime-smearing' as just described.

However we have just touched upon the algebraic approach very lightly, it has a very rigorous mathematical framework. The completion of this mathematical framework is even today a topic of current research [47]. It should also be obvious that the entire domain of quantum field theory in curved spacetime greatly extends the discussion of this section. But for the purpose of this thesis it is not necessary to go into further detail on these matters.

### 2.2 The Unruh effect

Surprisingly enough, as a first treatment of quantum field theory in curved spacetime we restrict our attention to Minkowski spacetime. The matter being treated here is nevertheless closely related to particle creation by black holes.
Although we saw in the previous section that the choice of the vacuum state is not unique in general, there is a natural vacuum state if the spacetime is static. Then, it is natural to let the positive frequency solutions have a \( t \)-dependence of the form \( e^{-i\omega t} \), where the \( \omega \) are positive constants interpreted as the energy of the particle with respect to the future-directed Killing vector field \( \partial/\partial t \). If the spacetime is globally hyperbolic and static, then this choice of positive frequency modes leads to a well-defined and natural vacuum state that preserves the time translation symmetry. This state is called the static vacuum.

Minkowski spacetime has global time-like Killing vector fields which generate time translations in various inertial frames. The sets of positive frequency modes corresponding to these Killing vectors are the same and are the usual positive frequency modes proportional to \( e^{-ikt} \) with \( k_0 > 0 \), where \( t \) is the time parameter with respect to one of the inertial frames. Thus, all these Killing vector fields define the same vacuum state.

Now, consider the boost Killing vector field

\[
b = z \frac{\partial}{\partial t} + t \frac{\partial}{\partial z},
\]

(2.77)

where \( z \) is one of the spatial coordinates. In the region defined by \( |t| < z \) in Minkowski spacetime, \( b \) is time-like and future-directed. Hence, this region called the right Rindler wedge is a static spacetime with \( b \) being the generator of time translations. Thus, one can define the corresponding static vacuum state. However, this vacuum state is not the same as the state obtained by restricting the usual Minkowski vacuum to this region. This observation is crucial in understanding the Unruh effect, as will be explained in the next subsections.

### 2.2.1 Rindler spacetime

Minkowski spacetime with the metric

\[
ds^2 = dt^2 - dx^2 - dy^2 - dz^2
\]

(2.78)

is of course a static globally hyperbolic spacetime. It can be divided in four distinct parts:

1) \( |t| < z \): right Rindler wedge, is a static globally hyperbolic spacetime
2) \( |t| < -z \): left Rindler wedge, also a static globally hyperbolic spacetime
3) \( t > |z| \): expanding degenerate Kasner universe, globally hyperbolic but not static
4) \( t < -|z| \): contracting degenerate Kasner universe, globally hyperbolic but also not static

These regions are shown on figure 2.1. The curves with arrows are the integral curves of the boost Killing vector field \( b = z(\partial/\partial t) + t(\partial/\partial z) \). The direction of increasing \( U = t - z \) and that of increasing \( V = t + z \) are also indicated.

Minkowski spacetime is invariant under the boost

\[
t \rightarrow t \cosh \beta + z \sinh \beta
\]

(2.79)

\[
z \rightarrow t \sinh \beta + z \cosh \beta
\]

(2.80)
where \( \beta \) is the boost parameter. That these transformations are generated by the Killing vector field \( b \) can be seen as follows. The integral curves of \( b \) are solutions of the set of coupled first order differential equations

\[
\frac{dt}{d\lambda} = z, \\
\frac{dz}{d\lambda} = t, 
\]

with \( \lambda \) an arbitrary parameter along the integral curve. This set of coupled first order equations can be rewritten as a decoupled set of second order equations

\[
\frac{d^2 t}{d\lambda^2} = t, \\
\frac{d^2 z}{d\lambda^2} = z. 
\]

The most general solution of the first second order differential equation is given by

\[
t = a \cosh \lambda + b \sinh \lambda.
\]

Applying the appropriate boundary conditions and taking \( \lambda = \beta \) results in (2.79). (2.80) is analogous.

The boost invariance of Minkowski spacetime motivates the following coordinate transformation

\[
t = \rho \sinh \eta, \\
z = \rho \cosh \eta,
\]
where \( \rho \) and \( \eta \) take any real value. Then, the Killing vector field \( b \) is

\[
b = \rho \cosh \eta \left( \frac{\partial}{\partial t} \frac{\partial}{\partial \rho} + \frac{\partial}{\partial \eta} \frac{\partial}{\partial \eta} \right) + \rho \sinh \eta \left( \frac{\partial}{\partial z} \frac{\partial}{\partial \rho} + \frac{\partial}{\partial \eta} \frac{\partial}{\partial \eta} \right) = \rho \cosh \eta \left( -\sinh \eta \frac{\partial}{\partial \rho} + \cosh \eta \frac{\partial}{\partial \eta} \right) + \rho \sinh \eta \left( \cosh \eta \frac{\partial}{\partial \rho} - \sinh \eta \frac{\partial}{\partial \eta} \right)
\]

(2.85)

(2.86)

and the metric takes the form

\[
ds^2 = \rho^2 \, d\eta^2 - d\rho^2 - dx^2 - dy^2,
\]

(2.88)

which is independent of \( \eta \) as expected. The world lines with fixed values of \( \rho, x \) and \( y \) are the trajectories of the boost transformation of (2.79) and (2.80). Each world line has a constant proper acceleration given by \( \rho^{-1} \) = constant. This can be seen using the general formula for the proper acceleration four-vector on an orbit of a vector field \( \xi \) as used in section 1.6 of chapter 1

\[
a^\mu = \frac{\xi^\nu \nabla_\nu \xi^\mu}{\xi^\nu \xi_\nu}.
\]

(2.89)

Here, one has \( \xi = b = \partial/\partial \eta \). Using the metric (2.88), one obtains

\[
\xi^\nu \xi_\nu = \rho^2 \quad (2.90)
\]

\[
\xi^\nu \nabla_\nu \xi^\mu = \Gamma^\mu_{\eta \eta} = -\frac{1}{2} g^{\mu \rho} \partial_\rho g_{\eta \eta} = -\rho g^{\mu \rho}.
\]

(2.91)

Because (2.88) is diagonal, one gets for the proper acceleration four-vector by combining (2.90), (2.91) and (2.89)

\[
a^\mu = \left( 0, \frac{1}{\rho}, 0, 0 \right) .
\]

(2.92)

So the proper acceleration becomes

\[
a = \sqrt{-a^\mu a_\mu} = \frac{1}{\rho} .
\]

(2.93)

The coordinates \( (\eta, \rho, x, y) \) cover only the regions with \( z^2 > t^2 \), i.e. the left and right Rindler wedges, as can readily be seen from (2.84).

The Killing vector field \( b \) becomes null on the hypersurfaces \( t = \pm z \) dividing Minkowski spacetime into the four regions. It also clearly is orthogonal to the these hypersurfaces, so they are Killing horizons of \( b \). To give a physical interpretation to these horizons, one can use the coordinates \( (\rho, \eta) \) of (2.84). The horizons are given by \( t^2 - z^2 = 0 \), which in the \( (\rho, \eta) \)-coordinates becomes \( \rho = 0 \). But from (2.93) it is clear that when \( \rho \to 0 \), \( a \to \infty \). So the Killing horizons at \( \rho = 0 \) are called acceleration horizons.
To discuss quantum fields in the right Rindler wedge, it is convenient to make a further coordinate transformation

\[ \rho = \frac{1}{a} e^{\alpha \chi} \]  
\[ \eta = a \tau , \]  

or in terms of the original variables \( t \) and \( z \)

\[ t = \frac{1}{a} e^{\alpha \chi} \sinh a \tau \]  
\[ z = \frac{1}{a} e^{\alpha \chi} \cosh a \tau , \]

where \( a \) is a positive constant. Then, the metric takes the form

\[ ds^2 = e^{2\alpha \chi} (d\tau^2 - d\chi^2) - dx^2 - dy^2 . \]

This coordinate system will be useful because the world line with \( \chi = 0 \) has a constant acceleration of \( a \). The coordinates \((\bar{\tau}, \bar{\chi})\) for the left Rindler wedge are given by

\[ t = \frac{1}{a} e^{\alpha \bar{\chi}} \sinh a \bar{\tau} \]  
\[ z = -\frac{1}{a} e^{\alpha \bar{\chi}} \cosh a \bar{\tau} , \]

In the next subsection it will be shown that the usual vacuum state for quantum field theory in Minkowski spacetime restricted to the right Rindler wedge is a thermal state with \( \tau \) playing the role of time, and similarly for the left Rindler wedge.

### 2.2.2 Accelerating observers and the thermal bath

The two-dimensional massless scalar field in Minkowski spacetime is problematic because of infrared divergences [51]. Nevertheless, this theory is a very good model for explaining the Unruh effect, and it is not necessary to deal with the infrared divergences for this purpose. It also turns out that the Unruh effect in scalar field theory in higher dimensions can be derived in essentially the same manner as in this model. So it captures all the necessary physics to be used in the next sections of this chapter. The model is presented in analogy to [52].

The massless scalar field in two dimensions \( \psi(t, z) \) satisfies the Klein-Gordon equation

\[ \left( \frac{\partial^2}{\partial t^2} - \frac{\partial^2}{\partial z^2} \right) \psi = 0 . \]

This field can be expanded as

\[ \psi(t, z) = \int_0^{\infty} \frac{dk}{\sqrt{4\pi k}} \left( b_{-k} e^{-ik(t-z)} + b_k e^{-ik(t+z)} + b_{-k}^\dagger e^{ik(t-z)} + b_k^\dagger e^{ik(t+z)} \right) . \]
The annihilation and creation operators satisfy

\[ [b_{\pm k}, b_{\pm k'}^\dagger] = \delta(k - k'), \quad (2.103) \]

with all other commutators vanishing. By using the definitions

\[ U = t - z \quad (2.104) \]
\[ V = t + z \quad (2.105) \]

one can write

\[ \psi(t, z) = \psi_-(U) + \psi_+(V), \quad (2.106) \]

where

\[ \psi_+(V) = \int_0^\infty dk [b_k f_k(V) + b_k^\dagger f_k^\ast(V)], \quad (2.107) \]

with

\[ f_k(V) = \frac{e^{-ikV}}{\sqrt{4\pi k}}, \quad (2.108) \]

and similarly for \( \psi_-(U) \). Since the left and right-moving sectors of the field, i.e. \( \psi_+(V) \) and \( \psi_-(U) \), do not interact with one another, only the left moving sector \( \psi_+(V) \) is discussed. Thus, the Unruh effect for the theory consisting only of the left-moving sector will be treated. The Minkowski vacuum state \( |0\rangle_M \) is defined by

\[ b_k |0\rangle_M = 0, \quad (2.109) \]

for all \( k \).

Using the metric in the right Rindler wedge given by (2.98), one finds a field equation of the same form as (2.101)

\[ \left( \frac{\partial^2}{\partial \tau^2} - \frac{\partial^2}{\partial \chi^2} \right) \psi = 0. \quad (2.110) \]

The solutions to this differential equation can be classified again into left and right-moving modes which depend only on

\[ v = \tau + \omega \quad (2.111) \]
\[ u = \tau - \omega, \quad (2.112) \]

respectively. These variables are related to \( U \) and \( V \) as follows

\[ U = t - z = -\frac{1}{a} e^{-au} \quad (2.113) \]
\[ V = t + z = \frac{1}{a} e^{au}. \quad (2.114) \]

The Lagrangian density leading to the Klein-Gordon equation is invariant under the coordinate transformation \((t, z) \rightarrow (\tau, \chi)\). As a result, going through the quantization procedure, one finds exactly the same theory as in the whole of Minkowski spacetime with \((t, z) \) replaced by \((\tau, \chi)\).
Thus, one has for $0 < V$

$$
\psi_+ (V) = \int_0^\infty d\omega \left[ a^R_\omega g_\omega (v) + a^{R\dagger}_\omega g_\omega ^* (v) \right],
$$

(2.115)

where

$$
g_\omega (v) = \frac{e^{-i\omega v}}{\sqrt{4\pi\omega}},
$$

(2.116)

and where

$$
[a^R_\omega , a^{R\dagger}_\omega ] = \delta (\omega - \omega'),
$$

(2.117)

with all other commutators vanishing. Notice that the functions $g_\omega (v)$ are eigenfunctions of the boost generator $\partial/\partial \tau$.

The field $\psi_+ (V)$ can be expressed in the left Rindler wedge with the condition $V < 0 < U$, by using the left Rindler coordinates $(\bar{\tau}, \bar{\chi})$ of (2.100). Defining $\bar{v} = \bar{\tau} - \bar{\chi}$, one obtains equations (2.115) - (2.117) with $v$ replaced by $\bar{v}$ and with the annihilation and creation operators $a^R_\omega$ and $a^{R\dagger}_\omega$ replaced by a new set of operators $a^L_\omega$ and $a^{L\dagger}_\omega$. The variable $\bar{v}$ is related to $V$ by

$$
V = -\frac{1}{a} e^{-a\bar{v}}.
$$

(2.118)

The static vacuum state in the left and right Rindler wedges, the Rindler vacuum state $|0\rangle_R$, is defined by

$$
a^R_\omega |0\rangle_R = a^L_\omega |0\rangle_R = 0,
$$

(2.119)

for all $\omega$.

To understand the Unruh effect, one needs to find the Bogoliubov coefficients $\alpha^R_{\omega k}$, $\beta^R_{\omega k}$, $\alpha^L_{\omega k}$ and $\beta^L_{\omega k}$, where

$$
\theta (V) g_\omega (v) = \int_0^\infty \frac{dk}{\sqrt{4\pi k}} (\alpha^R_{\omega k} e^{-ikV} + \beta^R_{\omega k} e^{ikV}),
$$

(2.120)

and

$$
\theta (-V) g_\omega (\bar{v}) = \int_0^\infty \frac{dk}{\sqrt{4\pi k}} (\alpha^L_{\omega k} e^{-ikV} + \beta^L_{\omega k} e^{ikV}),
$$

(2.121)

where $\theta (x)$ is the Heaviside function. To find $\alpha^R_{\omega k}$, one multiplies (2.120) by $e^{ikV}/2\pi$ with $k > 0$ and integrates over $V$. Thus, with (2.116), one finds

$$
\alpha^R_{\omega k} = \frac{1}{\sqrt{4\pi k}} \int_0^\infty \frac{dV}{2\pi} g_\omega (V) e^{ikV}
$$

$$
= \frac{k}{\omega} \int_0^\infty \frac{dV}{2\pi} (aV)^{-i\omega/a} e^{ikV}.
$$

(2.122)
Now introduce a cut-off for this integral for large \( V \) by letting \( V \to V + i\epsilon, \epsilon \to 0+ \). Then, changing the integration path to the positive imaginary axis by putting \( V = ix/k \), one finds

\[
\alpha^R_{\omega k} = \frac{ie^{\pi\omega/2a}}{2\pi\sqrt{\omega k}} \left( \frac{a}{k} \right)^{-i\omega/a} \int_0^{\infty} \frac{dx}{2\pi} e^{-ix/a} e^{-x} dx
\]

\[
= \frac{ie^{\pi\omega/2a}}{2\pi\sqrt{\omega k}} \left( \frac{a}{k} \right)^{-i\omega/a} \Gamma(1 - i\omega/a), \tag{2.123}
\]

where \( \Gamma(x) \) represents the gamma-function.

To find the coefficients \( \beta^R_{\omega k} \), one replaces \( e^{ikV} \) in (2.122) by \( e^{-ikV} \). Then, the appropriate substitution is \( V = -ix/k \). As a result, one obtains

\[
\beta^R_{\omega k} = -\frac{ie^{-\pi\omega/2a}}{2\pi\sqrt{\omega k}} \left( \frac{a}{k} \right)^{-i\omega/a} \Gamma(1 + i\omega/a). \tag{2.124}
\]

A similar calculation leads to

\[
\alpha^L_{\omega k} = -\frac{ie^{\pi\omega/2a}}{2\pi\sqrt{\omega k}} \left( \frac{a}{k} \right)^{i\omega/a} \Gamma(1 + i\omega/a), \tag{2.125}
\]

\[
\beta^L_{\omega k} = \frac{ie^{-\pi\omega/2a}}{2\pi\sqrt{\omega k}} \left( \frac{a}{k} \right)^{i\omega/a} \Gamma(1 + i\omega/a). \tag{2.126}
\]

So one finds following crucial relations for the derivation of the Unruh effect

\[
\beta^L_{\omega k} = -e^{-\pi\omega/a}\alpha^R_{\omega k} \tag{2.127}
\]

\[
\beta^R_{\omega k} = -e^{-\pi\omega/a}\alpha^L_{\omega k}. \tag{2.128}
\]

By substituting these relations in (2.120) and (2.121), one finds that the following functions are linear combinations of positive-frequency modes \( e^{-ikV} \) in Minkowski spacetime

\[
G_\omega(V) = \theta(V)g_\omega(v) + \theta(-V)e^{-\pi\omega/a}g^*_\omega(\bar{v}) \tag{2.129}
\]

\[
\tilde{G}_\omega(V) = \theta(-V)g_\omega(\bar{v}) + \theta(V)e^{-\pi\omega/a}g^*_\omega(v). \tag{2.130}
\]

One can show that these functions are purely positive-frequency solutions in Minkowski spacetime by an analyticity argument as well: since a positive-frequency solution is analytic in the lower half plane of the complex \( V \)-plane, the solution \( g_\omega(v) = (4\pi\omega)^{-1/2}V^{-i\omega/a} \) with \( V < 0 \) should be continued to the negative real line avoiding the singularity at \( V = 0 \) around a small circle in the lower half plane, thus leading to \( (4\pi\omega)^{-1/2}e^{-\pi\omega/a}(-V)^{-i\omega/a} \) for \( V < 0 \).

Equations (2.129) and (2.130) can be inverted to

\[
\theta(V)g_\omega(v) \propto G_\omega(V) - e^{-\pi\omega/a}\tilde{G}^*_\omega(V) \tag{2.131}
\]

\[
\theta(-V)g_\omega(\bar{v}) \propto \tilde{G}_\omega(V) - e^{-\pi\omega/a}G^*_\omega(V). \tag{2.132}
\]

By substituting these equations in

\[
\psi_\pm(V) = \int_0^{\infty} d\omega \left[ \theta(V)\{a^R_\omega g_\omega(v) + a^R_{\omega\dagger}g^*_\omega(v)\} + \theta(-V)\{a^L_\omega g_\omega(\bar{v}) + a^L_{\omega\dagger}g^*_\omega(\bar{v})\} \right], \tag{2.133}
\]
one finds that the integrand here is proportional to

\[
G_\omega(V) [a^R_\omega e^{-\pi\omega/a} a^L_\omega] + \bar{G}_\omega(V) [a^L_\omega e^{-\pi\omega/a} a^R_\omega] + \text{h.c.} \quad (2.134)
\]

Because it was derived that the functions \(G_\omega(V)\) and \(\bar{G}_\omega(V)\) are positive-frequency solutions with respect to the usual time translation in Minkowski spacetime, one has

\[
(a^R_\omega e^{-\pi\omega/a} a^L_\omega) |0\rangle_M = 0 \quad (2.135)
\]

\[
(a^L_\omega e^{-\pi\omega/a} a^R_\omega) |0\rangle_M = 0 \quad (2.136)
\]

These relations uniquely determine the Minkowski vacuum state \(|0\rangle_M\) as will be explained below.

To explain how the state \(|0\rangle_M\) is formally expressed in the Fock space on the Rindler vacuum state \(|0\rangle_R\) and to show that the state \(|0\rangle_M\) is a thermal state when it is probed only in the right (or left) Rindler wedge, one uses the approximation where the Rindler energy levels \(\omega\) are discrete. The rigorous treatment would be to do the calculation in a box and then let the volume of the box go to infinity. But here, a physical and straightforward version of this procedure will be used, not worrying too much about technical restrictions. To do so, write \(\omega_i\) instead of \(\omega\) and let

\[
[a^R_i, a^R_j] = [a^L_i, a^L_j] = \delta_{ij},
\]

with all other commutators vanishing. Using the discrete version of (2.135) and the commutators (2.137), one finds

\[
\langle 0_M | a^R_i a^R_i | 0_M \rangle = e^{-2\pi\omega_i/a} \langle 0_M | a^L_i a^L_i | 0_M \rangle + e^{-2\pi\omega_i/a}. \quad (2.138)
\]

The same relation with \(a^R_i\) and \(a^R_i\) replaced by \(a^L_i\) and \(a^L_i\), respectively and vice versa, can be found using (2.137). By solving these two relations simultaneously, one finds

\[
\langle 0_M | a^R_i a^R_i | 0_M \rangle = \langle 0_M | a^L_i a^L_i | 0_M \rangle = \frac{1}{e^{2\pi\omega_i/a} - 1}. \quad (2.139)
\]

Hence, the expectation value of the Rindler-particle number is that of a Bose-Einstein particle in a thermal bath of temperature \(T = a/2\pi\). Therefore, a uniformly accelerating observer in Minkowski spacetime will detect a thermal bath of particles, which is the Unruh effect.

Equation (4.42) can be expressed without discretization. Define

\[
a^R_f = \int_0^\infty d\omega f(\omega) a^R_\omega, \quad (2.141)
\]

with

\[
\int_0^\infty d\omega |f(\omega)|^2 = 1. \quad (2.142)
\]

Then

\[
\langle 0_M | a^R_f a^R_f | 0_M \rangle = \int_0^\infty d\omega \frac{|f(\omega)|^2}{e^{2\pi\omega/a} - 1}. \quad (2.143)
\]

Exactly the same formula applies to the left Rindler number operator.
2.3 Particle creation by black holes

As mentioned in chapter 1, it is possible to classically extract energy out of a black hole (the Penrose process) and to have induced emission in the case of rotating black holes (Superradiance). Some experience with quantum mechanics learns that in circumstances where there is induced emission, there also is spontaneous emission. So when the development of quantum field theory in curved spacetime arose in the mid-sixties, people tried to find a quantum mechanical mechanism for this spontaneous emission – i.e. spontaneous particle creation from the vacuum.

First, it should be noted that there is nothing wrong with using quantum field theory in a black hole background as long as one stays far enough from the singularity. As mentioned at the beginning of this chapter, quantum field theory in a curved spacetime is known to be only an approximation to a better and yet to be found physical theory of quantum gravity, but one that is reliable when avoiding Planck-scale phenomena. In a Schwarzschild spacetime the components of the Riemann curvature tensor are of order

$$R(\text{Horizon}) \sim \frac{1}{M^2 G^2}$$

at the horizon. For a large mass black hole they are typically very small. So, however an event horizon is an intrinsically general relativistic phenomenon, there is no danger in using quantum field theory in that region because there are no violent gravitational effects there.

A first notion of particle creation by black holes was made in [53], where it was pointed out that a Reissner-Nordström black hole of sufficiently small mass has an electric field that would create electron-positron pairs through the Heisenberg-Euler-Schwinger process. This process was worked out in complete detail in [54]. Further progress was made on particle creation by rotating black holes by Starobinsky [55] and Unruh [56]. The fact that spontaneous particle creation occurs near rotating black holes did not cause much surprise or excitement. The effect is negligible small for macroscopic black holes such as those that would be produced by the collapse of rotating stars. So, unless tiny black holes were produced in the early universe, the effect is not of astrophysical importance. While it is an interesting phenomenon as a matter of principle, it was not surprising or unexpected in view of the ability to extract energy from a rotating black hole by classical processes.

Unruh did the calculation of particle creation by a rotating black hole in the idealized spacetime representing the stationary final state of the black hole. This spacetime necessarily contains also a "time-reversed black hole", i.e. a white hole, although white holes are not expected to occur in nature (something which is undoubtedly closely related to the second law of thermodynamics). A white hole is a region of spacetime to which nothing can enter, starting from infinity. So for Unruh to get a result, initial conditions had to be imposed on the white hole horizon, expressing that no particles are emerging from the white hole. In this calculation, a seemingly natural choice of the "in" vacuum state on the white hole horizon was made. But it was not obvious that this choice was physically correct.

And then, in 1974, Hawking realized in his now classic papers [57, 58] that the difficulty of
Unruh’s calculation could be overcome by considering the more physically relevant spacetime describing gravitational collapse to a black hole rather than the idealized spacetime describing a stationary black hole (and white hole). Going through the calculation, he found that the results were significantly altered from the results obtained by Unruh. Remarkably, Hawking found that even for a non-rotating black hole, particle creation occurs and produces a steady flux of particles to infinity at late times. And even more remarkably he found that, for a non-rotating black hole, the spectrum of particles emitted to infinity at late times is precisely thermal, at a temperature \( T = \kappa/2\pi \), where \( \kappa \) denotes the surface gravity of the black hole.

The implications of Hawking’s results were enormous. They establish that black holes are perfect black (or actually gray) bodies in the thermodynamic sense at non-zero temperature. This tied in perfectly with the mathematical analogy that had previously been discovered between certain laws of black hole physics and the laws of thermodynamics in chapter 1, giving clear evidence that the similarity of these laws is much more than a mere mathematical analogy.

In the following section Hawking’s original results are given for the Schwarzschild black hole and the rotating Kerr black hole, as derived in [58].

### 2.3.1 Original derivation of the Hawking radiation

The derivation of the Hawking flux takes place in the spacetime of a gravitational collapse as discussed in Chapter 1. This means that at early times the mass that is later to form the black hole is widely dispersed and of sufficiently low density so that the early part of the spacetime is nearly flat. The thermal flux of particles is caused by the formation of an event horizon if the matter collapses. In the calculation the backreaction of particle creation on the metric is neglected. The flux of particles coming from the black hole will make its mass decrease and Schwarzschild radius shrink. But this process is expected to take place sufficiently slow so that when considering particle creation during an amount of time that is small enough, the metric can be taken time-independent. However, after a very long time, the black hole will become explosive because the surface gravity and temperature increase during the shrinking process. At some point, the surface gravity will be so big that the quantum field description is no longer valid. So there are still many open questions about the end state of black hole evaporation.

The field used to derive the Hawking radiation is a massless Hermitian scalar field, satisfying the generally covariant wave equation \( \Box \psi = 0 \), or

\[
(-g)^{-1/2} \partial_\mu [(-g)^{1/2} g^{\mu\nu} \partial_\nu \psi] = 0,
\]

(2.144)

because the determinant of the Schwarzschild metric is negative. The created particles observed at late times are created at a short affine distance from the event horizon. Their spectrum is not affected by the regions, such as that inside the collapsing body, where the metric is not stationary. In the spacetime of a body that collapses to form a Schwarzschild of Kerr black hole, one can write the field in the entire spacetime in the form

\[
\psi = \int d\omega \left( a_\omega f_\omega + a^\dagger_\omega f^*_\omega \right),
\]

(2.145)
where the $f_\omega$ and $f_\omega^*$ are a complete set of solutions of the field equation (2.144), with normalization

$$(f_{\omega_1}, f_{\omega_2}) = \delta(\omega_1 - \omega_2),$$

(2.146)

with the scalar product is defined as in the previous section. The $a_\omega$ are time-independent operators. Then the canonical commutation relations of the field $\psi$ imply that the $a_\omega$ are annihilators and $a_\omega^\dagger$ are creation operators obeying

$$[a_{\omega_1}, a_{\omega_2}^\dagger] = \delta(\omega_1 - \omega_2)$$

(2.147)

and

$$[a_{\omega_1}, a_{\omega_2}] = [a_{\omega_1}^\dagger, a_{\omega_2}^\dagger] = 0$$

(2.148)

The physical interpretation of the $a_\omega$ depends on the choice of the complete set of solutions $f_\omega$.

Far outside the collapsing body at early times, the definition of the physical particles that would be detected by inertial observers, or equivalently of positive frequency solutions of the field equation (2.144) is unambiguous. Let the $f_\omega$ be chosen such that at early times and large distances they form a complete set of incoming positive frequency solutions of energy $\omega$. Their asymptotic form on past null infinity, $I^-$, is

$$f_\omega \sim \omega^{-1/2} r^{-1} \exp(-i\omega v) S(\theta, \phi),$$

(2.149)

where discrete quantum numbers $(l, m)$ are suppressed, and $v = t + r$ is the incoming null coordinate at $I^-$. The factor $\omega^{-1/2}$ is required by the normalization of the scalar product. In that case, the operators $a_\omega$ are annihilators of particles on $I^-$. At late times, the situation is different because a black hole event horizon has formed. To define a unique solution of the field equation (2.144) outside the black hole, boundary conditions have to be given both on the event horizon and on future null infinity, $I^+$. This feature is not a mathematical detail, but really a cornerstone on which the entire derivation is built. It is again an example of how the role of boundary conditions in quantum field theory in curved spacetime cannot be underestimated.

On $I^+$, just as on $I^-$, the definition of positive frequency solutions is unambiguous. Let the $p_\omega$ be the solutions of the field equation (2.144) that have zero Cauchy data on the event horizon and are asymptotically outgoing and positive frequency at $I^+$. Assume that $p_\omega$ and $p_\omega^*$ form a complete set of solutions on $I^+$, satisfying the normalization condition

$$(p_{\omega_1}, p_{\omega_2}) = \delta(\omega_1 - \omega_2).$$

(2.150)

The asymptotic form of $p_\omega$ on $I^+$ is

$$p_\omega \sim \omega^{-1/2} r^{-1} \exp(-i\omega u) S(\theta, \phi),$$

(2.151)

where again the quantum numbers $(l, m)$ have been suppressed, and $u = t - r$ is the outgoing null coordinate at $I^+$. A wave packet formed by a superposition of the $p_\omega$ is outgoing and localized at large $r$ at late times.
The most general solution of the wave equation will have a part that is incoming at the event horizon at late times. Therefore, another set of solutions $q\omega$ must be introduced such that a superposition of them at late times is localized near the event horizon and has zero Cauchy data on $I^+$. The precise form of the $q\omega$ will not affect observations on $I^+$, since those observations can only depend on the $p\omega$. Let the $q\omega$ and $q^{\ast}\omega$ form a complete set on the horizon with normalization

$$
(q_{\omega_1}, q_{\omega_2}) = \delta(\omega_1 - \omega_2) .
$$

(2.152)

Since wave packets formed from the $p\omega$ and the $q\omega$ are in disjoint regions at late times, their conserved scalar product must vanish:

$$
(q_{\omega_1}, p_{\omega_2}) = 0 .
$$

(2.153)

One also has

$$(q_{\omega_1}, q^{\ast}_{\omega_2}) = 0$$

$$(q_{\omega_1}, p^{\ast}_{\omega_2}) = 0$$

$$(p_{\omega_1}, p^{\ast}_{\omega_2}) = 0 .$$

The field $\psi$ can now be expanded in the entire spacetime as

$$
\psi = \int d\omega \left\{ b_{\omega} p_{\omega} + c_{\omega} q_{\omega} + b^{\dagger}_{\omega} p^{\ast}_{\omega} + c^{\dagger}_{\omega} q^{\ast}_{\omega} \right\} .
$$

(2.154)

Again using the canonical commutation relations for the field, gives

$$
[b_{\omega_1}, b^{\dagger}_{\omega_2}] = \delta(\omega_1 - \omega_2)
$$

$$
[c_{\omega_1}, c^{\dagger}_{\omega_2}] = \delta(\omega_1 - \omega_2) ,
$$

(2.155)

with all other commutators between $b_{\omega_1}$ and $c_{\omega_2}$ and their Hermitian conjugates vanishing.

The derivation is done in the Heisenberg picture, so the state vector is independent of time. Let this state vector, $|0\rangle$, be chosen to have no particles of the field incoming from $I^-$. Thus, $|0\rangle$ is annihilated by the $a_{\omega}$ corresponding to particles incoming from $I^-$:

$$
a_{\omega}|0\rangle = 0 , \quad \forall \omega .
$$

(2.156)

As in the cosmological model of the previous section, the spectrum of the created particles is determined by the coefficients of the Bogoliubov transformation relating the annihilation operators at early times to the annihilation and creation operators at late times. It is in that spirit that the steps below are made.

The $f_{\omega}$ and $f^{\ast}_{\omega}$ are a complete set for expanding any solution of the field equation, so one can write

$$
p_{\omega} = \int d\omega' (\alpha_{\omega\omega'} f_{\omega'} + \beta_{\omega\omega'} f^{\ast}_{\omega'}) ,
$$

(2.157)
where $\alpha_{\omega\omega'}$ and $\beta_{\omega\omega'}$ are complex numbers, independent of the coordinates. From (2.150), (2.153) and (2.154) it follows that

$$b_\omega = (p_\omega, \psi).$$

(2.158)

Then, expressing $\psi$ and $p_\omega$ in terms of $f_{\omega'}$ and $f_{\omega'}^*$ according to (2.145) and (2.157), one gets

$$b_\omega = \int d\omega'(\alpha_{\omega\omega'}^* a_{\omega'} - \beta_{\omega\omega'}^* a_{\omega'}^\dagger),$$

(2.159)

where it was used that $(f_{\omega'}^*, f_{\omega''}^*) = -\delta(\omega' - \omega'')$. Furthermore, using the expansion of $p_\omega$ (2.157) it follows that

$$(p_{\omega_1}, p_{\omega_2}) = \int d\omega' (\alpha_{\omega_1\omega'}^* \alpha_{\omega_2\omega'} - \beta_{\omega_1\omega'}^* \beta_{\omega_2\omega'}).$$

(2.160)

The coefficients in the expansion of $p_\omega$ (2.157) can be expressed as

$$\beta_{\omega\omega'} = -(f_{\omega}', p_\omega)$$

(2.161)

$$\alpha_{\omega\omega'} = (f_{\omega}', p_\omega).$$

(2.162)

Now all the necessary general concepts are introduced. First, the Hawking flux will be calculated explicitly for a Schwarzschild black hole, and then for a rotating Kerr black hole.

### 2.3.1.1 The Schwarzschild black hole

The aim of this section is to calculate the coefficients $\alpha_{\omega\omega'}$ and $\beta_{\omega\omega'}$, from which the spectrum of the created particles will follow, for a non-rotating Schwarzschild black hole. The relevant geodesics were discussed in the first chapter. In figure 2.2 the Penrose diagram of the spacetime of a gravitational collapse is shown.

![Figure 2.2: The Penrose diagram for matter collapsing to a Schwarzschild black hole.](image)

A wave packet from superposition of the $p_\omega$ for a range of frequencies near a given value $\omega$ can be constructed. The coefficients in the superposition can be chosen so that the outgoing wave packet approaches $\mathcal{I}^+$ along a null geodesic characterized by a large constant value of $u$ (i.e. at late times). The components of this wave packet can expressed in terms of the $f_{\omega'}$ and $f_{\omega'}^*$ by
means of (2.157). Now imagine this wave packet propagating backward in time. Part of it will be scattered back toward infinity by the curved spacetime, and will reach $I^-$ as a superposition of the $f_{\omega'}$ with frequencies near the original frequency $\omega$. Another part of the wave packet will pass through the center of the collapsing body (ignoring interaction with the matter of the collapsing body, or assuming that the interaction is negligible at sufficiently high frequencies) and reach $I^-$ as a superposition of the $f_{\omega'}$ and $f^*_{\omega'}$ having highly blueshifted values of $\omega' \gg \omega$. This is because when particles leave the near-event horizon region to escape to infinity, they get heavily redshifted. So a particle being present at large distances and large times, travelling back in time towards the horizon, will then undergo the reverse process and get heavily blueshifted. And because this process takes place in the spacetime of a gravitational collapse, no black hole is present at early times. So when the particle propagates even further back in time, going to $I^-$, no redshift (or a certainly smaller redshift) occurs due to the absence of the black hole. Therefore, the $p_\omega$ in this latter part of the wave packet can be expressed in terms of the $f_{\omega'}$ and $f^*_{\omega'}$ with coefficients $\alpha_{\omega\omega'}$ and $\beta_{\omega\omega'}$ having $\omega' \gg \omega$. Furthermore, the relevant values of $\omega'$ become arbitrarily large at sufficiently late times (i.e. as $u \to \infty$) because in the limiting case, namely a particle originating from the black hole horizon, the redshift is infinite. Thus, the late time spectrum of outgoing particles is determined by the asymptotic form of the coefficients for arbitrarily large $\omega'$. It is here that it appears essential to use a gravitational collapse spacetime in the derivation of the Hawking flux. It is clear that the entire spacetime is important in the process. So an idealized, stationary black hole spacetime, as used by Unruh, could never yield the same results.

To determine these coefficients, one traces the latter part of $p_\omega$, the one going through the collapsing body, back in time along an outgoing geodesic having a very large value of $u$. The geodesic passes through the center of the collapsing body just before the event horizon has formed, and emerges as an incoming geodesic characterized by a value of $v$ close to $v_0$ as can be seen on figure 2.2. The value of $v$ at which the packet reaches $I^-$ is related to the value of $u$ that it had at $I^+$, by

$$u(v) = -4MG \ln \left( \frac{v_0 - v}{K} \right),$$

(2.163)

as was derived in section 1.9. Here $K$ is a positive constant characterizing the affine parametrization of the geodesic when it is near $I^+$ and $I^-$. The asymptotic form of $p_\omega$ near $I^+$ is already given by (2.151). The location of the center of this wave packet formed from $p_\omega$ with a small range of frequencies near the value of $\omega$ is determined by the principle of stationary phase. It follows that at early times, the components $p_\omega$ forming the part of the wave packet that passes back through the collapsing body and reaches $I^-$ at $v$ have (to within a normalization constant) the form on $I^-$

$$p_\omega \sim \omega^{-1/2} r^{-1} \exp(-i\omega u(v)) S(\theta, \phi),$$

(2.164)

with $u(v)$ given by (2.163) and $v < v_0$, because otherwise the wave packet would end up in the black hole. The $f_{\omega'}$ in the expansion of $p_\omega$ (2.157) have an asymptotic form near $I^-$ given by (2.167) with $v < v_0$, because this part of the wave packet cannot reach $I^-$ at $v > v_0$. 


Using these early time asymptotic forms for $p_\omega$ and $f_\omega'$, one can show with Fourier’s theorem that

$$\alpha_\omega' = C \int_{-\infty}^{v_0} dv \left( \frac{\omega'}{\omega} \right)^{1/2} e^{i\omega'v} e^{-i\omega u(v)}$$

(2.165)

$$\beta_\omega' = C \int_{-\infty}^{v_0} dv \left( \frac{\omega'}{\omega} \right)^{1/2} e^{-i\omega'v} e^{-i\omega u(v)},$$

(2.166)

where $C$ is a constant. Now substituting (2.163) for $u(v)$ and introducing the new variable $s \equiv v_0 - v$ in the expression for $\alpha_\omega'$ and $s \equiv v - v_0$ in the expression for $\beta_\omega'$ one gets

$$\alpha_\omega' = C \int_0^\infty ds \left( \frac{\omega'}{\omega} \right)^{1/2} e^{-i\omega's} e^{i\omega'v_0} \exp[i\omega 4MG \ln \left( \frac{s}{K} \right)]$$

(2.167)

$$\beta_\omega' = C \int_0^{-\infty} ds \left( \frac{\omega'}{\omega} \right)^{1/2} e^{i\omega's} e^{-i\omega'v_0} \exp[i\omega 4MG \ln \left( -\frac{s}{K} \right)].$$

(2.168)

In the equation for $\alpha_\omega'$ (2.167), the contour of integration along the real axis from 0 to $\infty$ can be closed by a a quarter circle at infinity and by the contour along the imaginary axis from $-\infty$ to 0. Because there are no poles in the enclosed quadrant of the complex plane, and the integrand vanishes at infinity, the integral from 0 to $\infty$ along the real $s$-axis equals the integral from 0 to $-i\infty$ along the imaginary $s$-axis.

Similarly, in the expression for $\beta_\omega'$ (2.168), the integral along the real axis in the complex $s$ plane from $-\infty$ to 0 can be joined by a quarter circle at infinity to the contour along the imaginary $s$-axis from $-i\infty$ to 0, thereby resulting in a closed contour. One gets that the integral from $-\infty$ to 0 equals the integral from $-i\infty$ to 0, for the same reasons as before. Therefore, putting $s \equiv is'$, it follows

$$\alpha_\omega' = -iC \int_{-\infty}^0 ds' \left( \frac{\omega'}{\omega} \right)^{1/2} e^{-i\omega's'} e^{i\omega'v_0} \exp[i\omega 4MG \ln \left( \frac{is'}{K} \right)]$$

(2.169)

$$\beta_\omega' = iC \int_{-\infty}^0 ds' \left( \frac{\omega'}{\omega} \right)^{1/2} e^{i\omega's'} e^{-i\omega'v_0} \exp[i\omega 4MG \ln \left( -\frac{is'}{K} \right)].$$

(2.170)

Now the multiple-valued complex logarithm has to be dealt with. One gets a single-valued natural logarithm function by taking the cut in the complex plane along the negative real axis. So for $s' < 0$, as in the integrals above, the complex logarithm be written as

$$\ln(is'/K) = \ln(-i|s'|/K) = -i(\pi/2) + \ln(|s'|/K),$$

and

$$\ln(-is'/K) = \ln(i|s'|/K) = i(\pi/2) + \ln(|s'|/K).$$

This is because to get from the negative part of the imaginary axis to the positive part of the real axis, one has to perform a counterclockwise rotation over $\pi/2$, and to get from the positive part of the imaginary axis to the positive part of the real axis, a clockwise rotation over $\pi/2$ is
required.

So (2.169) and (2.170) become

\[ \alpha_{\omega\omega'} = -iCe^{i\omega'v_0e^{2\pi\omega MG}}\int_{-\infty}^{0} ds' \left( \frac{\omega'}{\omega} \right)^{1/2} e^{i\omega's'} \exp[i\omega'4MG \ln\left( \frac{|s'|}{K} \right) ] \]  

(2.171)

\[ \beta_{\omega\omega'} = iCe^{-i\omega'v_0e^{-2\pi\omega MG}}\int_{-\infty}^{0} ds' \left( \frac{\omega'}{\omega} \right)^{1/2} e^{i\omega's'} \exp[i\omega'4MG \ln\left( \frac{|s'|}{K} \right) ] . \]  

(2.172)

And this leads to the important result

\[ |\alpha_{\omega\omega'}|^2 = \exp(8\pi MG\omega)|\beta_{\omega\omega'}|^2 , \]

(2.173)

for the part of the wave packet that was propagated back in time through the collapsing body just before it formed a black hole.

For the components \( p_\omega \) of this part of the wave packet, one has the scalar product,

\[ (p_{\omega_1}, p_{\omega_2}) = \Gamma(\omega_1)\delta(\omega_1 - \omega_2) , \]  

(2.174)

where \( \Gamma(\omega_1) \) is the fraction of an outgoing packet of frequency \( \omega_1 \) at \( \mathcal{I}^+ \) that would propagate backward in time through the collapsing body to \( \mathcal{I}^- \). One can see this is the following way. Let \( p_\omega^{(2)} \) denote the components of this part of the wave packet, and let \( p_\omega^{(1)} \) denote the components of the part of the wave packet that if propagated backward in time would be scattered from the spacetime outside the collapsing body and would travel back in time, reaching \( \mathcal{I}^- \) with the same frequency \( \omega \) as when it had when it started from \( \mathcal{I}^+ \). This is because this latter part of the wave packet stays at all time in the outside region of the black hole (and later the mass that collapsed to form the black hole) so that its blueshift when approaching the mass and its redshift when going away from the mass cancel each other exactly.

Because \( p_\omega^{(1)} \) and \( p_\omega^{(2)} \) propagate to disjoint regions on \( \mathcal{I}^- \) (i.e. \( v > v_0 \) and \( v < v_0 \) respectively), they are orthogonal. With \( p_\omega = p_\omega^{(1)} + p_\omega^{(2)} \), it then follows that

\[ (p_{\omega_1}, p_{\omega_2}) = (p_{\omega_1}^{(1)}, p_{\omega_2}^{(1)}) + (p_{\omega_1}^{(2)}, p_{\omega_2}^{(2)}) . \]  

(2.175)

So from this and (2.150) one has

\[ (p_{\omega_1}^{(1)}, p_{\omega_2}^{(1)}) = \Gamma(\omega_1)\delta(\omega_1 - \omega_2) \]  

(2.176)

\[ (p_{\omega_1}^{(2)}, p_{\omega_2}^{(2)}) = (1 - \Gamma(\omega_1))\delta(\omega_1 - \omega_2) \]  

(2.177)

where \( \Gamma(\omega_1) \) is the fraction of the packet of frequency \( \omega_1 \) at \( \mathcal{I}^+ \) that would propagate back through the collapsing body to reach \( \mathcal{I}^- \)

It then follows from (2.160) and (2.175) that

\[ \Gamma(\omega_1)\delta(\omega_1 - \omega_2) = \int d\omega' (\alpha_{\omega_1\omega'}^*\alpha_{\omega_2\omega'} - \beta_{\omega_1\omega'}^*\beta_{\omega_2\omega'}) , \]  

(2.178)
where $\alpha_{\omega \omega'}$ and $\beta_{\omega \omega'}$ now refer to the coefficients in the expansion of $p^{(2)}_\omega$ in terms of the $f_{\omega'}$ and $f^*_{\omega'}$ as in (2.157).

The part of $b_\omega$ in (2.158) that is of interest is

$$b^{(2)}_\omega = (p^{(2)}_\omega, \psi).$$  \hspace{1cm} (2.179)

To simplify the notation, from now on $b_\omega$ will refer only to $b^{(2)}_\omega$.

The information about the particles that are created in the collapse of the body to form a black hole should be contained in $b_\omega$, but one encounters an infinity by straightforward evaluation of

$$\langle 0 | b^\dagger_\omega b_\omega | 0 \rangle = \int d\omega' |\beta_{\omega \omega'}|^2.$$  \hspace{1cm} (2.180)

This infinity is a consequence of the $\delta(\omega_1 - \omega_2)$ that appears in (2.178). Since $\langle 0 | b^\dagger_\omega b_\omega | 0 \rangle$ is the total number of created particles per unit frequency that reach $I^+$ at late times in the wave $p^{(2)}_\omega$, this total number is infinite (neglecting the change in the mass of the black hole of course) because there is a steady flux of particles reaching $I^+$ at late times.

One way to see this is to replace $\delta(\omega_1 - \omega_2)$ in (2.178) by

$$\delta(\omega_1 - \omega_2) = \lim_{T \to \infty} \frac{1}{2\pi} \int_{-T/2}^{T/2} dt e^{i(\omega_1 - \omega_2)t}.$$  \hspace{1cm} (2.181)

Then, for $\omega_1 = \omega_2 = \omega$ (2.178) can be written as

$$\lim_{T \to \infty} \Gamma(\omega)(T/2\pi) = \int d\omega' \left( |\alpha_{\omega \omega'}|^2 - |\beta_{\omega \omega'}|^2 \right) = [\exp(8\pi M G \omega) - 1] \int d\omega' |\beta_{\omega \omega'}|^2,$$  \hspace{1cm} (2.182)

where (2.173) was used. Hence,

$$\langle 0 | b^\dagger_\omega b_\omega | 0 \rangle = \lim_{T \to \infty} \frac{(T/2\pi)\Gamma(\omega)[\exp(8\pi M G \omega) - 1]^{-1}}{1}.$$  \hspace{1cm} (2.184)

The interpretation of this is that at late times, the number of created particles per unit angular frequency and per unit time that passes through a surface $r = R$, with $R$ much larger than the Schwarzschild radius, is

$$\frac{\Gamma(\omega)}{2\pi} \frac{1}{\exp(8\pi M G \omega) - 1}.$$  \hspace{1cm} (2.185)

(Note that the number per unit frequency per unit time has no factor $(2\pi)^{-1}$.)

Recall that the quantity $\Gamma(\omega)$ is the fraction of a purely outgoing wave packet that when propagated from $I^+$ backward in time would enter the collapsing body just before it had formed a black hole. At sufficiently late times this fraction is the same as the fraction of the wave packet that would enter the black hole past event horizon if the collapsing body were replaced in the spacetime by the analytic extension of the black hole spacetime. This means that $\Gamma_{in}(\omega)$ is also the probability that a purely incoming wave packet that starts from $I^-$ at late times will enter
the black hole event horizon, that is, will be absorbed by the black hole.

Therefore (2.185) implies that a Schwarzschild black hole emits and absorbs radiation exactly like a gray body of absorptivity $\Gamma(\omega)$ and temperature $T$ given by

$$kT = \frac{1}{8\pi MG}$$

$$\kappa = \frac{\kappa}{2\pi}$$

(2.186)

where $k$ is Boltzmann’s constant, and $\kappa = 1/4MG$ is the surface gravity of a Schwarzschild black hole as derived in section 1.6.

2.3.1.2 The Kerr black hole

Calculating the Hawking flux for a rotating Kerr black hole is essentially the same as in the non-rotating case, with two basic changes.

First, the radial geodesics in the Schwarzschild spacetime are replaced by the principal null congruence of geodesics in the Kerr spacetime, as derived in chapter 1. This means that as one traces back in time from $\mathcal{I}^+$ to $\mathcal{I}^-$ the part of an outgoing wave packet that passes through the collapsing body just before the event horizon has formed, the value of $u$ that the wave packet has on $\mathcal{I}^+$ is related to the value of $v$ it had on $\mathcal{I}^-$ by

$$u(v) \approx -\frac{1}{\kappa} \ln \left[ \frac{v - v_0}{K} \right] ,$$

(2.187)

as derived in section 1.9.2, where

$$\kappa = \kappa_+ = \frac{r_+ - r_-}{2(r_+^2 + a^2)}$$

(2.188)

is the surface gravity of the Kerr black hole as calculated in appendix B.

The second difference is that the event horizon of the Kerr black hole has angular velocity $d\phi/dt = \Omega_H$. As one approaches arbitrarily close to the null generators of the event horizon at $r_+$, both $\phi$ and $t$ diverge, but the angular coordinate $\tilde{\phi}_+ = \phi - \Omega_H t$ is well behaved in the vicinity of $r_+$. In tracing an outgoing wave packet with components $p^{(2)}_\omega$ back in time into the collapsing body just before it has fallen within the event horizon, the angular coordinate $\tilde{\phi}_+$ is appropriate as the wave packet passes into the collapsing body.

The result is that if $p^{(2)}_\omega$ has the form $\exp[-i\omega u + im\phi]$ at $\mathcal{I}^+$, then it has the form $\exp[-i(\omega - m\Omega_H)u(v) + im\phi']$ at $\mathcal{I}^-$, where $\phi'$ is the azimuthal angular coordinate in an inertial coordinate system far outside the collapsing body at early times. $m$ is the azimuthal quantum number, which may have either sign.

As a consequence of these two differences between the non-rotating and rotating cases, the quantity $\omega$ in the right-hand sides of (2.164) through (2.171) is replaced by the quantity $\omega - m\Omega_H$,
and \(u(v)\) is replaced by expression (2.187), so \(4MG\) in (2.167) - (2.171) is replaced by \(\kappa^{-1}\). Hence, for a rotating black hole one finds

\[
|\alpha_{\omega\nu}|^2 = \exp[2\pi\kappa^{-1}(\omega - m\Omega_H)]|\beta_{\omega\nu}|^2
\]

(2.189)

instead of (2.173).

It then follows, as in the previous section, that the average number of particles created in a wave packet that reaches \(I^+\) with energy \(\omega\) and angular momentum quantum numbers \(l, m\) is

\[
\langle N_{\omega lm} \rangle = \Gamma_{lm}(\omega)\{\exp[2\pi\kappa^{-1}(\omega - m\Omega_H)] - 1\}^{-1},
\]

(2.190)

where the surface gravity \(\kappa\) is given by (2.188), and \(\Gamma_{lm}(\omega)\) is the same as the fraction of a similar wave packet incident on a Kerr black hole that would be absorbed by the black hole. Thus, the Kerr black hole acts like a gray body at temperature

\[
kT = \frac{\kappa}{2\pi}
\]

(2.191)

where \(k\) is again Boltzman’s constant. This is the same equation as for a Schwarzschild black hole, but only the expression for the surface gravity is different.

If (2.190) is to make sense, i.e. \(\langle N_{\omega lm} \rangle > 0\), \(\Gamma_{lm}(\omega)\) has to be negative when \(\omega < m\Omega_H\). This means that when an incoming wave packet with \(\omega < m\Omega_H\) is sent towards a Kerr black hole, the backscattered part of the wave packet returns with a larger amplitude than the original incoming packet. This is the superradiant scattering phenomenon that was discussed in section 1.10.3. Superradiance can be thought of as stimulated pair production caused by the incoming boson.

For fermions one finds the expression

\[
\langle N_{\omega lm} \rangle_{\text{ferm}} = \Gamma_{lm}(\omega)\{\exp[2\pi\kappa^{-1}(\omega - m\Omega_H)] + 1\}^{-1}.
\]

(2.192)

The +-sign now implies that \(\Gamma_{lm}(\omega)\) remains positive at all frequencies. So there is no radiant scattering for fermions because of the Pauli exclusion principle.

For a charged rotating black hole, it can also be shown [58] that the average number of particles of charge \(e\) emitted in mode \(\omega, l, m\) has the same form of (2.190), but with \(\omega - m\Omega_H - e\Phi\) appearing in the exponential, where \(\Phi\) is the electrostatic potential of the black hole, and with the expressions for the surface gravity \(\kappa\) and gray body factors \(\Gamma\) appropriate for a rotating charged black hole. The temperature of such a black hole satisfies again equation (2.191) with the appropriate expression for the surface gravity.

### 2.3.1.3 Final remarks

The above derivation of the Hawking radiation can easily be generalized to the case of non-spherical symmetric gravitational collapse. The late time emission depends only on the final...
state of the black hole. The detailed nature of the collapse and the manner in which the black hole 'settles down' to its final state are not relevant. So one can conclude from the uniqueness theorems of section 1.8.1 that we have actually treated the most general case of Hawking radiation, at least, with respect to the spacetime in which the emission takes place.

The generalization to physical relevant interacting fields is not so evident. To address this issue, we mention that the existence of the Hawking flux has also been derived in the algebraic framework of quantum field theory in curved spacetime as described in section 2.1.3.3 [59]. There, the derivation of the thermal behavior of the quantum field at asymptotically late times is shown to arise from the singularity structure of the two-point function at arbitrary short distances. However, even ignoring possible new effects arising from the quantum nature of gravity itself at distance scales smaller than the Planck length, it is unreasonable just to assume that the simple linear field model considered in the derivation above will provide an accurate model to a realistic field theory at ultra-short distance scales. Thus, one might question whether the particle creation effect will occur for nonlinear fields even if these fields can be treated as non-interacting on large distance scales or equivalently, at low energies. In response to this issue, it should be noted that the Unruh effect, which has the same physical and mathematical origin as the Hawking effect, is proven to continue to hold for nonlinear fields in Minkowski spacetime by a theorem of Bisognano and Wichmann [60]. Furthermore, there is strong evidence based upon the analytic continuation of propagators to a Euclidean curved spacetime that the Unruh effect even continues to hold for nonlinear fields in static curved spacetimes [2]. So although there is no conclusive proof that the Hawking effect continues to hold for nonlinear fields, all the evidence currently available points to the fact that it does. Together with it’s role in completing black hole thermodynamics (see section 2.5 below) this makes that there is very little doubt about the validity of the Hawking effect for interacting fields.

Although the emission of Hawking radiation has a very low intensity, especially for large black holes, after a sufficient amount of time backreaction effects on the metric will become relevant. By conservation energy it is clear what will happen: the mass of the black hole, and thereby its Schwarzschild radius, will decrease because of the energy that is being emitted under the form of Hawking particles. As the black hole gets smaller, it gets hotter and so starts to radiate faster. As the temperature rises, it exceeds the rest mass of subsequently more and more massive particles. So at first, only photons and neutrino’s will be emitted, then the temperature increases and particles such as electrons and muons would will begin to constitute the Hawking flux until eventually all types of particles will take place in the radiation process. At the time the black hole temperature reaches the strong interaction energy scale, a large amount of energy will be emitted at time scales of $10^{-23}$s. So whatever theory dictates the laws of physics at the Planck scale, it is very likely that the evaporation process will end with an explosion, completely erasing the black hole.

After the evaporation process, the energy that was originally in the black hole will be uniformly spread throughout space. Because of the low emission rate of the Hawking radiation the energy density will be negligible and the final state of the evaporation process will be flat spacetime. So after an appropriate 'gluing job' (see section 1.7.2) between the Penrose diagram of a gravitational collapse spacetime and that of Minkowski spacetime one gets the Penrose diagram of a spacetime for gravitational collapse of matter to a black hole and the subsequent
evaporation process leading to flat spacetime. This diagram is given on figure 2.3, where \( B \) represents the boundary of the collapsing body.

![Diagram](image)

**Figure 2.3:** The Penrose diagram of a spacetime for gravitational collapse and black hole evaporation.

### 2.3.2 Alternative views on the Hawking radiation

Now the original derivation of the Hawking effect is presented, its relation to other physical mechanisms is given. The aim is to create a context for black hole radiation and to show how it perfectly connects with other ideas presented in this thesis.

#### 2.3.2.1 Static observers and the Unruh effect

Return to the Schwarzschild solution

\[
d s^2 = \left(1 - \frac{2GM}{r}\right) dt^2 - \left(1 - \frac{2GM}{r}\right)^{-1} dr^2 - r^2 d\Omega^2, \tag{2.193}
\]

and let

\[
 r - 2GM = \frac{\rho^2}{8GM}. \tag{2.194}
\]

Then

\[
1 - \frac{2GM}{r} = \frac{(\kappa \rho)^2}{1 + (\kappa \rho)^2}, \tag{2.195}
\]

where \( \kappa = 1/4GM \) for the Schwarzschild black hole was used. In the region near the horizon, i.e. \( \rho \ll 1 \), one finds

\[
1 - \frac{2GM}{r} \approx (\kappa \rho)^2. \tag{2.196}
\]
From (2.194) one also has

\[ dr = \frac{\rho}{4GM}d\rho. \]

And therefore

\[ dr^2 = (\kappa \rho)^2 d\rho^2. \]  

(2.198)

So in a small region outside the horizon, (2.193) can be written as

\[ ds^2 \approx (\kappa \rho)^2 dt^2 - d\rho^2 + \frac{1}{4\kappa^2}d\Omega^2, \]

(2.199)

where the last term represents a 2-sphere of radius \(1/2\kappa\). The first two terms can be rewritten as

\[ ds'^2 = \rho^2 d(\kappa t)^2 - d\rho^2, \]

(2.200)

which after a comparison with (2.88) appears to be nothing but two-dimensional Rindler space. More specifically, region I of the maximally extended Schwarzschild spacetime (see section 1.5 of chapter 1) can be identified with the right Rindler wedge. So in this near-horizon Rindler description, the black hole horizon is an acceleration horizon. From the discussion of section 2.2 about the Unruh effect, one could therefore suspect that an observer on an orbit of \(\partial/\partial(\kappa t)\), i.e. a static observer just outside the horizon, would detect a thermal bath of particles. So the Unruh effect and the Hawking effect are perfectly consistent with each other. Of course, one should not take this analogy too literally since the Unruh effect takes place in flat Minkowski spacetime and the Hawking effect in a curved black hole spacetime. Nevertheless, the same physical principle seems to be at work in both cases.

Although the proper acceleration of an \(\rho = \text{constant}\) worldline diverges as \(\rho \to 0\), its acceleration as measured by another \(\rho = \text{constant}\) observer will remain finite. Since

\[ d\tau^2 = \rho^2 d(\kappa t)^2, \]

(2.201)

with \(\rho = a^{-1}\) constant, the acceleration as measured by an observer whose proper time is \(t\) is

\[ \left( \frac{d\tau}{dt} \right) \frac{1}{\rho} = (\kappa \rho) \frac{1}{\rho} = \kappa. \]

(2.202)

But in Schwarzschild spacetime, an observer with proper time \(t\) is one at spatial infinity. This points out the equivalency between the Unruh temperature and the Hawking temperature and confirms the physical interpretation of the surface gravity given in section 1.6.

\subsection*{2.3.2.2 Heuristic arguments}

To end the discussion on the origin of the Hawking radiation, two heuristic arguments are given which make the effect blend in with other physical phenomena.

First, recall that in section 1.4.1 it was mentioned that there exists an analytically solved model describing gravitational collapse of a spherically symmetric uniform dust cloud to a black hole. The solution existed of a matching of the Friedman-Lemaître solution on the interior of the cloud to the Schwarzschild solution on the outside. In the derivation of the Hawking radiation
above it also became clear that the structure of spacetime just prior to the horizon formation is of crucial importance for the existence of the Hawking particles. And finally, in section 2.1.2 it was shown that there is particle creation in an expanding or contracting spacetime. It is now clear that these three ideas, presented in different contexts throughout this thesis, are perfectly consistent with the idea of Hawking radiation. So they present another viewpoint on the creation of Hawking particles.

Another viewpoint that was already presented in Hawking’s original paper is that of negative energy flux across the horizon. One might picture this as follows. Just outside the event horizon there will be virtual pairs of particles, one with negative energy and one with positive energy. The negative particle is in a region which is classically forbidden but it can tunnel through the event horizon to the interior region. As seen chapter 1, the Killing vector field \( k \) representing time translations at infinity is space-like in this region. So the particle can exist as a real particle with a timelike momentum vector even though its energy relative to infinity as measured by the Killing vector field \( k \) is negative. The other particle of the pair, having a positive energy, can escape to infinity where it constitutes a part of the Hawking radiation. The probability of the negative energy particle tunnelling through the horizon is governed by the surface gravity since this quantity measures the gradient of the magnitude of the Killing vector. Or in other words, how fast the Killing vector is becoming spacelike. Instead of thinking of negative energy particles tunnelling through the horizon in the positive sense of time, one could regard them as positive energy particles crossing the horizon on past-directed world-lines and then being scattered onto future-directed world-lines by the gravitational field. However, it should be emphasized that this interpretation should not be taken to literally, certainly when recalling the problems of the particle interpretation of quantum field theory in curved spacetime as explained in section 2.1.3.

A final viewpoint is that a black hole is an excited state of the gravitational field which decays quantum mechanically and energy should be able to tunnel out of its potential well because of quantum fluctuations of the metric.

### 2.3.3 Trans-Planckian physics in Hawking radiation

After Hawking published his paper deriving the thermal spectrum of the radiation created by a black hole [57, 58], questions were raised about the use of paths from \( I^{-} \) to \( I^{+} \). The frequencies of massless particles receive arbitrarily large redshifts along such paths as they pass through the collapsing dust cloud just prior to formation of the event horizon. The range of frequencies that can be seen by distant observers at late times would have had to originate at \( I^{-} \) with ultrahigh frequencies, including frequencies above the Planck scale. Local Lorentz invariance would be violated if such frequencies would be arbitrarily cut off. So the question was if the Hawking thermal spectrum would nevertheless survive the breaking of local Lorentz invariance. There is no conclusive answer to this question, but in the remaining of this section some models are presented that strongly hint that the physics at the Planck scale does not influence the Hawking spectrum.

In the context of black holes, Unruh [61] considered a definite model of sound waves propagating in a moving fluid that simulates the behavior of the event horizon of a black hole (see the
water analogy in the introduction of chapter 1). By numerical methods he found that despite
the breaking of Lorentz invariance in his fluid model, the sonic black hole nevertheless produced
a spectrum of sound waves that was very close to a thermal spectrum. He demonstrated that
the ultrahigh frequencies are not responsible for the thermal spectrum produced by a sonic black
hole. This supports the viewpoint that the ultra high frequencies that appear in the derivation
of the Hawking thermal spectrum in black hole evaporation are not necessarily essential for
obtaining the thermal spectrum. In this context, related models with dispersion relations that
break Lorentz invariance have been considered by, for example, Jacobson [62].

In [63] the Hadamard form of the two-point correlation function of the field at very short
distances characterized by an invariant Planck length was altered. The invariance of the Planck
length appearing in the two-point function is enforced by means of a non-linear physical real-
ization of the Lorentz group. It was shown that this alteration of the Hadamard form at the
invariant Planck scale has negligible effect on the thermal spectrum of Hawking radiation. This
conclusion extends to spectral frequencies much higher than the energy scale set by the Hawking
temperature of the black hole. Thus, the thermal spectrum of an evaporating black hole of radius
above the Planck scale appears to be insensitive to such changes in physics near the Planck scale.

In Deser and Levin [64, 65], the spacetime of a four-dimensional black hole is embedded in a
six-dimensional Minkowski spacetime in a global way, in the sense that the embedding in the six-
dimensional flat spacetime covers the usual Kruskal maximal extension of Schwarzschild space-
time (with a white hole in it) without encountering a coordinate singularity at the Schwarzschild
radius of the black hole. In this embedding, a detector held at rest at constant Schwarzschild
radial distance $r$ is mapped to a detector moving at constant acceleration in the six-dimensional
Minkowski spacetime. It is shown that the temperature $a/2\pi$ of the thermal spectrum measured
by this uniformly accelerated detector is the temperature that the detector would detect as a
result of the Hawking radiation. This correspondence makes no use of trans-Planckian frequen-
cies and thus supports the view that they are not essential to the thermal spectrum of Hawking
radiation.

The string theory derivation of the Hawking radiation for a nearly extremal supersymmetric
black hole [66] makes use of the Minkowski spacetime limit of the black hole in terms of $D$-branes
and oppositely moving string excitations that interact and produce the Hawking thermal spec-
trum of radiation, including the gray-body factor, without appealing to large red- or blueshifts.
This again suggests that the thermal spectrum is not dependent on very high frequency modes
of the radiation field.

2.4 Angular momentum and gray body factors

In this section the role of the gray body factors of the black hole spectrum $\Gamma_{lm}(\omega)$, that were
encountered in the derivation of the Hawking radiation, will be discussed. In particular, the
focus will be on their relation with the angular momentum of the particles. The derivation is
based upon [9].
Again, a massless scalar field $\psi$ is considered in a Schwarzschild background. It is of advantage in this section to use the tortoise coordinates as introduced in chapter 1. With the metric in tortoise coordinates (1.52), the action for $\psi$ can be written as

$$S = \frac{1}{2} \int d^4x \left( -g^{\mu\nu} \partial_\mu \psi \partial_\nu \psi \right)^{1/2} \right) - \frac{1}{r^2} \left( \frac{\partial \psi}{\partial \theta} \right)^2 - \frac{1}{r^2 \sin^2 \theta} \left( \frac{\partial \psi}{\partial \phi} \right)^2 \right) F r^2 \sin \theta \right)$$

(2.203)

With $F = (1 - \frac{2MG}{r})$. By defining

$$\chi = r\psi$$

the action takes the form

$$S = \frac{1}{2} \int dt dr^* d\theta d\phi \left[ (\partial_t \chi)^2 - \left( \frac{\partial \chi}{\partial r^*} - \frac{\partial \ln r}{\partial r^*} \chi \right)^2 - F \left( \sin \theta \left( \frac{\partial \chi}{\partial \theta} \right)^2 + \frac{1}{\sin \theta} \left( \frac{\partial \chi}{\partial \phi} \right)^2 \right) \right]$$

(2.204)

which, after an integration by parts and the introduction of the spherical harmonic decomposition becomes

$$S = \sum_{lm} \frac{1}{2} \int dt dr^* \left[ (\dot{\chi}_{lm})^2 - \left( \frac{\partial \chi_{lm}}{\partial t} \right)^2 - \left( \frac{\partial \chi_{lm}}{\partial r^*} \right)^2 - V_l(r^*) \chi_{lm}^2 \right]$$

(2.205)

Using the relation between $r$ and $r^*$ (1.51), one gets for each $l, m$ an action

$$S_{lm} = \frac{1}{2} \int dt dr^* \left[ \left( \frac{\partial \chi_{lm}}{\partial t} \right)^2 - \left( \frac{\partial \chi_{lm}}{\partial r^*} \right)^2 - V_l(r^*) \chi_{lm}^2 \right]$$

(2.206)

where the potential $V_l(r^*)$ is given by

$$V_l(r^*) = \frac{r - 2MG}{r} \left( \frac{l(l+1)}{r^2} + \frac{2MG}{r^3} \right)$$

(2.207)

The equation of motion is

$$\frac{\partial^2 \chi_{lm}}{\partial t^2} = \frac{\partial^2 \chi_{lm}}{\partial r^{*2}} - V_l(r^*) \chi_{lm}$$

(2.208)

For a mode of frequency $\nu$ this becomes

$$- \frac{\partial^2 \chi_{lm}}{\partial r^{*2}} + V_l(r^*) \chi_{lm} = \nu^2 \chi_{lm}$$

(2.209)

The potential $V_l(r^*)$ (2.207) is shown in figure 2.4 as a function of the Schwarzschild coordinate $r$.

For $r \gg 3MG$ the potential is repulsive. The potential can be seen as the relativistic generalization of the repulsive centrifugal barrier. However, closer to the horizon, the gravitational attraction takes over and the potential becomes attractive. So a wave packet gets pulled towards the horizon there. The maximum of the potential, which separates the two regions of repulsion
Figure 2.4: The effective potential for a massless scalar field in a Schwarzschild background and attraction, depends only weakly on the angular momentum $l$. It is given by

$$r_{\text{max}} = 3MG \left( \frac{1}{2} \left( 1 + \sqrt{1 + \frac{14l^2 + 14l + 9}{9l^2(l + 1)^2}} \right) - \frac{1}{2l(l + 1)} \right). \quad (2.210)$$

For $l \to +\infty$ the maximum occurs at $r_{\text{max}}(\infty) = 3MG$.

The same potential governs the motion of massless classical particles. The points $r_{\text{max}}(l)$ represent unstable circular orbits, and the innermost such orbit is at $r = 3MG$. Any particle that starts with vanishing radial velocity in the region $r < 3MG$ will spiral into the horizon. In the region of large negative $r^*$ where the horizon is approached, the potential is unimportant and the field behaves like a free field. The eigenmodes in this region have the form of plane waves which propagate with unit velocity ($c = 1$)

$$\frac{dr^*}{dt} = \mp 1 \quad \chi \to e^{ik(r^* \pm t)} \quad (2.211)$$

Now the link with the derivation of the Hawking radiation can be made. There, geodesic paths from $I^+$ to $I^-$ were used along which a wave packet was propagated back in time. Then, it was said that a fraction of this wave packet would be scattered back to $I^-$ and a fraction would travel through the collapsing body to $I^-$. The fraction of the total wave packet that would travel through the collapsing body was denoted by $\Gamma_{lm}(\omega)$ and played the role of the gray body factor in the thermal spectrum of the Hawking radiation. To find the link between this gray body factor and the angular momentum, the discussion of the effective potential can be used.

Consider a field quantum of frequency $\nu$ and angular momentum $l$ propagating from $I^+$ towards the potential barrier at $r \approx 3MG$. Using the fact that equation (2.209) has the form of a Schrödinger equation for a particle of energy $\nu^2$ in a potential $V_l(r^*)$, and the time-reversal symmetry of the Schrödinger equation, we can derive an estimate for the effect of the gray body factors. The field quantum has enough energy to overcome the barrier without tunneling if $\nu^2$
is larger than the maximum height of the barrier, which can be approximated by

\[ V_{\text{max}} \approx \frac{1}{27} \frac{l^2}{M^2 G^2}. \]  

(2.212)

So the threshold energy for passing over the barrier is

\[ \nu \sim \frac{1}{\sqrt{27} MG}. \]  

(2.213)

Less energetic particles must tunnel through the barrier. Thus, the effect of the gray body factors is that particles of low angular momentum are more easily emitted by the black hole. The black hole radiation will therefore have a dominant contribution of low angular momentum quanta.

We can make this statement a little bit more concrete. To do so, first, conventional units are restored. A black body spectrum is peaked at \( \hbar \omega \approx 3kT \), and filling in this peak frequency in (2.213) together with the expression for the temperature of a black hole (2.247) gives

\[ \frac{3\kappa}{(2\pi)^2 c} \sim \frac{1}{\sqrt{27} MG}. \]  

(2.214)

So we can say that the black hole radiation will have a negligible contribution of quanta with angular momentum

\[ l > \frac{3\sqrt{27}}{16\pi^2} \approx 0.1, \]  

(2.215)

where \( \kappa = c^4/4MG \) for a Schwarzschild black hole was used. From this we can conclude that the Hawking radiation will be heavily dominated by s-wave quanta.

### 2.5 The generalized second law

In chapter 1, it was showed that there is a striking mathematical analogy between certain laws applying to black hole mechanics and the laws of thermodynamics. In this correspondence of laws, the mass of the black hole plays the same mathematical role as the total energy of a thermodynamic system. Since mass and energy represent the same physical quantity, this suggests that the analogy of laws might have some physical content.

However, classically this physical analogy breaks down: the quantity in black hole physics which plays the role mathematically analogous to the temperature in thermodynamics is the surface gravity \( \kappa \), but the physical temperature of a classical black hole is absolute zero. However, as shown in the previous section, the treatment of a quantum field in the black hole spacetime implies that \( \kappa/2\pi k \) truly is the physical temperature of a black hole. Hence, this suggests the possibility that the laws of black hole mechanics truly are the ordinary laws of thermodynamics applied to a system containing a black hole. In this section the generalized second law will be described, which strongly suggests that \( A/4G \) should be regarded as the physical entropy of a black hole. This neatly falls into place with the quantum mechanical derivation of \( \kappa/2\pi \) as the physical temperature of a black hole and the first law of black hole mechanics (1.246).
First, it should be noted that there are some difficulties with the ordinary second law of thermodynamics and with the area theorem. A difficulty with the ordinary second law arises when a black hole is present. One can take some matter and dump it into a black hole in which case, at least according to classical general relativity, it will disappear into the singularity within the black hole. In this manner, the total entropy of matter in the universe can be decreased. On the other hand, the area theorem clearly must be violated in the quantum particle creation process since the mass \( M \) of the black hole and hence its area must decrease in the process if energy is to be conserved. This violation of the area theorem can occur because the expectation value of the energy-momentum tensor of the quantum field violates the null energy condition at the horizon of the black hole. This violation is caused by the indeterminacy of particle number and energy of a quantum field in a curved spacetime. However, when the total entropy \( S_m \) of matter outside of black holes is decreased by dumping matter into a black hole, \( A \) will tend to increase. Similarly, when \( A \) is decreased during the particle creation process, thermal matter is created outside the black hole, so \( S_m \) increases. Thus, although \( S_m \) and \( A \) each can decrease individually, it is possible that the generalized entropy \( S' \) defined by

\[
S' = S_m + \frac{1}{4G} A
\]

never decreases. The conjecture that \( \Delta S' > 0 \) was first put forth by Bekenstein [67] and is known as the generalized second law (historically, this was done prior to the discovery of particle creation by black holes).

If valid, the generalized second law would have a very natural interpretation. Presumably, it simply would be the ordinary second law of thermodynamics applied to a system containing a black hole. If so, then there would be no question that \( A/4G \) truly represents the physical entropy of a black hole. Thus, a key issue in the subject of black hole thermodynamics is whether the generalized second law holds.

### 2.5.1 The lowering of matter in a static black hole

For simplicity, consider a static black hole. In that case, the Killing vector field which is timelike at infinity \( k \) coincides with the Killing vector field \( \xi \) which is normal to the horizon of the black hole. Far from the black hole, put matter of energy \( E \) and entropy \( S \) into a box and then lower the box quasistatically on a rope towards the black hole. When the horizon is reached, open the box and allow the matter to fall into the black hole. Since no entropy need be generated in the lowering process, the entropy of matter outside the black hole will be decreased by \( S \) in this process, i.e. \( \Delta S_m = -S \). We consider \( E \) to be much smaller than the black hole mass so that the dumping of the matter can be treated as a perturbation.

On the other hand, the area change of the black hole can be calculated as follows. The force exerted by the distant observer who holds the rope is given by

\[
F_\infty = E \frac{d[\xi]}{dy},
\]
where $|\xi| = \sqrt{\xi^2} = \sqrt{\xi_\mu\xi^\mu}$ is the redshift factor, which for the Schwarzschild black hole reduces to $(1 - 2GM/r)^{1/2}$ and corresponds to (1.47) as previously derived in section 1.4.2 with $r_1 \to \infty$. $y$ denotes the proper distance along the path followed by the box in the (quasi-)static hypersurface. It is assumed that the dimension of the box in the $y$-direction is negligible. The expression for the force readily follows from its definition as the gradient of the potential energy.

So it follows that the work done by the observer at infinity during the lowering of the box is given by

$$W_\infty = -\int_0^y dy F_\infty = (1 - |\xi|)E,$$

where the integral is taken from infinity to the point where the matter is released out of the box and it is used that the redshift factor at infinity is 1. Thus, by conservation of energy, the energy delivered to the black hole is

$$\Delta M = E - W_\infty = |\xi|E.$$  

(2.219)

By the first law of black hole mechanics (1.246), the area increase of the black hole in this process is given by

$$\Delta A = \frac{8\pi G}{\kappa} \Delta M = \frac{8\pi G}{\kappa}|\xi|E.$$  

(2.220)

However, at the horizon $|\xi| = 0$, so by lowering the box sufficiently close to the horizon, one can make $\Delta A$ arbitrary small. Thus, it would appear that one can make $\Delta S' = -S + \Delta A/4G$ negative, in violation of the generalized second law.

The problem with the derivation above is that it does not take into account quantum effects and the corresponding Hawking radiation. This might be surprising because the set-up of the problem is truly macroscopic and the black hole mass can be chosen so large that the Hawking radiation seen at infinity is negligible and there are no important nonclassical effects on freely falling bodies. Nevertheless, it will appear that the Unruh effect makes a large quantum correction to the behavior of a body which is quasi-statically lowered towards the horizon of the black hole.

As mentioned in section 2.2, when a quantum field is in the natural vacuum state associated with observers on orbits of $\xi$, a static observer will see himself immersed in a thermal bath at the locally measured temperature

$$T = \frac{\kappa}{2\pi|\xi|}.$$  

(2.221)

Since the redshift factor is not constant, there will be a nonzero gradient of the locally measured temperature as seen by static observers. By the Gibbs-Duhem relation of thermodynamics in the case of vanishing chemical potential, there will be a pressure gradient associated with the
thermal bath given by
\[ \nabla_\mu P = s \nabla_\mu T, \tag{2.222} \]
where \( s \) is the entropy density of the thermal bath. Consequently, there will be a force exerted on the box lowered quasi-statically towards the horizon of the black hole, much as though the box were being lowered into an ordinary fluid body. Taking into account this force, the total force (2.217) is modified to become
\[ F_\infty = E \frac{d|\xi|}{dy} + V \frac{d(|\xi|P)}{dy}, \tag{2.223} \]
where \( V \) denotes the volume of the box. Integrating this equation, one finds for the work done during the lowering process
\[ W_\infty = (1 - |\xi|)E - |\xi|PV, \tag{2.224} \]
so that the energy delivered to the black hole is now given by
\[ \Delta M = |\xi|(E + PV). \tag{2.225} \]
Thus, more energy is delivered to the black hole than was found in the above classical calculation. Indeed, since \(|\xi|P\) becomes large near the horizon, the optimal place to release the matter into the black hole is no longer at the black hole horizon. Rather, the optimal place now occurs at the value of \( y \) at which the increase in mass becomes minimal
\[ 0 = \frac{d(\Delta M)}{dy} = -\frac{dW_\infty}{dy} = -F_\infty, \tag{2.226} \]
i.e. at the 'floating point' of the box. By means of (2.223), (2.222) and (2.221), the 'floating point' condition is
\[
0 = E \frac{d|\xi|}{dy} + PV \frac{d|\xi|}{dy} + V|\xi| \frac{dP}{dy} \\
= (E + PV) \frac{d|\xi|}{dy} + V|\xi|s \frac{dT}{dy} \\
= (E + PV - VsT) \frac{d|\xi|}{dy}.
\]
Since \( d|\xi|/dy \neq 0 \), the floating point condition becomes
\[ E + PV - VsT = 0. \tag{2.228} \]
Now one can use the integrated form of the Gibbs-Duhem relation for the thermal bath
\[ eV + PV - sTV = 0, \tag{2.229} \]
where \( e \) denotes the energy density of the thermal bath, to write the condition for the box to float as
\[ E = eV, \tag{2.230} \]
which agrees with a result previously found by Archimedes.
With this result one obtains the minimum energy that can be delivered to the black hole in this process. Use (2.230) to rewrite (2.225) as

$$\Delta M_{\text{min}} = |\xi| sTV,$$

(2.231)

and use (2.221) to obtain

$$\Delta M_{\text{min}} = \frac{\kappa}{2\pi} V s.$$  

(2.232)

So by (2.220) one gets for the minimum increase in the area

$$\Delta A_{\text{min}} = \frac{8\pi G}{\kappa} \Delta M_{\text{min}} = \frac{4GVs}{\kappa}.$$  

(2.233)

Thus, the net change in the generalized entropy in the process is given by

$$\Delta S' = \Delta S_m + \frac{1}{4G} \Delta A \geq \Delta S_m + \frac{1}{4G} \Delta A_{\text{min}} = -S + sV,$$

(2.234)

where $s$ is the entropy density of the thermal bath at the floating point. But, by definition, at a given energy and volume, the entropy is maximum in a thermal state. Therefore, if follows from (2.230) that

$$sV \geq S,$$  

(2.235)

and thus

$$\Delta S' \geq 0.$$  

(2.236)

So the generalized second law cannot be violated by this process.

Note that in the above calculation of the extra force on the box, an energy density $e$ and pressure $P$ were attributed to the thermal bath of the radiation. In fact, this is not correct. For a macroscopic black hole, the true expectation value of the energy momentum tensor $\langle T_{\mu\nu} \rangle$ of the quantum field is negligibly small near the horizon as expected on physical grounds. The thermal bath values $e$ and $P$ used in the above calculation actually measure the expected energy and pressure relative to the natural vacuum state defined by observers on the static isometries. These static isometries are the orbits of $\partial/\partial t$ and because the box is lowered quasistatically, it follows to good approximation such an orbit. Therefore, starting at infinity, the natural zero-energy reference point is taken to be the vacuum as seen by observers on orbits of $\partial/\partial t$. Thus, it follows that for a macroscopic black hole, the expected energy density and pressure and the natural vacuum state are nearly $-e$ and $-P$ respectively. Since only the energy-momentum tensor differences between the outside and the inside of the box are relevant to the calculation of the forces on the box, this shift in the zero-point of $\langle T_{\mu\nu} \rangle$ has no effect on the above results.

This reasoning suggests that the process is more accurately described by saying that, rather than feeling an externally applied force, the box fills up with negative energy and pressure according to the Moore or 'moving mirrors' effect where particles are created in the box by moving
perfectly reflecting boundaries [68] as it is slowly lowered. In this description, the floating point occurs when a sufficient amount of negative energy has flowed into the box so that the total energy in the box is zero. The difference between the behavior of a slowly lowered box, which feels a large force of quantum origin, and a freely falling box also is readily explained in this viewpoint since the freely falling box does not fill up with negative energy.

2.5.2 A more general argument

In this section a more general argument is given for the validity of the generalized second law in the case of processes which can be treated as small perturbations of a stationary black hole [2, 10].

Consider a process where one starts with a stationary black hole and perturb it infinitesimally by some process, e.g. by dropping matter into it. The aim is to calculate the net change in the generalized entropy resulting from this process. In comparing the perturbed spacetime with the unperturbed black hole, it is convenient in such a way that the black hole horizons coincide and have the same null generators. In addition, one identifies the spacetimes so that in a neighborhood of the horizon of the perturbed spacetime, the image under this identification of the Killing vector field $\xi$ normal to the horizon has the same norm as it has in the unperturbed spacetime. This can be achieved by composition of any horizon preserving identification with an additional diffeomorphism which moves points along the orbits of $\xi$, thereby compressing or stretching $\xi$ as needed. One then defines $\xi$ on the perturbed spacetime to be the image of $\xi$ under this identification of the Killing vector field $\xi$. Thus, in this choice of 'gauge', one automatically has $\delta\xi^\mu = 0$ on the perturbed spacetime as well as $\delta|\xi| = 0$ in a neighborhood of the horizon.

Consider the family of observers outside the black hole that follow orbits of $\xi$. In the unperturbed spacetime such observers see a thermal bath of particles, and relative to the stationary vacuum state $|0\rangle_s$ associated with $\xi$, they would assign a thermal bath energy density $e$ to the quantum field given by

$$e = T_{\mu\nu} \frac{\xi^\mu \xi^\nu}{|\xi|^2},$$

(2.237)

where $T_{\mu\nu}$ denotes the difference between the actual expectation value of the energy-momentum tensor and the expectation value of the energy-momentum tensor in the state $|0\rangle_s$. Such observers would naturally assign to the quantum field a thermal bath entropy current of the form

$$S^\mu = \frac{\xi^\mu}{|\xi|}.$$  

(2.238)

Then the local entropy density $s$ is given in terms of $S^\mu$ by

$$s = -\frac{S^\mu \xi^\mu}{|\xi|},$$

(2.239)
Now consider the perturbed spacetime and the observers following orbits of $\xi$. The perturbation in the energy and entropy densities they would assign to the quantum field are given by

$$
\delta e = \delta \left[ T_{\mu\nu} \frac{\xi^\mu \xi^\nu}{\xi^2} \right] = (\delta T_{\mu\nu}) \frac{\xi^\mu \xi^\nu}{\xi^2},
$$

(2.240)

$$
\delta s = -\delta \left[ S_\mu \frac{\xi^\mu}{|\xi|} \right] = - (\delta S_\mu) \frac{\xi^\mu}{|\xi|}.
$$

(2.241)

However, $\delta s$ would be maximized for a given $\delta e$ if the perturbed field remained locally in a thermal state. Hence, one must have

$$
\delta s \leq (\delta s)_{\text{th}} = \frac{\delta e}{T} = \frac{2\pi |\xi|}{\kappa} \delta e,
$$

(2.242)

where the ordinary first law of thermodynamics for the thermal bath was used as well as (2.221) for the locally measured temperature. Multiplying this equation by $|\xi|$ and taking the limit as one approaches the horizon, one gets using (2.240) and (2.241)

$$
- (\delta S_\mu) \xi^\mu |_{\text{horizon}} \leq \frac{2\pi}{\kappa} (\delta T_{\mu\nu}) \xi^\mu \xi^\nu |_{\text{horizon}}.
$$

(2.243)

Integrating this relation over the horizon with respect to the Killing parameter $v$, the left side can be interpreted as the total flux of matter entropy into the black hole, whereas the right side is proportional to the same combination of energy and angular momentum fluxes as appeared in the derivation of the first law in chapter 1. So using (1.244) one can write

$$
- \Delta S_m \leq \frac{2\pi}{\kappa} (\Delta M - \Omega H \Delta J)
$$

(2.244)

Therefore, it follows from the first law of black hole mechanics (1.246) that

$$
- \Delta S_m \leq \frac{1}{4G} \Delta A
$$

(2.245)

and thus

$$
\Delta S' = \Delta S_m + \frac{1}{4G} \Delta A \geq 0,
$$

(2.246)

which again confirms the generalized second law.

In chapter 4 an even more general argument for the validity of the second law will be given. If the generalized law is accepted to be true, then by far the most natural interpretation of the laws of black hole thermodynamics is that they simply are the ordinary laws of thermodynamics applied to a black hole. In that case $A/4G$ truly would represent the physical entropy of a black hole, and $S'$ simply would be the total entropy of the universe, including contributions from both ordinary matter and from black holes. In the absence of a complete quantum theory of gravity, it is hard to imagine how a more convincing case could be made for this conclusion.
There are some major puzzles involving black hole entropy. First of all, the main idea underlying ordinary thermodynamics and the usual interpretation of entropy is the 'ergodic principle' which states the equivalence between time averages and phase space averages. In view of the nature of 'time' in general relativity, it is hard to see how this notion would be applicable to a system containing a black hole, and if it is not, what idea would replace it. In addition, the fact that a black hole cannot causally influence its exterior makes it difficult to understand the underlying mechanism by which thermal equilibrium could be achieved between a black hole and a material body. Secondly, why is the entropy so directly related to the area of the horizon? A formula of this type could only arise if all the degrees of freedom of a black hole were concentrated in a Planck length 'skin' around the horizon. Namely, if a finite number of states are assigned to each Planck volume in this region, then the logarithm of the total number of states would be proportional to $A$. However, ideas relating the degrees of freedom to the horizon run counter to the notion in classical general relativity of the black hole horizon as being a globally defined mathematical surface, possessing no local physical significance as was argued in section 1.4.1. We will come back to this idea in the context of black hole complementarity in chapter 5.

It is noteworthy that the temperature and entropy of a black hole involve Planck’s constant. In conventional units (with $\hbar$ and $c$ restored) we have

$$kT = \frac{\hbar \kappa}{2\pi c}$$

$$S = \frac{k c^3}{4G\hbar} A.$$  

(2.247)

(2.248)

The appearance of $\hbar$ in the expressions for the temperature and entropy of a black hole, which is a classical object from the point of view of the theory of general relativity, suggests that the study of black hole thermodynamics may lead to a deeper understanding of how gravitation and quantum theory are interrelated.

The ideas of thermodynamics seem to be deeply embedded in the theory of gravity and they have really shaped the search for a quantum description of gravity. This has resulted in thought-provoking papers that explain Einstein’s equation as an equation of state [69] and describe gravity as an emergent force, which is done in the so-called entropic gravity theory [70].

### 2.6 Euclidean path integral methods

Having promoted the mathematical analogy between black hole mechanics and thermodynamics to a real equivalence, this insight can be used to gain more understanding of the link between geometry and concepts as temperature and entropy. Still aware of the fact that there is not yet a satisfactory unification of gravity and quantum theory, we use Euclidean path integrals in a semiclassical approximation and their natural link with thermodynamics to get more clues about the principles of quantum gravity.
2.6.1 Hawking temperature derivation

In Minkowski spacetime, using Euclidean path integrals involves setting

\[ t = i\tau, \]  

(2.249)

and continuing \( \tau \) from imaginary to real values. Thus \( \tau \) is 'imaginary time' in this section. The Minkowski metric then becomes the ordinary Euclidean metric

\[ ds^2 = dt^2 + dx^2 + dy^2 + dz^2, \]  

(2.250)

where the metric is redefined to have positive coefficients. The invariance group is then not the Poincaré group with the Lorentz group as its homogeneous part, but it now contains the orthogonal group \( SO(4) \). Thus, Lorentz transformations are replaced by ordinary rotations

\[ z \rightarrow z \cos \beta + \tau \sin \beta, \]
\[ \tau \rightarrow -z \sin \beta + \tau \cos \beta, \]  

(2.251)

under which the metric (2.250) is invariant.

In the Schwarzschild spacetime the substitution \( t = i\tau \) leads to a continuation of the Schwarzschild metric to the Euclidean Schwarzschild metric

\[ ds^2_E = \left(1 - \frac{2GM}{r}\right)d\tau^2 + \left(1 - \frac{2GM}{r}\right)^{-1}dr^2 + r^2d\Omega^2. \]  

(2.252)

This metric is singular at \( r = 2GM \). To examine the region near \( r = 2GM \), one sets

\[ r - 2GM = \frac{\rho^2}{8GM}, \]  

(2.253)

to get

\[ ds^2_E \approx (\kappa \rho)^2d\tau^2 + d\rho^2 + \frac{1}{4\kappa^2}d\Omega^2. \]  

(2.254)

Not surprisingly, the first two terms of the metric near \( r = 2GM \) are that of Euclidean Rindler spacetime

\[ ds^2_E = d\rho^2 + \rho^2d(\kappa \tau)^2. \]  

(2.255)

This is just the Euclidean 2-plane if one makes the periodic identification

\[ \tau \sim \tau + \frac{2\pi}{\kappa}, \]  

(2.256)

which means that the singularity of the Euclidean Schwarzschild metric at \( r = 2GM \) is just a coordinate singularity provided that the imaginary time coordinate \( \tau \) is periodic with period \( 2\pi/\kappa \). So in Euclidean space, the transition to Rindler spacetime is nothing but a transition to cylindrical coordinates. This implies that the Euclidean functional integral must be taken over fields that are periodic with period \( 2\pi/\kappa \), i.e. \( \psi(x^i, \tau) = \psi(x^i, \tau + 2\pi/\kappa) \). Now, the Euclidean functional integral is

\[ Z = \int D\psi e^{-I_E[\psi]}, \]  

(2.257)
where
\[ I_E = \int dt \left( -i\pi \dot{\psi} + H \right), \tag{2.258} \]
with \( \pi \) the conjugate field, is the Euclidean action. If the functional integral is taken over fields that are periodic in imaginary time with period \( \hbar \beta \) the it can be written as [12]
\[ Z = \text{tr} e^{-\beta H}, \tag{2.259} \]
which is the partition function for a quantum mechanical system with Hamiltonian \( H \) at temperature \( T \) given by \( \beta = (kT)^{-1} \), where \( k \) is Boltzman’s constant.

But was just shown that \( \hbar \beta = 2\pi/\kappa \) for a Schwarzschild spacetime, so one deduces that a quantum field can be in equilibrium with a black hole only at the Hawking temperature. At any other temperature, the Euclidean Schwarzschild black hole has a conical singularity so there can be no equilibrium. It must be noted that the equilibrium at the Hawking temperature is unstable since if a black hole absorbs radiation its mass increases and its temperature decreases, so the a black hole has a negative heat capacity. However, this result should not be surprising on physical grounds, since an ordinary self-gravitating virialized star in Newtonian gravity also has a negative heat capacity. If one removes energy from a star, it contracts and heats up. As in the case of an ordinary star, this heat capacity does not imply any fundamental difficulty in describing the thermodynamics of black holes, since the microcanonical ensemble still should be well defined for a finite system containing a black hole, and a black hole can exist in a stable, thermal equilibrium in a sufficiently small box with walls that perfectly reflect radiation.

### 2.6.2 Black hole entropy derivation

From the full equivalence between black hole mechanics and thermodynamics, it follows that one should identify \( A/4G \) as the black hole entropy. One would now like to calculate this entropy from first principles, but this is not yet possible with the current theories. However, the Euclidean path integral provides a way to get an idea of where this entropy comes from.

We consider the entropy of a single static black hole. It will appear that the reason for gravitational configurations to be able have nonzero entropy is that the Euclidean solutions can have nontrivial topology. In other words, if one start with a static spacetime and identifies imaginary time with period \( \beta \), the manifold need not have topology \( S^1 \otimes \Sigma \) where \( \Sigma \) is some three manifold. In fact, for non-extreme black holes, the topology is \( S^2 \otimes \mathbb{R}^2 \), as was shown in the previous section.

To obtain the black hole entropy, the canonical partition function for the gravitational field is defined by a sum over all smooth Riemannian geometries [71], which satisfy some conditions to be specified below,
\[ Z(\beta) = \int \mathcal{D}g e^{-I[g]}, \tag{2.260} \]
where \( I[g] \) is the classical action of the geometry.

Suppose that it is a priori known that the spacetime includes a black hole. This imposes following conditions on the metrics considered in the path integral [72]:
1) \( g_{\mu\nu} \) possesses a Killing vector field \( \partial_\tau \),

2) There exists a surface \( \Sigma \), the horizon, which is a fixed point of the isometry generated by \( \partial_\tau \), i.e. where the Killing vector becomes null. In the asymptotically flat context, this means that the integration in (2.260) includes all asymptotically flat geometries with an isometry along a compact direction whose proper size at infinity is \( \beta \).

3) The asymptotic fall-off of the metric at large values of radial coordinate \( r \) is fixed by the mass \( M \) and electric charge \( Q \) of the configuration.

There are problems with the definition of this Euclidean path integral: these include the non-renormalizable UV divergences of gravity and the indefiniteness of the gravitational action, which is not even bounded from below. One should therefore view it as merely a semi-classical tool. That is, one should not view the sum over geometries as a fundamental definition of the theory.

Instead, we are interested in seeing what insight we can gain from considering the saddle-point approximation to this integral, which means that one puts

\[
\ln Z \approx -I_s,
\]

where \( I_s \) is the classical action of a Euclidean solution which satisfies the conditions above. There may be more than one such solution. One considers therefore the dominant contribution, which comes from the solution of least action

\[
I_s = I[g_s] \quad \text{with} \quad \frac{\delta I}{\delta g}[g_s] = 0.
\]

(2.262)

So the saddle-point approximation takes only the 'zero-loop' contribution into account. The expectation is that this approximation should give useful results if the classical solution is weakly curved, whatever the fundamental quantum theory may be. Since \( Z(\beta) \) is the canonical partition function, one has \( Z(\beta) = e^{-\beta F} = e^{-\beta \langle E \rangle + S} \). So the energy and entropy can be evaluated by the standard formulae

\[
\langle E \rangle = -\frac{\partial}{\partial \beta} \ln Z \approx \frac{\partial}{\partial \beta} I_s
\]

(2.263)

\[
S = \beta \langle E \rangle + \ln Z
\]

\[
= \left( \beta \frac{\partial}{\partial \beta} - 1 \right) \ln Z
\]

\[
= \left( \beta \frac{\partial}{\partial \beta} - 1 \right) I_s.
\]

(2.264)

There is an important topological difference between the Euclidean solutions which do and do not involve black holes. In for example the Euclidean flat space

\[
ds^2 = d\tau^2 + dr^2 + r^2d\Omega_{d-2}
\]

(2.265)
the Killing vector $\partial_\tau$ is non-vanishing throughout the entire spacetime. The radial coordinate ranges over $r \geq 0$ and $S^{d-2}$ shrinks to zero size at $r = 0$. One can identify $\tau$ periodically with any period one likes to choose. For cases with no black hole one can exploit the fact that global time is a Killing symmetry to write the action as

$$I = \int d^d x \, L = \int d\tau \int d^{d-1} x L = \beta H ,$$

where $H$ is the Hamiltonian. This can be done because constant time surfaces are well defined and one can consider Hamiltonian evolution from one surface to another. Hence, when such a geometry provides the dominant saddle point, $I_s$ is linear in $\beta$, and

$$S \approx \left( \beta \frac{\partial}{\partial \beta} - 1 \right) I = 0 .$$

That is, there is no classical contribution to the entropy for this solution, as expected.

On the other hand, for solutions with a black hole such a foliation by surfaces of constant time will necessarily break down in the interior of the horizon, where the $S^1$ degenerates. Another way to see this is that $\tau$ is no longer a time-like coordinate inside the horizon since the corresponding vector field $\partial_\tau$ becomes spacelike there. So one cannot make a foliation of constant time surfaces, needed for Hamiltonian evolution, which are expressed by some coordinate being constant to write down something like (2.266). For this reason, one should only restrict to the outer region $r \leq r_+$ to obtain a Hamiltonian description. One can split up the integration over the spacetime on the outer region into an integral over a small disc around the horizon at $r = r_+$ and the remaining, as shown in figure 2.5. This remaining integration over the bulk can be foliated with surfaces of constant $t$ and its contribution to the action will be linear in $\beta$ according to (2.266).

![Figure 2.5: Decomposition of the calculation of the action into a small region near the horizon and the remainder.](image)

One might think that the integration over the small disc would vanish in the limit as one takes the size of the disc to zero, since this is a smooth region of spacetime. However, this appears not to be the case. That is because in order to be able to write the integration over the bulk of the spacetime in Hamiltonian form, one has to be careful about how one breaks up the integration. More specifically, it appears that the Einstein-Hilbert action is not a good description of general relativity when boundaries are involved. This can be seen as follows. Take the usual (Lorentzian) Einstein-Hilbert Lagrangian density

$$L = \sqrt{-g} R$$
and apply a variation
\[ \delta L = \sqrt{-g} (\delta R_{\mu \nu}) g^{\mu \nu} + \sqrt{-g} R_{\mu \nu} \delta g^{\mu \nu} + R \delta (\sqrt{-g}). \] (2.269)

\[ g^{\mu \nu} \delta R_{\mu \nu} = \nabla^\mu v_\mu, \] (2.270)
with
\[ v_\mu = \nabla^\nu (\delta g_{\mu \nu}) - g^{\rho \sigma} \nabla_\mu (\delta g_{\rho \sigma}), \] (2.271)
and
\[ \delta (\sqrt{-g}) = \frac{1}{2} \sqrt{-g} g^{\mu \nu} \delta g_{\mu \nu} \\
= -\frac{1}{2} \sqrt{-g} g_{\mu \nu} \delta g^{\mu \nu}, \] (2.272)
the variation of the Einstein-Hilbert action can then be written as
\[ \delta I = \int d^d x \sqrt{-g} \nabla^\mu v_\mu + \int d^d x \sqrt{-g} \left( R_{\mu \nu} - \frac{1}{2} R g_{\mu \nu} \right) \delta g^{\mu \nu}. \] (2.273)

The second term on the right side will give rise to the Einstein equations. But the first term on the right hand side stands in the way. This term does not vanish for general variations where \( g^{\mu \nu} \) is held fixed on the boundary, although it does vanish for variations where the first derivatives of \( g_{\mu \nu} \) also are held fixed.

By Stoke’s theorem, this first term on the right side of (2.273) can be written as [11]
\[ \int_U d^d x \sqrt{-g} \nabla^\mu v_\mu = \int_{\partial U} d^{d-1} \sqrt{-h} v_\mu n^\mu, \] (2.274)
where \( U \) represents a general integration volume, \( n^\mu \) is the unit normal to the boundary \( \partial U \) and \( h_{\mu \nu} = g_{\mu \nu} \pm n_\mu n_\nu \) is the induced metric on \( \partial U \). Using the definition of \( v_\mu \) (2.271), one has
\[ v_\mu n^\mu = n^\mu g^{\rho \sigma} [\nabla_\rho (\delta g_{\mu \nu}) - \nabla_\mu (\delta g_{\rho \sigma})] \\
= n^\mu h^{\rho \sigma} [\nabla_\rho (\delta g_{\mu \nu}) - \nabla_\mu (\delta g_{\rho \sigma})] \\
= -n^\mu h^{\rho \sigma} \nabla_\mu (\delta g_{\rho \sigma}), \] (2.275)
where it was used that \( h^{\rho \sigma} \nabla_\sigma (\delta g_{\mu \nu}) = 0 \) because \( \delta g_{\mu \nu} = 0 \) on \( \partial U \). Now we define the trace of the extrinsic curvature of the boundary as
\[ K \equiv K_\mu^\mu = h^{\mu \nu} \nabla_\mu n^\nu. \] (2.276)

So the variation of \( K \) is
\[ \delta K = h^{\mu \nu} (\delta \Gamma)^\nu_{\mu \rho} n^\rho \\
= \frac{1}{2} n^\rho h^{\mu \nu} g^{\nu \lambda} [\partial_\rho (\delta g_{\mu \lambda}) + \partial_\nu (\delta g_{\mu \lambda}) - \partial_\lambda (\delta g_{\mu \nu})] \\
= \frac{1}{2} n^\rho h^{\mu \lambda} \partial_\rho (\delta g_{\mu \lambda}). \] (2.277)
So combining (2.277) and (2.275), the variation of the Einstein-Hilbert action (2.273) under variations of the metric for which \( \delta g_{\mu\nu} = 0 \) can be written as

\[
\delta I = -2 \int_{\partial U} d^{d-1} \sqrt{-h} \delta K + \int_U d^d x \sqrt{-g} G_{\mu\nu} \delta g^{\mu\nu}.
\] (2.278)

In fact, (2.278) continues to hold if one allows variations of \( g_{\mu\nu} \) for which only the induced metric on the boundary is held fixed, \( \delta h_{\mu\nu} = 0 \). This can be verified directly or deduced from the fact that if \( \delta h_{\mu\nu} = 0 \) on the boundary, one can find a gauge transformation \( \nabla_{\mu} l_{\nu} + \nabla_{\nu} l_{\mu} = 0 \) on the boundary which makes \( \delta g_{\mu\nu} = 0 \). Since (2.278) holds for all variations with \( \delta g_{\mu\nu} = 0 \) on \( \partial U \) and since all terms in (2.278) are invariant under such gauge transformations, this equation must continue to hold for variations which merely satisfy \( \delta h_{\mu\nu} = 0 \).

It follows from (2.278) that the unwanted term in the variation of the Einstein-Hilbert action can be removed by modifying the action. We define

\[
I' = I + 2 \int_{\partial U} d^{d-1} x \sqrt{-h} K.
\] (2.279)

Then the extremization of \( I' \) yields the desired result. Thus, when boundary terms are taken into account, \( I' \) is the appropriate action to use for general relativity.

So the action for the small disc of figure 2.5 is

\[
I_d = \frac{1}{16 \pi G} \int_D d^d x \sqrt{g} R + \frac{1}{8 \pi G} \int_{\partial D} d^{d-1} y \sqrt{h} K,
\] (2.280)

where the determinant of the metric is positive now because the Euclidean metric is used. The surface term can be rewritten as

\[
\int_{\partial D} d^{d-1} y \sqrt{g} K = -\frac{\partial}{\partial n} \int_{\partial D} d^{d-1} y \sqrt{h}
\] (2.281)

For the small disc near the horizon, one can use the approximate metric (2.254), so one obtains

\[
\int_{r = r_+ + \epsilon} d^{d-1} y \sqrt{h} = 2 \pi \epsilon A,
\] (2.282)

where \( A \) is the area of the horizon. Therefore, it follows that

\[
\frac{\partial}{\partial n} \int_{\partial D} d^{d-1} y \sqrt{h} = 2 \pi A.
\] (2.283)

Hence, in the limit \( \epsilon \to 0 \), the small disc around \( r = r_+ \) makes a contribution

\[
I_{\text{disc}} = -\frac{1}{4G} A,
\] (2.284)

which gives

\[
S = \left( \beta \frac{\partial}{\partial \beta} - 1 \right) I_s = \left( \beta \frac{\partial}{\partial \beta} - 1 \right) I_{\text{disc}} = \frac{1}{4G} A.
\] (2.285)

This calculation provides a direct link between geometry and entropy. As in the calculation of
the Hawking temperature for a quantum field in the previous section, regularity of the geometry at the horizon plays a crucial role in the derivation. Note that the explicit form of the geometry was not used in this derivation, just the fact that the geometry is smooth there. Thus, this derivation explains the universality of the relation between entropy and area. Note also that the explicit form of the action is used, so the result depends on the gravitational dynamics, unlike the calculation of the temperature.

The Euclidean path integral method describes a canonical ensemble. But as already mentioned in the previous section, a black hole has a negative heat capacity so it cannot exist in a stable thermal equilibrium with an ordinary heat bath at fixed temperature as measured at infinity. So this presents a problem. This problem also manifests itself by the fact that $A = 16^2 M^2$ for a Schwarzschild black hole, so assuming the usual interpretation of entropy, the density of states of a Schwarzschild black hole should grow with $M$ as $\exp(4\pi M^2)$. However, in that case, the sum in (2.259) would not converge. Thus, there appears to be a logical inconsistency in the Euclidean path integral calculation of the black hole entropy, since the result of the calculation would seem to invalidate the method used to derive it. But it is shown that these problems can be overcome by redefining the canonical ensemble or by using the microcanonical ensemble [73–75].

The saddle-point calculation of the black hole entropy does not offer any insight into the nature of the microstates the entropy is counting. However, there is evidence from black hole pair creation that the black hole entropy is really counting microstates [71]. To explicitly identify these microstates, a concrete microscopic theory of quantum gravity is needed.
Chapter 3

The membrane paradigm

"It’s by logic that we prove, but by intuition that we discover.”
- H. Poincaré (1908)

At this point we’ve established the viewpoint in which black holes are truly thermodynamical objects. Although there were already some hints about their thermal nature in the classical description, this remains a remarkable feature. This property again emphasises the special nature of black holes and how important it is to find a correct way to think about these objects without losing some crucial physical aspects.

Based on their thermodynamical behavior, and some parallel discoveries we will discuss in this chapter, a new mental picture of black holes emerged. For reasons explained below it is called the membrane paradigm and it enables us to describe in a very intuitive way how the physics of an outside observer is influenced by the presence of a black hole. The membrane paradigm is very powerful to describe black holes as dynamical objects which interact with their environment.

Although the membrane paradigm is founded completely on general relativity, it will play a crucial role in the quantum description of black holes in later chapters.

3.1 The stretched horizon

As in section 1.8.3, we will again consider the field lines of a charged particle near a black hole. The analytic solution for the electric field of the particle at rest on the polar axis \((θ = 0)\) at radius \(r_0\) outside a Schwarzschild black hole is [10]

\[
\vec{E} = \frac{Q}{r_0^2} \left[ \frac{GM}{D} \left( 1 - \frac{r_0 - GM + GM \cos θ}{D} \right) \right] \vec{e}_r \\
+ r \left[ (r - M)(r_0 - GM) - G^2M^2 \cos θ \right] \left[ r - GM - (r_0 - GM) \cos θ \right] \frac{1}{D^3} \vec{e}_r \\
+ \left[ \frac{Q(r_0 - 2GM)\sqrt{1 - 2GM/r \sin θ}}{D^3} \right] \vec{e}_θ, \\
\tag{3.1}
\]

115
with

\[ D \equiv \left( (r - GM)^2 + (r_0 - GM)^2 - G^2M^2 - 2(r - GM)(r_0 - GM) \cos \theta + G^2M^2 \cos^2 \theta \right)^{1/2}. \] (3.2)

If the particle is reasonably far out (for example at radius \( r_0 = 5MG \) in diagrams (a) and (b) of figure 3.1), then its field lines are only modestly distorted by the hole. But if the particle is very close to the horizon (for example, at \( r_0 = 2.1GM \) in diagram (d)), its field lines are so strongly distorted that more distant observers see a nearly radial field emerging from the black hole’s center, not from the particle’s position.

![Figure 3.1: A point charge and its field lines at different distances from the horizon.](image)

Now the idea of the membrane paradigm is to stretch the horizon a little bit outwards (the dashed line in diagram (d)) so that it entirely covers up the particle. In this way one produces a picture in which the field lines emerge radially from the stretched horizon, as though it were endowed with a uniform charge density and the particle had totally disappeared down the black hole.
The electric field of a dynamically infalling particle behaves similarly to this sequence of static fields. Although the particle does not cross the horizon at any finite Schwarzschild time $t$, soon after it passes the stretched horizon its field behaves as though its electric charge had been deposited on and smeared uniformly over the stretched horizon. In the next sections we will show that the membrane paradigm even has sufficient power to describe the dynamical evolution of the (apparent) charge on the stretched horizon as it smears itself out.

The membrane paradigm is mathematically equivalent to the standard, full, general relativistic theory of black holes, so far as all physics outside the horizon is concerned. It adopts a frozen-star-like view of physics outside the horizon, but it contains within itself a simple prescription for ignoring 'irrelevant' near-horizon details in astrophysical problems. More specifically, in this viewpoint particles and fields very near the horizon possess a highly complex, frozen, boundary-layer structure which is essentially a relic history of the black hole's past. This complex boundary layer has no influence on the present or future evolution of particles and fields above the boundary layer. In a way the membrane viewpoint stretches the horizon to cover up the boundary layer and then imposes simple membrane-like boundary conditions on the stretched horizon. This sweeping away of irrelevancies entails small and in practice negligible errors, but it results in a remarkably powerful formalism.

Next to the electrical behavior described above, the horizon appeared to have other interesting properties. It was discovered in [76] that external gravitational fields can tidally deform the horizon of a black hole and the motion of the deformation produces entropy just as if the horizon were viscous. So combining the electrical, viscous and thermodynamical behavior, the horizon appears to behave like a hot, charged fluid. But, there is a difficulty with describing processes very near the horizon because of the 'freezing' of motion at the horizon. This difficulty is resolved by the stretching the horizon, where the null horizon is replaced with a time-like physical membrane endowed with electrical, mechanical and thermodynamical properties. So the role of the stretched horizon is two-sided: covering up irrelevancies and allowing real dynamics which give rise to a fluid interpretation.

It is important to always keeps in mind that the membrane viewpoint is a very convenient mental picture to describe the observations of an outside observer. As dictated by the equivalence principle, an infalling observer will just see ordinary, flat spacetime at the horizon. The hot fluid at the horizon only exists for outside observers.

It should also be noted that the membrane paradigm uses a 3+1 split of spacetime. This means a preferred family of 3-dimensional space-like hypersurfaces is chosen as surfaces of constant time and then is treated as though they were a single 3-dimensional space that evolves as time passes. So 4-dimensional spacetime is decomposed into 3-dimensional space plus 1-dimensional time. The general relativistic physics of black holes, plasmas and accretion disks takes place in this 3-dimensional space. And the relativistic laws that govern them, written in 3-dimensional language, resemble the nonrelativistic laws. Thus, the 3+1 formulation is well suited to carrying physicists’ nonrelativistic intuition about plasmas and hydrodynamics into the arena of black holes and general relativity.

A first indication that the membrane paradigm could also be of importance in the quantum
description of black holes was given in [77], where it was suggested that the entropy of a black hole could be the logarithm of the total number of quantum mechanically distinct configurations that can exist in the covered-up boundary layer. In chapter 5, this idea will appear to be one of the founding principles of black hole complementarity.

In the next sections we will calculate some properties of the stretched horizon, again focusing on its electrical behavior. The membrane paradigm of course greatly extends the electromagnetic applications presented here and for an excellent overview is referred to [10].

3.2 A conducting surface

As mentioned in the previous section, stretching the horizon has the very useful benifit that one describes a time-like system instead of a light-like system. This means that real dynamics and evolution can take place on the stretched horizon. In this section we will study the near-horizon dynamics by considering the electromagnetic field equations.

Because we will work only on the outside of and very close to the horizon, we can use the approximate Rindler metric

$$ds^2 = \rho^2 d\omega^2 - d\rho^2 - dx_{\perp}^2,$$

where we used (2.199) in Cartesian coordinates and defined the dimensionless Rindler time as $\omega = \kappa t$. We take $\rho$ along the $z$-direction and $x_{\perp} = (x, y)$.

The stretched horizon is defined as the surface

$$\rho = \rho_0,$$

where $\rho_0$ is very small (we will later take it to be the Planck length $l_p = \sqrt{\hbar G/c^3}$).

The action for the electromagnetic field in Rindler spacetime is [9]

$$I = \int \left[ -\frac{\sqrt{-g}}{16\pi} F^{\mu\nu} F_{\mu\nu} + J^\mu A_\mu \right] d\omega d\rho d^2x_{\perp}. \quad (3.5)$$

As usual, $J$ is a conserved current in the sense that $\partial_\mu J^\mu = 0$.

By using the metric (3.3) we can calculate

$$-\frac{\sqrt{-g}}{16\pi} F^{\mu\nu} F_{\mu\nu} = -\frac{\sqrt{-g}}{16\pi} g^{\mu\alpha} g^{\nu\beta} F_{\mu\nu} F_{\alpha\beta}$$

$$= \frac{\rho}{8\pi} \left( \frac{1}{\rho^2} F_{\omega\rho} F_{\omega\rho} + \frac{1}{\rho^2} F_{\omega\rho} F_{\omega\rho} - F_{i\rho} F_{i\rho} \right), \quad (3.6)$$

where a summation over $i = x, y$ is understood. By putting $A_\mu = (-\phi, A_\rho, A_x, A_y)$ we can write (3.6) as

$$-\frac{\sqrt{-g}}{16\pi} F^{\mu\nu} F_{\mu\nu} = \frac{1}{8\pi} \left( \frac{1}{\rho} (\dot{A}_\rho + \partial_\rho \phi)^2 + (\dot{A}_i + \partial_i \phi)^2 - \rho (\partial_\rho A_\rho - \partial_i A_i)^2 \right), \quad (3.7)$$
where $\dot{A}$ represents $\frac{\partial \vec{A}}{\partial \omega}$. So the action (3.5) becomes
\[
I = \int \left[ \frac{1}{8\pi} \left( \frac{1}{\rho} (\dot{A} + \vec{\nabla} \phi)^2 - \rho (\vec{\nabla} \times \vec{A})^2 \right) + J \cdot A \right] d\omega d\rho d^2x_\perp ,
\] (3.8)

The electric and magnetic field are defined in the conventional way
\[
\vec{E} = -\vec{\nabla} \phi - \dot{\vec{A}} \tag{3.9}
\]
\[
\vec{B} = \vec{\nabla} \times \vec{A} \tag{3.10}
\]

In terms of the electric and magnetic field, the action becomes
\[
I = \int \left[ \frac{1}{8\pi} \left( \frac{1}{\rho} |\vec{E}|^2 - |\rho \vec{B}|^2 \right) + J \cdot A \right] d\omega d\rho d^2x_\perp , \tag{3.11}
\]

and the Maxwell field equations are
\[
\frac{1}{\rho} \dot{\vec{E}} - \vec{\nabla} \times (\rho \vec{B}) = -4\pi \vec{J} \tag{3.12}
\]
\[
\dot{\vec{B}} + \vec{\nabla} \times \vec{E} = 0 \tag{3.13}
\]
\[
\vec{\nabla} \cdot \left( \frac{1}{\rho} \vec{E} \right) = 4\pi J^0 \tag{3.14}
\]
\[
\vec{\nabla} \cdot \vec{B} = 0 . \tag{3.15}
\]

We first consider electrostatics. By electrostatics is meant the study of fields due to stationary or slowly moving charges placed outside the horizon. Since the charges are slowly moving in Rindler coordinates, this means that they are experiencing proper acceleration. We will also assume all length scales associated with the charges are much larger than $\rho_0$. In particular, the distance of the charges from the stretched horizon is macroscopic.

The surface charge density on the stretched horizon is defined as the component of the electric field perpendicular to the stretched horizon
\[
\sigma = \frac{1}{4\pi \rho} E^\rho \bigg|_{\rho=\rho_0} \tag{3.16}
\]
\[
= -\frac{1}{4\pi \rho} \partial_\rho \phi \bigg|_{\rho=\rho_0} . \tag{3.17}
\]

If we work in the Coulomb gauge $\vec{\nabla} \cdot \vec{A} = 0$, (3.14) becomes
\[
\vec{\nabla} \cdot \left( \frac{1}{\rho} \vec{E} \right) = -\vec{\nabla} \cdot \left( \frac{1}{\rho} \vec{\nabla} \phi \right) = 0 , \tag{3.18}
\]

because $J^0 = 0$ near the horizon. Thus
\[
\partial_\rho^2 \phi - \frac{1}{\rho} \partial_\rho \phi = -\nabla_\perp^2 \phi \tag{3.19}
\]
This equation can be solved near the horizon by the ansatz $\phi \sim \rho^\alpha$. The right hand side will be smaller than the left hand side by two powers of $\rho$ and can therefore be ignored. We then find

$$\alpha(\alpha - 1)\rho^{\alpha-2} - \alpha\rho^{\alpha-2} = 0,$$

so $\alpha$ has to be either 2 or 0. So we can write the general solution as

$$\phi = F(x_\perp) + \rho^2 G(x_\perp) + \text{terms higher order in } \rho.$$  

(3.21)

Filling in this form for $\phi$ in equation (3.19) and evaluating at $\rho = \rho_0$ gives

$$\nabla^2_{\perp} F + \rho_0^2 \nabla^2_{\perp} G = 0.$$  

(3.22)

Since $\rho_0$ is much smaller than all other length scales this becomes

$$\nabla^2_{\perp} F = 0.$$  

(3.23)

Since the black hole horizon is compact, this implies that the only possible solution on the horizon is

$$\phi = \text{constant},$$  

(3.24)

which confirms that the horizon behaves like an electrical conductor. From this we can deduce that the field lines of a point charge need to be perpendicular to the stretched horizon, just as they would be with a normal metal object. This is shown on figure 3.2.

**Figure 3.2:** The field lines of a point charge near the horizon.

We can now even try to determine the resistivity of the stretched horizon. To do so, we identify the surface current density. By taking the time derivative of the charge density (3.17) and using the Maxwell equation (3.12) with $\vec{J} = 0$ one gets

$$4\pi \dot{\sigma} = \frac{1}{\rho_0} \dot{E}_\rho = (\vec{\nabla} \times \rho \vec{B})_\rho.$$  

(3.25)

This equation can be interpreted as a continuity equation if one defines the current as

$$4\pi j_x = -\rho B_y,$$

$$4\pi j_y = \rho B_x.$$  

(3.26) (3.27)
Now consider an electromagnetic wave propagating towards the stretched horizon along the $\rho$ axis. From Maxwell’s equations one obtains

\begin{align*}
\dot{B}_x &= \partial_\rho E_y \\
\dot{B}_y &= -\partial_\rho E_x \\
\frac{1}{\rho} \dot{E}_x &= -\partial_\rho (\rho B_y) \\
\frac{1}{\rho} \dot{E}_y &= \partial_\rho (\rho B_x).
\end{align*}

(3.28) (3.29) (3.30) (3.31)

One can make these equations more familiar by redefining the magnetic field

$$\rho \tilde{B} = \tilde{\beta}$$

(3.32)

and using the coordinate

$$u = \log \rho.$$  

(3.33)

One then gets

\begin{align*}
\dot{\beta}_x &= \partial_u E_y \\
\dot{\beta}_y &= -\partial_u E_x \\
\dot{E}_x &= \partial_u \beta_y \\
\dot{E}_y &= -\partial_u \beta_x.
\end{align*}

(3.34) (3.35) (3.36) (3.37)

These equations allow solutions in which the wave can propagate in either direction along the $u$-axis. However, the physics only makes sense for waves propagating towards the horizon from outside the black hole. For such waves, these equations give

\begin{align*}
\beta_x &= E_y \\
\beta_y &= -E_x.
\end{align*}

(3.38) (3.39)

So from (3.26) and (3.27) we get for the surface current

\begin{align*}
j_x &= \frac{1}{4\pi} E_x \\
j_y &= \frac{1}{4\pi} E_y.
\end{align*}

(3.40) (3.41)

This allows us to conclude that the resistivity of the stretched horizon is $4\pi$. One can take this role of a conductor for the stretched horizon very literally. If a circuit is constructed as in figure 3.3, a current will flow precisely as if the horizon were a conducting surface.

### 3.3 Spreading of a charge

One could now drop a charged particle onto the horizon and compute the time for the charge to equilibrate. Since the horizon is an electric conductor the charge density will quickly become uniform. Without loss of generality, we can take the charge to be at rest at position $z_0$ in
Minkowski coordinates. The freely falling point charge is depicted in Minkowski coordinates on figure 3.4.

The calculation is easy because at any given time the Rindler coordinates are related to the Minkowski coordinates by a boost along the $z$-axis. Since the component of the electric field along the boost direction is invariant, one can write the standard Coulomb field

$$E_\rho = E_z$$

$$= \frac{e(z - z_0)}{\left[(z - z_0)^2 + x_\perp^2\right]^{3/2}}$$

$$= \frac{e(\rho \cosh \omega - z_0)}{\left[(\rho \cosh \omega - z_0)^2 + x_\perp^2\right]^{3/2}},$$

where relation (2.84) between the Minkowski $z$-coordinate and the Rindler coordinates was used.
Using the definition of the surface density (3.16), one finds

\[ \sigma = \frac{e}{4\pi \rho_0} \frac{\rho_0 \cosh \omega - z_0}{[(\rho_0 \cosh \omega - z_0)^2 + x_{\perp}^2]^{3/2}}. \]  

(3.45)

Now let’s consider the surface density for large Rindler time

\[ \sigma = \frac{e}{4\pi \rho_0} \frac{\rho_0 e^\omega}{[(\rho_0 e^{2\omega} + x_{\perp}^2)^{3/2}}. \]  

(3.46)

It is convenient to rescale \( x_{\perp} \) using \( x_{\perp} = e^{\omega} y_{\perp} \) to obtain

\[ \sigma = \frac{e}{4\pi} \frac{e^{-2\omega}}{[(\rho_0^2 + y_{\perp}^2)^{3/2}}. \]  

(3.47)

We can now use this expression to calculate how fast the charge gets spread across the entire stretched horizon. We will assume the Rindler time is big enough so that we can neglect \( y_{\perp}^2 \) in the denominator of (3.47). We then get that the charge is uniform when

\[ 4\pi \rho_0^3 e^{2\omega} = 4\pi R_s^2 \rho_0, \]  

(3.48)

where \( R_s \) is the Schwarzschild radius of the black hole. This can be solved for \( \omega \)

\[ \omega = \log \left( \frac{R_s}{\rho_0} \right), \]  

(3.49)

or in terms of the Schwarzschild time

\[ t = \frac{1}{\kappa} \log \left( \frac{R_s}{\rho_0} \right) \]  

(3.50)

\[ = 4MG \log \left( \frac{R_s}{\rho_0} \right) \]  

(3.51)

\[ \sim R_s \log \left( \frac{R_s}{\rho_0} \right). \]  

(3.52)

This exponential spreading of the charge is characteristic of an Ohm’s law conductor. To see this, use Ohm’s law \( j = \text{conductivity} \ E \). By taking the divergence one gets

\[ \nabla \cdot j \sim \nabla \cdot \bar{E} \sim \sigma. \]  

(3.53)

By using the continuity equation \( \dot{\sigma} + \nabla \cdot j = 0 \) one finds

\[ \dot{\sigma} \sim -\sigma, \]  

(3.54)

which evidently predicts the surface charge density will decrease exponentially. Conservation of charge will then cause the charge to spread exponentially.

The result of this section can be extended to more general situations. In particular, we can consider (3.52) as the typical timescale for a black hole to reestablish equilibrium after a small perturbation. So (3.52) gives the timescale at which an outside observer looses track of the particle that fell down the black hole. It therefore states how fast a black hole looses its hair.
When restricting to the electromagnetic field, the Ohmic behavior of the stretched horizon is actually completely equivalent to the statement that black holes have no hair.
Chapter 4

Entanglement and information

If you don’t see the use of it, I certainly won’t let you clear it away. Go away and think. Then, when you can come back and tell me that you do see the use of it, I may allow you to destroy it.
- G.K. Chesterton on paradoxes (1929)

In chapter 1, we saw that the no hair conjecture implies that black holes effectively destroy information at the classical level. This wasn’t a problem since a classical black hole would last forever because of the area theorem and the information could be thought of as preserved inside it, but just not very accessible. Also, the loss of classical information is not in conflict with any other principle of nature.

However, the situation changes drastically when quantum effects are taken into account. In chapter 2, it was shown that black holes lose energy because of the emission of particles to infinity. This causes them to shrink, and -most likely- to completely vanish after a long period of time. But now one can compare the situation before and after the presence of the black hole. More specifically, one can compare the state of the matter that collapsed and formed the black hole with the state of the radiation that is the end product of the evaporation process. Is the information about the initial matter still present in the final radiation? This may look like a far-fetched and irrelevant question, but in fact it is of crucial importance. Because as we will see in this chapter, the loss of information is incompatible with quantum mechanics.

Before we can adress these problems, a clear definition of the term ‘information’ in quantum theory is needed. We shall see that it is intimately related to other important concepts like entanglement and entropy. In this chapter, all these concepts are introduced and are used to give a more complete description of the quantum aspects of black holes. For the most important of them, the information paradox, the complete context and a detailed description is given.

4.1 Density matrices and entanglement

In this section we introduce the concepts of a density matrix, entanglement and entanglement entropy which have a fundamental role in the remainder of this thesis.
### 4.1.1 Ensembles

In quantum mechanics, there are two basic types of ensembles [78]. A pure ensemble is a collection of physical systems such that every member is characterized by the same ket $|\alpha\rangle$. In contrast, in a mixed ensemble, a fraction of the members with relative population $w_1$ are characterized by $|\alpha^{(1)}\rangle$, some other fraction with relative population $w_2$ by $|\alpha^{(2)}\rangle$, and so on. Roughly speaking, a mixed ensemble can be viewed as a mixture of pure ensembles, just as the name suggests. The fractional populations are constrained to satisfy the normalization condition

$$\sum_i w_i = 1. \quad (4.1)$$

It should be noted that the states $|\alpha^{(1)}\rangle$ and $|\alpha^{(2)}\rangle$ need not be orthogonal. Furthermore, the number of terms in the sum (4.1) need not coincide with the dimensionality $N$ of the Hilbert space, it can easily exceed it. For example, for spin $1/2$ systems with $N = 2$, one may consider 40% with spin in the positive $z$-direction, 30% with spin in the positive $x$-direction and the remaining 30% with spin in the negative $y$-direction.

The expectation value of an operator $A$ in a mixed ensemble is given by

$$\langle A \rangle = \sum_i w_i \langle \alpha^{(i)} | A | \alpha^{(i)} \rangle = \sum_i \sum_{\lambda} w_i |\langle \lambda | \alpha^{(i)} \rangle|^2 \lambda, \quad (4.2)$$

where $|\lambda\rangle$ is the eigenbasis of $A$. Notice how probabilistic concepts enter twice in this equation: first in $|\langle \lambda | \alpha^{(i)} \rangle|^2$ for the quantum mechanical probability for the state $|\alpha^{(i)}\rangle$ to be found in the eigenstate $|\lambda\rangle$, and second in the probability factor $w_i$ for finding in the ensemble a state $|\alpha^{(i)}\rangle$.

We can now rewrite the ensemble average (4.2) using a more general basis $\{|k\rangle\}$

$$\langle A \rangle = \sum_i w_i \sum_{k,l} \langle \alpha^{(i)} | k \rangle \langle k | A | l \rangle \langle l | \alpha^{(i)} \rangle = \sum_{k,l} \left( \sum_i w_i \langle l | \alpha^{(i)} \rangle \langle \alpha^{(i)} | k \rangle \right) \langle k | A | l \rangle. \quad (4.3)$$

The number of terms in the sum over $k,l$ is just the dimensionality of the Hilbert space, whereas the number of terms in the sum over $i$ depends on how the mixed ensemble is viewed as a mixture of pure ensembles. Notice that in this form, the basic property of the ensemble that does not depend on the particular observable $A$ is factored out. This is the motivation to define the density operator as

$$\rho \equiv \sum_i w_i |\alpha^{(i)}\rangle \langle \alpha^{(i)}|. \quad (4.4)$$

With this definition, we can now write the ensemble average (4.3) as

$$\langle A \rangle = \text{tr}(\rho A). \quad (4.5)$$
Because the trace is independent of representations, \( \text{tr}(\rho A) \) can be evaluated using any convenient basis.

The density operator has two very important properties. First, from its definition (4.4) it is immediately clear that \( \rho \) is Hermitian. Second, the density operator satisfies the normalization condition

\[
\text{tr}(\rho) = \sum_i \sum_k w_i \langle \alpha^{(i)} \rangle \langle \alpha^{(i)} | k \rangle \\
= \sum_i w_i \langle \alpha^{(i)} | \alpha^{(i)} \rangle \\
= 1. \tag{4.6}
\]

A pure ensemble is specified by \( w_i = 1 \) for some \( |\alpha^{(i)}\rangle \), with \( i = n \) for example, and \( w_i = 0 \) for all other conceivable states. The corresponding density operator is written as

\[
\rho = |\alpha^{(n)}\rangle \langle \alpha^{(n)}|. \tag{4.7}
\]

Clearly, the density operator for a pure ensemble is idempotent

\[
\rho^2 = \rho. \tag{4.8}
\]

Thus, for a pure ensemble one has

\[
\text{tr}(\rho^2) = 1 \tag{4.9}
\]

Because \( \rho \) is idempotent for a pure ensemble, it also follows that its eigenvalues are zero or one. It can be shown that \( \text{tr}(\rho^2) \) is maximal when the ensemble is pure. For a mixed ensemble, \( \text{tr}(\rho^2) \) is a positive number less than 1.

One should not conclude from its definition (4.4) that \( \rho \) is always diagonal. This is because the \( |\alpha^{(i)}\rangle \) don’t have to be an orthogonal set. The density matrix in a basis \( \{|k\rangle\} \) is obtained via

\[
\sum_i |\alpha^{(i)}\rangle \langle \alpha^{(i)}| = \sum_{k,l} \left( \sum_i \langle k | \alpha^{(i)} \rangle \langle \alpha^{(i)} | l \rangle \right) |k\rangle \langle l|. \tag{4.10}
\]

### 4.1.2 Quantum statistical mechanics

The density operator formalism is the basis of quantum statistical mechanics. To establish the connection, first consider a completely random ensemble. The density matrix for such an ensemble can be written in some orthonormal basis \( \{|k\rangle\} \) as

\[
\rho = \sum_k \frac{1}{N} |k\rangle \langle k|, \tag{4.11}
\]

where \( N \) is again the dimension of the Hilbert space. So all its eigenvalues are equal and given by \( 1/N \). In fact, the representation (4.11) is independent of the choice of basis. So (4.11) represents an ensemble where all states are equally populated.
We saw in the previous section that the density matrix of a pure ensemble has only a single nonzero eigenvalue which is equal to one. So the density matrix of a pure and random ensemble cannot look more different. It would be desirable to construct a quantity that characterizes this difference. Thus we define
\[
S = -\text{tr}(\rho \ln \rho).
\] (4.12)

The logarithm of an operator is defined via a Taylor expansion. But a more straightforward evaluation is available when working with the basis in which \(\rho\) is diagonal. Denoting the eigenvalues of \(\rho\) by \(\lambda_i\), we obtain
\[
S = -\sum_i \lambda_i \ln \lambda_i.
\] (4.13)

So we get for a pure and a random ensemble
\[
S_{\text{pure}} = 0 \quad (4.14)
\]
\[
S_{\text{random}} = \ln N. \quad (4.15)
\]

It is now argued that physically, \(S\) can be regarded as a quantitative measure of disorder. A pure ensemble is an ensemble with a maximum amount of order because all members are characterized by the same state. For such a state \(S\) is zero. At the other extreme, a completely random ensemble, in which all states are equally likely, has maximum disorder. For a random ensemble \(S\) is very large, we will show later that \(\ln N\) is even the maximum possible value for \(S\) subject to the normalization condition \(\sum_i \lambda_i = 1\). So we conclude \(S\) can be identified with the entropy (note we take \(k = 1\), which is done at all times throughout this thesis).

It is now shown how the density matrix can be obtained for an ensemble in thermal equilibrium. The basic assumption is that nature tends to maximize \(S\) subject to the constraint the the ensemble average of the Hamiltonian has a certain prescribed value. Once thermal equilibrium is established, one has
\[
\frac{\partial \rho}{\partial t} = 0. \quad (4.16)
\]

And because of the Heisenberg evolution equation it follows that
\[
[H, \rho] = 0, \quad (4.17)
\]

which means that \(\rho\) and \(H\) can be simultaneously diagonalized. So we will use the energy eigenbasis to represent the density operator. With this choice, \(\lambda_k\) represents the fractional population for an energy eigenstate with energy eigenvalue \(E_k\).

The expectation value of the Hamiltonian is given by
\[
\langle H \rangle = \text{tr}(\rho H) = U, \quad (4.18)
\]

where \(U\) is the internal energy per constituent. So the energy constraint is
\[
\delta \langle H \rangle = \sum_k \delta \lambda_k E_k = 0. \quad (4.19)
\]
The normalization constraint is
\[ \delta(\text{tr}\rho) = \sum_k \delta\lambda_k = 0. \] (4.20)

We now want to maximize \( S \) by requiring
\[ \delta S = 0, \] (4.21)
subject to the constraints (4.19) and (4.20). This is most readily accomplished by using Lagrange multipliers. One obtains
\[ \sum_k \delta\lambda_k[(\ln \lambda_k + 1) + \beta E_k + \gamma] = 0, \] (4.22)
which for an arbitrary variation is possible only if
\[ \lambda_k = \exp(-\beta E_k - \gamma - 1). \] (4.23)

By using the normalization condition \( \sum_k \lambda_k = 1 \), the final result is
\[ \lambda_k = \frac{e^{-\beta E_k}}{\sum_l e^{-\beta E_l}}, \] (4.24)
which directly gives the fractional population for an energy eigenstate with eigenvalue \( E_k \). The sum is over distinct eigenstates, if there is degeneracy one must sum over states with the same energy eigenvalue.

The density matrix element (4.24) corresponds to the canonical ensemble. Had we maximized \( S \) without the internal-energy constraint, we would have obtained
\[ \lambda_k = \frac{1}{N}. \] (4.25)

This is the density matrix element of a completely random ensemble. Comparing (4.24) and (4.25), it follows that the completely random ensemble can be seen as the high temperature limit \( \beta \to 0 \) of a canonical ensemble.

The denominator of (4.24) can be recognized as the partition function
\[ Z = \sum_k e^{-\beta E_k}. \] (4.26)

It can also be written as
\[ Z = \text{tr}(e^{-\beta H}). \] (4.27)

And finally, the density operator can be cast into the form
\[ \rho = \frac{e^{-\beta H}}{Z}. \] (4.28)
4.1.3 Reduced density matrix

In the previous sections the density operator was used to describe ensembles. The obtained entropy was the conventional entropy from thermodynamics. Because if we are ignorant about the state of the system, we assign a probability to each state. This lead to an entropy of ignorance, also referred to as the thermal entropy.

In this section however, we will consider a completely different type of entropy, which has a purely quantum mechanical origin. It is this form of entropy that is of most interest for the purposes of this thesis. And although it has a completely different origin than the entropy of ignorance or thermal entropy, it can be described using the same density matrix formalism.

The entropy that we will consider here results from the superposition principle and by considering subsystems of a larger system which is in a pure state. For example, take two spin 1/2’s labeled $a$ and $b$ who are in the singlet state

$$|\psi\rangle = \frac{1}{\sqrt{2}}(|\uparrow\rangle_a |\downarrow\rangle_b - |\downarrow\rangle_a |\uparrow\rangle_b).$$

This is our total system, described by the pure state $|\psi\rangle$. But now we are interesting in only a subsystem, say spin $a$. Notice that it is impossible to write $|\psi\rangle$ as a tensor product of two other states which describe $a$ and $b$ separately

$$|\psi\rangle \neq |\psi^{(1)}\rangle_a \otimes |\psi^{(2)}\rangle_a.$$  

If this were true we would say that $|\psi\rangle$ is a product state, in which case it would be very straightforward to describe the spin $a$ individually. If (4.30) holds, we say that $a$ and $b$ are entangled.

Since we cannot describe $a$ by a single state vector, we will have to assign it a density matrix. First, construct the density matrix corresponding to the total pure state

$$\rho_{ab} = |\psi\rangle\langle\psi|$$

$$\begin{align*}
\rho_{ab} &= \frac{1}{2}(|\uparrow\rangle_a |\downarrow\rangle_b + |\downarrow\rangle_a |\uparrow\rangle_b - |\uparrow\rangle_a |\downarrow\rangle_b - |\downarrow\rangle_a |\uparrow\rangle_b) \\
&= \frac{1}{2}(|\uparrow\rangle_a |\uparrow\rangle_b + |\downarrow\rangle_a |\downarrow\rangle_b)
\end{align*}$$

One can now construct the reduced density matrix for the spin $a$ by tracing out the spin $b$

$$\rho_a = \text{tr}_b(\rho_{ab})$$

$$\begin{align*}
\rho_a &= b(\uparrow | \rho_{ab} | \uparrow) + b(\downarrow | \rho_{ab} | \downarrow) + \frac{1}{2} (| \uparrow \rangle a (| \uparrow \rangle_a | \uparrow \rangle_b) + | \downarrow \rangle a (| \downarrow \rangle_a | \downarrow \rangle_b) \\
&= \frac{1}{2} | \uparrow \rangle a (| \uparrow \rangle_a + \frac{1}{2} | \downarrow \rangle_a | \downarrow \rangle_a.
\end{align*}$$

From $\rho_a$ we can now deduce that there is a probability of 1/2 to find $a$ as an up-spin and a probability of 1/2 to find it as a down-spin. This is no surprise when one looks at the original pure state $|\psi\rangle$ of the total system. Because the reduced density matrix $\rho_a$ is completely random,
Chapter 4. Entanglement and information

In the general case one considers a quantum system composed of two subsystems $A$ and $B$. Assume the Hilbert space $H$ is a tensor product space

$$H = H_A \otimes H_B.$$ (4.33)

If $\{|i\}\}$ is an orthonormal basis for $H_A$ and $\{|j\}\}$ is an orthonormal basis for $H_B$, then a general state $|\psi\rangle$ in $H$ may be written as

$$|\psi\rangle = \sum_{i,j} c_{ij} |i\rangle \otimes |j\rangle.$$ (4.34)

The reduced density matrix of the subsystem $A$ in the basis $\{|i\}\}$ is

$$\langle i | \rho_A | i' \rangle = \rho_A(i,i') = \sum_j c_{ij} c_{i'j}^*,$$ (4.35)

and that of $B$ is

$$\langle j | \rho_A | j' \rangle = \rho_B(j,j') = \sum_i c_{ij} c_{ij'}^*.$$ (4.36)

Note that we’ve again taken the total system to be in a pure state. The procedure above can of course also be applied to the situation where $A$ and $B$ are subsystems of a total system which is not pure. We will not consider this case explicitly here.

In complete analogy to (4.12), we can now associate an entropy with each subsystem via

$$S_A = -\text{tr}(\rho_A \ln \rho_A)$$ (4.37)

$$S_B = -\text{tr}(\rho_B \ln \rho_B).$$ (4.38)

This entropy is called entanglement entropy. It is of a completely different nature than the thermal entropy described above. Thermal entropy results from the human ignorance in describing a complex system. Entanglement entropy comes from an inherent indeterminacy in the state of a subsystem because of its quantum mechanical correlations with another subsystem. It should be noted that the second law of thermodynamics only concerns thermal entropy, so the entanglement entropy can increase or decrease with time.

The entanglement entropy of a subsystem is zero only if the state $|\psi\rangle$ of the total system is an uncorrelated product state. Denote the dimension of $H_B$ by $|B|$ and that of $H_A$ by $|A|$. If $|A| > |B|$, then the maximum value of $S_B$ is

$$S_B = \ln |B|,$$ (4.39)

which corresponds to a completely random state for $B$.

Entanglement entropy satisfies two important inequalities [79]. The first is called subadditivity and is given by

$$|S_A - S_B| \leq S_{AB} \leq S_A + S_B.$$ (4.40)
The second involves three subsystems $A$, $B$ and $C$ and states
\[ S_{ABC} + S_B \leq S_{AB} + S_{BC}. \]  
(4.41)
This inequality is called strong subadditivity.

Heuristically, entanglement entropy can also be thought of as the lack of information one has about the state of a (sub)system. Because a total pure state has entanglement entropy zero and two correlated subsystems each have nonzero entanglement entropy, this shows that for quantum information, the whole system contains more information than the sum of the information in the separate parts. The state of the total system contains information about the quantum mechanical correlations between the different subsystems. It is this information that gets lost by considering the density matrix of an individual subsystem. So by tracing out a subsystem one does not only remove the information contained within that subsystem, but also the information contained in the correlations between the two subsystems. Entanglement entropy will have a key role in the discussion of quantum black holes. In appendix E, the two manifestations of entanglement entropy which are most important for the purposes of this thesis are put forward and compared to each other.

### 4.2 Unruh density matrix

As a first application of the concepts introduced in the previous section, we come back to the Unruh effect of section 2.2.2. There, it was shown that a accelerating observer experiences the Minkowski vacuum as a thermal bath of particles
\[ \langle 0_M | a^\omega_i a^\omega_i | 0_M \rangle = \langle 0_M | a^L_i a^L_i | 0_M \rangle = \frac{1}{e^{2\pi\omega_i/a} - 1}, \]  
(4.42)
where a discretization was applied for convenience. This indicates that the Minkowski vacuum can be expressed as a thermal state in the right Rindler wedge with the boost generator as the Hamiltonian.

It should be emphasized that in section 2.2.2, the conclusion that the Minkowski vacuum restricted to the left or right Rindler wedge is a thermal state was actually taken too soon. Showing that the expectation value of the number operators has the correct form is not enough. It is necessary to show that the probability of each right/left Rindler-energy eigenstate corresponds to the grand canonical ensemble if the other Rindler wedge is disregarded. One can show this fact by using the discrete version of equations (2.135) and (2.137) of section 2.2.2, which are given here
\[ (a^\omega_i - e^{-\pi\omega_i/a} a^{L^i}_\omega)|0\rangle_M = 0 \]  
(4.43)
\[ (a^L_i - e^{-\pi\omega_i/a} a^{R^i}_\omega)|0\rangle_M = 0. \]  
(4.44)
Chapter 4. Entanglement and information

Multiplying (4.43) with \(a_{\omega_i}^R\) and (4.44) with \(a_{\omega_i}^L\) from the right, subtracting both equations and using the fact that \(a_{\omega_i}^R\) and \(a_{\omega_i}^L\) commute, results in

\[
(a_{\omega_i}^R a_{\omega_i}^R - a_{\omega_i}^L a_{\omega_i}^L)|0\rangle_M = 0.
\]

Thus, the number of left Rindler particles is the same as that of the right Rindler particles for each \(\omega_i\). This implies that one can write

\[
|0\rangle_M \propto \prod_i \sum_{n_i = 0}^\infty \frac{K_{n_i}}{n_i!} (a_{\omega_i}^R)^{n_i} (a_{\omega_i}^L)^{n_i} |0\rangle_R.
\]

One can find the recursion formula satisfied by \(K_{n_i}\) using the relations (4.43) and (4.44). First, one finds that

\[
e^{-\pi \omega_i / a} a_{\omega_i}^L |0\rangle_M \propto e^{-\pi \omega_i / a} \prod_i \sum_{n_i = 0}^\infty \frac{K_{n_i}}{n_i!} (a_{\omega_i}^R)^{n_i} (a_{\omega_i}^L)^{n_i} a_{\omega_i}^L |0\rangle_R.
\]

And secondly

\[
a_{\omega_i}^R |0\rangle_M \propto a_{\omega_i}^R \prod_i \sum_{n_i = 0}^\infty \frac{K_{n_i}}{n_i!} (a_{\omega_i}^R)^{n_i} (a_{\omega_i}^L)^{n_i} |0\rangle_R
\]

\[
= \prod_i \sum_{n_i = 0}^\infty \frac{K_{n_i}}{(n_i - 1)!} (a_{\omega_i}^R)^{n_i} (a_{\omega_i}^L)^{n_i} a_{\omega_i}^L |0\rangle_R
\]

\[
= \prod_i \sum_{n_i' = 0}^\infty \frac{K_{n_i'+1}}{n_i'!} (a_{\omega_i}^R)^{n_i'} (a_{\omega_i}^L)^{n_i'} a_{\omega_i}^L |0\rangle_R.
\]

So combining (4.43), (4.47) and (4.48), one gets

\[
K_{n_i+1} - e^{-\pi \omega_i / a} K_{n_i} = 0.
\]

Hence, \(K_{n_i} = e^{-\pi n_i \omega_i / a} K_0\) and

\[
|0\rangle_M = \prod_i \left( C_i \sum_{n_i = 0}^\infty e^{-\pi n_i \omega_i / a} n_i |n_i, R\rangle \otimes |n_i, L\rangle \right),
\]

where

\[
C_i = \sqrt{1 - e^{-2\pi \omega_i / a}}
\]

is a normalization constant. Here, the state with \(n_i\) left-moving particles with Rindler energy \(\omega_i\) in each of the left and right Rindler wedges is denoted by \(|n_i, R\rangle \otimes |n_i, L\rangle\), i.e.

\[
\prod_i |n_i, R\rangle \otimes |n_i, L\rangle = \left[ \prod_i \frac{1}{n_i!} (a_{\omega_i}^R a_{\omega_i}^L)^{n_i} \right] |0\rangle_R.
\]
Chapter 4. Entanglement and information

If one probes only the right Rindler wedge, then the Minkowski vacuum is described by the density matrix obtained by tracing out the left Rindler states, which leads to

$$\rho_R = \prod_i \left( C_i^2 \sum_{n_i=0}^{\infty} e^{-2\pi n_i \omega_i / a} |n_i, R\rangle \langle n_i, R| \right).$$

This is exactly the density matrix for a system of free bosons with temperature $T = a/2\pi$. Thus, now it is allowed to conclude that the Minkowski vacuum state $|0\rangle_M$ for the left-moving particles restricted to the left (or right) Rindler wedge is the thermal state with temperature $a/2\pi$ with the boost generator normalized on $z^2 - t^2 = 1/a^2$ as the Hamiltonian. This is the Unruh effect for the right-moving sector. It is clear that the Unruh effect for the left-moving sector can be derived in a similar manner.

### 4.3 Generalized second law for quasistationary semiclassical black holes

As a second application of the concepts introduced in section 4.1, we come back to the generalized second law of black hole thermodynamics, which was discussed in section 2.5. There, it was stated that the total entropy of a system containing a black hole does not decrease. However, no proof was given. Only two concrete processes were considered and verified to satisfy the generalized second law. Here, a more general argument or even a proof for the generalized second law is presented. The proof was first given by Page and Frolov in [80] and we follow their procedure.

The reasoning below proves the validity of the generalized second law for quasistationary changes of a generic charged, rotating black hole emitting, absorbing and scattering any sort of radiation in the semiclassical formalism, i.e. quantum fields in the classical spacetime background of a black hole whose conserved quantities change by the expectation value of the flux of radiation out or into it.

A quasistationary black hole may be considered to emit a density matrix $\rho_0$ of thermal radiation. These modes will be referred to as the $UP$ modes. Suppose that there is also radiation with density matrix $\rho_1$ incident on the black hole from far away, e.g. from past null infinity, in modes that are called $IN$ modes. These incoming modes are of positive frequency at $I^-$. The semiclassical approximation is used, and it is assumed that the radiation in these two sets of modes will be quantum mechanically uncorrelated, i.e. the initial density matrix is given by a product state

$$\rho_{\text{initial}} = \rho_{01} = \rho_0 \otimes \rho_1.$$  \hspace{1cm} (4.54)

This assumption is natural for an eternal black hole. For it, the $UP$ modes, which are defined to be of positive frequency with respect to the Killing vector field of which the past event horizon $H^-$ is a Killing horizon, vanish at $I^-$, whereas the $IN$ modes vanish at $H^-$ and $I^-$ and $H^-$ are causally disconnected.
In the case in which the black hole arises from gravitational collapse and becomes quasistationary, the \(UP\) modes are defined to be the same in the future stationary region as the \(UP\) modes of the eternal black hole with the same future stationary region. They are nonvanishing at \(I^-\) at the advanced time at which the black hole forms. This can be seen on figure 4.1.

![Figure 4.1](image)

**Figure 4.1:** The UP and IN modes and their region of support at \(I^-\).

However the \(IN\) and \(UP\) modes generally have a different region of support at \(I^-\), there is a small overlap around \(v_0\), i.e. the advanced time of the last geodesic that can escape to infinity. One might therefore worry that the \(UP\) modes in principle could be correlated with the \(IN\) modes which come from \(I^-\) at much later advanced time. However, after the hole has become quasistationary, the relevant \(UP\) modes trace back to such high energy modes at \(I^-\) that the state in those modes must be extremely close to being unpopulated there. Thus, in the quasistationary approximation, they will have totally negligible correlations with the \(IN\) modes coming in much later in advanced time. That is why, for the physics of the quasistationary region at late time, both pictures (that of eternal black holes and that of black holes arising from gravitational collapse) give very nearly the same results. For concreteness, the eternal black hole picture will be used in the following discussion.

After the initial state \(\rho_{01}\) interacts with the classical angular momentum and curvature barrier separating the horizon from infinity (see section 2.4), and possibly interacts with itself as well, it will have evolved unitarily into a -generally- correlated final state

\[
\rho_{\text{final}} = \rho_{23} \neq \rho_2 \otimes \rho_3 , \tag{4.55}
\]

where

\[
\rho_2 = \text{tr}_3 \rho_{23} \tag{4.56}
\]

is the density matrix of the radiation in the \(OUT\) modes escaping to future null infinity \(I^+\), and

\[
\rho_3 = \frac{1}{2} \rho_{23} \tag{4.57}
\]
is the density matrix of the $\text{DOWN}$ modes that are swallowed by the future horizon $H^+$. All the modes are depicted on figure 4.2.

![Figure 4.2: The UP, IN, OUT and DOWN modes.](image)

As seen in section 4.1, the entropy of each of these states is

$$S_i = -\text{tr}(\rho_i \ln \rho_i).$$

(4.58)

Because the evolution from $\rho_{01}$ to $\rho_{23}$ is unitary, one has that $S_{01} = S_{23}$. Furthermore, since $\rho_{01}$ is uncorrelated but $\rho_{23}$ is generically partially correlated, the entropies of these states obey the inequality

$$S_2 + S_3 \geq S_{23} = S_{01} = S_0 + S_1.$$

(4.59)

The first law of black hole mechanics (see section 1.11.3) for a black hole of mass $M$, angular momentum $J$, charge $Q$, angular velocity $\Omega_H$ and electrostatic potential $\Phi$ states that

$$\Delta S = \frac{1}{T_H}(\Delta M - \Omega_H \Delta J - \Phi \Delta Q) = \frac{1}{T} \Delta E,$$

(4.60)

where $T_H = \kappa/2\pi$ is the Hawking temperature and $T$ and $E$ are the local temperature and energy as measured by an observer corotating with the hole near the horizon.

If $E_0$ and $E_3$ are the expectation values of the local energies of the emitted state $\rho_0$ and the absorbed state $\rho_3$ respectively, then the semiclassical approximation, combined with (4.60), yields

$$\Delta S = \frac{1}{T}(E_3 - E_0),$$

(4.61)

assuming that the changes to the black hole are sufficiently small that $T$ stays approximately constant throughout the process, which is again the quasistationary approximation.
Now (4.61) and (4.59) imply that the change in the generalized, total entropy is

$$\Delta S' = \Delta S + \Delta S_{\text{rad}}$$

$$= \frac{1}{T}(E_3 - E_0) + S_2 - S_1$$

$$\geq (S_0 - \frac{E_0}{T}) - (S_3 - \frac{E_3}{T}). \quad (4.62)$$

Now for fixed $T$ and equivalent quantum systems, as are the $UP$ modes of $\rho_0$ and the corresponding $DOWN$ modes of $\rho_3$ by $CPT$ invariance, $S - T^{-1}E$ is a Massieu function, which is essentially the negative of the local free energy divided by the temperature, and is maximized by the thermal state. The calculation of the Hawking radiation of section 2.3.1 implies that $\rho_0$ is thermal, so it follows from (4.62) that

$$\Delta S' \geq 0, \quad (4.63)$$

which is the generalized second law. This is an explicit mathematical demonstration of the fact that the generalized second law is a special case of the ordinary second law, with the black hole as a hot, rotating, charged body that emits thermal radiation uncorrelated with what is incident upon it.

### 4.4 The information paradox

Thus far, the discovery of particle creation by black holes had nothing but positive consequences. It provided black holes with a nonvanishing physical temperature and promoted the analogy between thermodynamics and black hole physics to a true equivalence. But in this section, it will be shown that the black hole radiation process also has a very cumbersome downside when one considers 'life beyond the black hole'. The problem is called the 'information paradox' and was first put forth by Hawking in [81].

To give the essential features of this paradox, a toy model for particle creation by black holes is presented which will give an outline to what the mechanism creating the paradox is. It is especially useful in showing why there is something like the information paradox in black hole evaporation, but not in the black body radiation of a burning piece of coal. It also very easily demonstrates a common misconception about the information paradox.

Finally, we leave the toy model for what it is and say a few things about the true physical situation, arguing that black hole formation and evaporation truly suffers from the problems presented in the toy model.

#### 4.4.1 A toy model

First, some concepts that were just silently assumed before will now be defined exactly. More specifically, we will define what is implicitly assumed when a quantum field is put in a curved
background. After that, another view on particle creation is presented that will be used to expose the difficulties of black hole evaporation. This will be done according to [82].

4.4.1.1 Nice slices

The reason why we trust the outcomes of putting a quantum field on a curved background is that we believe there is an appropriate limit where the effects of quantum gravity becomes small, and a local, well defined approximate evolution equation becomes possible. This limit underlies all of our physical thinking. This low energy limit is called the semiclassical approach.

In this subsection, a set of 'niceness conditions' are introduced such that under these conditions physics can be described by a known, local evolution equation. This implies that under the niceness conditions, one can specify the quantum state on an initial space-like slice, and then a Hamiltonian evolution operator gives the state on later slices. This viewpoint is based upon the Hamiltonian formulation of general relativity, which is presented in appendix D. Furthermore, locality implies that the influence of the state in one region on the evolution in another region must go to zero as the distance between these regions goes to infinity.

The niceness conditions are

1) The quantum state is defined on a space-like slice $\Sigma$ which intrinsic three-curvature $R^{(3)}$ should be much smaller than the Planck scale everywhere: $R^{(3)} \ll l_p^{-2}$.

2) $\Sigma$ is nicely embedded in an 4-dimensional spacetime, i.e. its extrinsic curvature $K$ is small everywhere: $K \ll l_p^{-2}$.

3) The four-curvature of the full spacetime in the neighbourhood of the slice should be small everywhere: $R \ll l_p^{-2}$.

4) Any quanta on the slice should have wavelength much longer than the Planck length, $\lambda \gg l_p$, and the energy density $e$ and momentum density $p$ should be small everywhere compared to the Planck density: $e, p \ll l_p^{-4}$. The matter on the slice also satisfies the usual energy conditions.

5) The state on $\Sigma$ will be evolved to later slices. All slices encountered should be 'good' as above. Further, the the lapse and shift vectors needed to specify the evolution should change smoothly with position: $\frac{dN^i}{ds} \ll l_p^{-1}$, $\frac{dN}{ds} \ll l_p^{-1}$.

It will be shown below that these niceness conditions, together with requiring locality leads to 'unacceptable' physical evolution for black hole evaporation. One must therefore either agree to this 'unacceptable evolution', or find a way to add new conditions to the set above in such a way that these conditions still allow us to define a proper low energy limit incorporating some idea of locality. But first, we come back to the process of particle creation by quantum fields in curved backgrounds.
4.4.1.2 Particle creation revisited

Now that we have given a (hopefully complete) set of conditions such that we can ignore quantum gravity effects in our spacetime, we reconsider the process of particle creation in the light of these ‘nice slices’.

Start with the vacuum state on the lower slice in figure 4.3(a). Consider the evolution to the upper slice shown in the figure. The later slice is evolved forward in the right hand region more than in the left hand region. This is of course allowed, in general relativity time is ‘many-fingered’, in the language of Wheeler, so one can evolve in any way that he likes. The slices are of course assumed to satisfy the niceness conditions.

![Image](image_url)

Figure 4.3: Space-like slices in an evolution with particle creation.

The evolution of the geometry will lead to particle creation in the region where the geometry of the slice is being deformed, this happens because as was shown before in chapter 2, the vacuum state on one slice will not in general be the natural vacuum state on a later slice. Let the geometry in the deformation region be characterized by the length and time scale $L$. Then the particle pairs created have wavelengths $\lambda \sim L$, and the number of such created pairs is $n \sim 1$. Why it is possible to say this, and why the creation process may be located at the deformation region will be explained below. The particle pair is depicted by $c, b$ in figure 4.3(b). As seen in chapter 2, the state of the created pair is of the thermal form

$$\langle \Psi \rangle_{\text{pair}} = Ce^{\gamma c^\dagger b^\dagger} |0\rangle_c |0\rangle_b,$$

where $\gamma$ is a number of order unity. The essence of the entanglement in this state can be obtained by assuming the following simple form for the state

$$\langle \Psi \rangle_{\text{pair}} = \frac{1}{\sqrt{2}} (|0\rangle_c |0\rangle_b + |1\rangle_c |1\rangle_b).$$

There also is some matter in a state $|\psi\rangle_M$ on the space-like slice, but the crucial point is that this matter is very far away, at a distance $L' \gg L$, from the place where the pair creation is taking place.
If one now assumes locality on the space-like slices, then the complete state on the space-like slice would be

$$|\Psi\rangle_{\text{pair}} \approx |\psi\rangle_M \otimes \frac{1}{\sqrt{2}} (|0\rangle_c|0\rangle_b + |1\rangle_c|1\rangle_b). \quad (4.66)$$

Even though the matter is far away from the place where the pairs are being created, there will always be some effect of $|\psi\rangle_M$ on the state of the created pairs. This is why there is an $\approx$ written in (4.66).

Now let the state of matter $|\psi\rangle_M$ consist of a single spin which can be up or down. Let’s take

$$|\psi\rangle_M = \frac{1}{\sqrt{2}} (|\uparrow\rangle + |\downarrow\rangle). \quad (4.67)$$

Then if there was no effect of the matter state on the state of the created pairs, the state on the slice would be

$$|\Psi\rangle \approx \frac{1}{\sqrt{2}} (|\uparrow\rangle + |\downarrow\rangle) \otimes \frac{1}{\sqrt{2}} (|0\rangle_c|0\rangle_b + |1\rangle_c|1\rangle_b). \quad (4.68)$$

It is crucial to understand that locality allows small departures from (4.68), for example

$$|\Psi\rangle = \frac{1}{\sqrt{2}} (|\uparrow\rangle + |\downarrow\rangle) \otimes \left( (\frac{1}{\sqrt{2}} + \epsilon)|0\rangle_c|0\rangle_b + (\frac{1}{\sqrt{2}} - \epsilon)|1\rangle_c|1\rangle_b \right), \quad (4.69)$$

but not a completely different state like

$$|\Psi\rangle = \frac{1}{\sqrt{2}} (|\uparrow\rangle|0\rangle_c + |\downarrow\rangle|0\rangle_c) \otimes \frac{1}{\sqrt{2}} (|0\rangle_b + |1\rangle_b). \quad (4.70)$$

### 4.4.1.3 Slicing the black hole geometry

The discussion below will apply to all black holes, but for concreteness, consider the Schwarzschild metric

$$ds^2 = \left(1 - \frac{2GM}{r}\right) dt^2 - \left(1 - \frac{2GM}{r}\right)^{-1} dr^2 - r^2 d\Omega^2. \quad (4.71)$$

Aan essential property of black holes for the discussion below, is the no-hair conjecture discussed in section 1.8. There is no information about the hole in the vicinity of the horizon. Or in other words, the horizon is ‘information-free’. To make this more precise, around every point at the horizon one can find a neighborhood which is the vacuum. This means that the evolution of field modes with wavelengths $l_p << \lambda < M$ is given by the semiclassical evolution of quantum fields on empty curved space up to terms that vanish as $m_p/M \to 0$.

Note that it was stated in chapter 2 that there is no unique definition of particles in a general curved spacetime. But if the curvature radius is $R$ then for wavemodes with wavelength $\lambda < R$, one can get a definition of particles in which one can say what the vacuum is.

Now we would like to define a family of nice slices for the black hole geometry. Is is clear that one should avoid the singularity if one wants to keep the niceness conditions satisfied. A space-like slice in a black hole geometry which satisfies the niceness conditions is constructed as follows.
Chapter 4. *Entanglement and information*

1) For \( r > 4GM \), let the slice be \( t = t_1 = \text{constant} \).

2) Inside \( r < 2GM \), the space-like slices are \( r = \text{constant} \) rather than \( t = \text{constant} \). Let the slice be \( r = r_1 \), with \( M/2 < r_1 < 3M/2 \), so that this part of the slice is not near the horizon \( r = 2MG \) and not near the singularity \( r = 0 \).

3) The parts of (1) and (2) are joined by a smooth "connector" segment which obeys the niceness conditions.

4) The Schwarzschild metric gives an eternal black hole, but we will be interested in black holes resulting from gravitational collapse. With such a spacetime, one can follow the \( r = r_1 \) part of the slice down to early times before the hole was formed, and then smoothly extend it to \( r = 0 \) when there was no singularity.

This makes one complete nice space-like slice, which is called \( S_1 \) in figure 4.4.

![Figure 4.4: Schematic representation of the Schwarzschild black hole with nice spacelike slices.](image)

Now let’s consider how to make a ‘later’ slice \( S_2 \).

1) At \( r > 4MG \), take \( t = t_1 + \Delta \).

2) The \( r = \text{constant} \) part will be \( r = r_1 - \delta_1 \), with \( \delta_1 << M \). Note that the time-like direction for this part of the geometry is in the decreasing \( r \) direction. Let \( \delta_1 \) be small, and later the limit \( \delta_1 \to 0 \) will be taken.

3) The parts from (1) and (2) are again joined by a smooth connector segment. In the limit \( \delta_1 \to 0 \), the geometry of the connector segment can be taken to be the same for all slices. Note
that the \( r = \text{constant} \) part of the later slice is longer than the \( r = \text{constant} \) part of \( S_1 \).

4) At early times, again bring the \( r = \text{constant} \) part smoothly down to \( r = 0 \), at a place where there is no singularity.

To describe the nature of the evolution from \( S_1 \) to \( S_2 \), choose lapse and shift vectors on the spacetime as follows. Take the slice \( S_1 \) and pick a point \( x^i \) on it. Now move along the time-like normal till a point on \( S_2 \) is reached. Let this point on \( S_2 \) have the same spatial coordinates \( x^i \). Thus, the shift vector is \( N^i = 0 \). With this choice, one can describe the evolution as follows:

1) In the \( t = \text{constant} \) part of the slice, there is no change in intrinsic geometry. This part of the slice just advances forward in time with a lapse function \( N = \left( 1 - \frac{2GM}{r} \right)^{1/2} \).

2) In the limit \( \delta_1 \to 0 \), the \( r = \text{constant} \) part of \( S_1 \) moves over to \( S_2 \) with no change in intrinsic geometry. The early time part which joins this segment \( r = 0 \) also remains unchanged.

3) The connector segment of \( S_1 \) has to stretch during this evolution since the corresponding points on \( S_2 \) will have to cover both the connector of \( S_2 \) and the extra part of the \( r = \text{constant} \) segment of \( S_2 \).

Thus, the stretching happens only in the region near the connector segment. This region has space and time dimensions of order \( GM \). Evolution from \( S_2 \) to later slices can be done in a completely analogous way.

Note that while the Schwarzschild metric looks time independent, this is only an illusion because the Schwarzschild coordinates break down at the horizon. Any slicing will necessarily be time-dependent. The crucial point is that although the geometry is independent of \( t \), yet one cannot make a space-like slicing which covers both the outside and the inside of the hole and is time-independent. This is because the Killing vector field \( \partial/\partial t \) is time-like at infinity, but is not time-like everywhere. Thus the \( t = \text{constant} \) surface is not space-like everywhere. If one does try to foliate the spacetime with space-like slices then one finds that these slices `stretch’ during the evolution. So actually, the geometry is not truly static since there is no global Killing vector field.

The interesting thing about the stretching between successive slices is that it happens in a given place, so that the Fourier modes of fields at this location keep getting stretched to larger wavelengths and particles will keep being produced. Thus, the time-dependence of the slices is the reason for particle creation in the black hole geometry because as a consequence the Fourier decomposition of a field is not invariant under evolution between the slices. This process is sketched on figure 4.5. The longer wavelengths will distort to a nonuniform shape first, and thereby create an entangled pair. The modes with shorter wavelength evolve for some more time before suffering the same distortion, and then create an entangled pair.

One cannot have such a set of slices in ordinary Minkwoski spacetime. If one tries to make such slices in Minkwoski spacetime, then after some point in the evolution the later slices will not be spacelike everywhere: the stretching part will become null and then timelike. But it is
the basic feature of black hole geometries that the space and time directions interchange roles inside the horizon, and one gets space-like slices having a stretching like that of figure 4.4. This interpretation also ties in with the fact that the temperature of a black hole is proportional to its surface gravity $\kappa$, since $\kappa$ is a measure of 'how fast' the Killing vector field generating time translations at infinity becomes space-like around the horizon.

![Figure 4.5](image)

**Figure 4.5:** A fourier mode on the initial space-like slice is evolved to later space-like slices. ($\tau$ is a schematic time coordinate since this is not a Penrose diagram illustrating the actual spacetime structure of the geometry)

### 4.4.1.4 From pure to mixed

The members of the particle pairs which are created according to the mechanism described above that float out to infinity are called the Hawking radiation. The pairs will form a state which is entangled in a very specific way, and this fact lies at the heart of the information paradox. It is crucial that the state of these pairs is a state unlike any that is created when a normal hot body radiates photons. It will appear that the essential difference arises from the fact that in the black hole case the particle pairs are the result of the stretching of a region of the space-like slice, i.e. these pairs are 'pulled out of the vacuum'. In normal hot bodies the radiation is emitted from the constituents making up the hot body. This is the essential difference between a hot body and the black hole.

![Figure 4.6](image)

**Figure 4.6:** The creation of Hawking pairs.
Consider an initial space-like slice. The shell that collapsed to make the hole is represented by a matter state $|\psi\rangle_M$. As seen above, in the evolution to the next space-like slice the middle part of the space-like slice stretches, while the left and right parts remain unchanged. The stretching creates correlated pairs, labelled $b_1$ and $c_1$, and the state on the complete slice is

$$|\Psi\rangle \approx |\psi\rangle_M \otimes \frac{1}{\sqrt{2}} (|0\rangle_{c_1}|0\rangle_{b_1} + |1\rangle_{c_1}|1\rangle_{b_1}) .$$  \hspace{1cm} (4.72)$$

The no-hair conjecture is of crucial importance to be able to write down this state. If a black hole did have hair, then the region where the pair was created would contain degrees of freedom capable of storing information about the collapsed matter. In that case, the leading order behavior would drastically deviate from the tensor product in (4.72) and the reasoning below leading to the information paradox would fail.

The entanglement of $b_1$ with the $M,c_1$ system is

$$S_{\text{ent}} = \ln 2 .$$  \hspace{1cm} (4.73)$$

If one computes the entanglement of the set $\{b_1,b_2\}$ with the system $\{M,c_1,c_2\}$, one gets

$$S_{\text{ent}} = 2\ln 2 .$$  \hspace{1cm} (4.75)$$

It is now easy to see that after $N$ such steps, the state on the slice becomes

$$|\Psi\rangle_M \approx |\psi\rangle_M \otimes \frac{1}{\sqrt{2}} (|0\rangle_{c_1}|0\rangle_{b_1} + |1\rangle_{c_1}|1\rangle_{b_1}) \otimes \frac{1}{\sqrt{2}} (|0\rangle_{c_2}|0\rangle_{b_2} + |1\rangle_{c_2}|1\rangle_{b_2}) \otimes \cdots \otimes \frac{1}{\sqrt{2}} (|0\rangle_{c_N}|0\rangle_{b_N} + |1\rangle_{c_N}|1\rangle_{b_N}) .$$  \hspace{1cm} (4.76)$$

and the entanglement entropy of the $\{b_i\}$ set with the $\{M,c_i\}$ system is

$$S_{\text{ent}} = N\ln 2 .$$  \hspace{1cm} (4.77)$$

As the quanta $\{b_i\}$ collect at infinity, the mass of the hole decreases. The slicing does not satisfy the niceness conditions after the point when the mass of the black hole approaches the Planck mass because then $R << l_p^{-2}$ is no longer true. We will therefore stop evolving the space-like slices when this point is reached. Although one can not say what will happen beyond the semiclassical approximation until a quantum theory of gravity is established, we will assume here that the black hole evaporates completely. In further sections, we will consider the possibility
that quantum gravitational effects halt the evaporation process.

According to (4.77), the quanta \( \{ b_i \} \) have an entanglement entropy of \( N \ln 2 \). But when the black hole evaporates completely, there is nothing left to be entangled with so the final state can not be described by any quantum wave function or pure state. The final state is mixed and can only be described by a density matrix. But this leads to a loss of unitarity since a pure state, i.e. the state of the matter that collapsed to form a black hole, evolves to a mixed state, which is in conflict with the principles of quantum mechanics. So we get

**The information paradox**  *If one tries to analyze the evolution of a black hole using the usual principles of relativity and quantum theory, one is led to a contradiction, for these principles forbid the evolution of a pure state to a mixed state.*

To recapitulate the outline above, what we have seen is that at each stage of the evolution the entanglement entropy of the \( b_i \) increases by \( \ln 2 \). The evolution is very unique to the black hole because the radiation is created by the stretching of connector segment of the space-like slices. When normal hot bodies radiate, the radiation quanta are not created by stretching of space-like slices. Thus for normal hot bodies the radiation quanta depend on the nature of the atomic state at the surface of the body. by contrast, in black hole evolution the matter making the hole stays far away from the place where the Hawking pairs are being created. In fact, with each successive stage of stretching, the matter is removed further away from the place where the next pair would be produced.

To see how far the matter is from the creation of the typical Hawking pair for a solar mass black hole, note that after each stage of stretching, the matter moves a distance of order \( GM \sim 3 \) km away from the place where the pairs are being created. The number of radiation quanta is \( (M/m_p)^2 \). Thus after about half the evolution, the distance of the matter measured along the space-like slice to the place where the pairs are being created is of order

\[
L' \sim M \left( \frac{M}{m_p} \right)^2 \approx 10^{77} \text{ light years}.
\]  

(4.78)

This shows the sharp contrast between the black body radiation of normal hot bodies and of black holes. For normal bodies, the distance between the matter in the body and the place where the radiation is created would be zero since the radiation leaves from the atoms in the body.

One might think that even though the matter is very far away from where the pairs are being created, the pairs which have been created recently are close to the new pair being created, and this may help to generate correlations. Again, one finds that this does not happen. For one thing, the earlier created quanta also move away from the pair creation region at each step. Thus the typical created quantum is also a distance of order \( 10^{77} \) light years from the place where the new pairs are being created. Of course the pairs have been created recently are at a distance \( \sim 3 \) km from the newly created pair. But the nature of the pair creation process is such that this nearness does not help. The new pair is created by the stretching of a new Fourier mode, and the earlier created pair is simply pushed away in this process.
It should be noted that the EPR pairs used in this toy model were used so that the reasoning behind the information paradox could be followed easily. As noted in [83, 84], they are not appropriate to describe the real physical situation since they have the possibility to teleport the information about the matter to the Hawking radiation through annihilations of the negative energy quanta $c_i$ with the positive energy quanta of the collapsed matter.

4.4.1.5 Mixed states and information

Even though the (seemingly) non-unitary evolution of black hole formation and evaporation is commonly called 'the information paradox', the problem raised is not really centered on information, but rather on the mixed nature of the radiation state. In fact one can make radiation states that have full information about the hole but are still mixed, and conversely, one can have the radiation state as a pure unmixed state and yet carry no information about the hole.

Suppose the matter state is $|\psi\rangle_M = \alpha|\uparrow\rangle + \beta|\downarrow\rangle$. Now assume that the process of evolution creates two pairs, with the full state being described as follows

$$
\alpha|\uparrow\rangle + \beta|\downarrow\rangle \to \frac{1}{\sqrt{2}} ([|\uparrow\rangle|0\rangle_{c_1} + |\downarrow\rangle|1\rangle_{c_1}) \otimes (\alpha|0\rangle_{b_1} + \beta|1\rangle_{b_1}) \\
\otimes \frac{1}{\sqrt{2}} ([1\rangle_{c_2}|0\rangle_{b_2} + |0\rangle_{c_2}|1\rangle_{b_2}) .
$$

Note that this evolution is purely hypothetical, the state on the right hand side is nowhere near the state predicted by the semiclassical evolution. But with this evolution, the quantum $b_1$ carries the full information about the initial state, so the information comes out. But there is a second quantum $b_2$ which is entangled with $c_2$ so that the Hawking radiation has entanglement entropy $\ln 2$. So if the black hole evaporates away, the final state of radiation will be a mixed state, implying loss of unitarity.

As a second example, consider again the initial matter state $|\psi\rangle_M = \alpha|\uparrow\rangle + \beta|\downarrow\rangle$ and let it evolve as

$$
|\psi\rangle_M = \alpha|\uparrow\rangle + \beta|\downarrow\rangle \to (\alpha|\uparrow\rangle|0\rangle_{c} + \beta|\downarrow\rangle|1\rangle_{c}) \otimes \frac{1}{\sqrt{2}} ([1\rangle_{b} + |0\rangle_{b}) .
$$

This time the state outside is a pure state with no entanglement with the state inside the black hole. But this state carries no information about the initial matter state, so if the black hole disappears we will be left with a pure state and yet lose information.

When a piece of coal is burnt one has normal quantum mechanical evolution, so the radiation is in an unmixed state and also has the information of the coal. In black hole evaporation, the leading order state (4.76) has both the problems of the examples above. The radiation state is entangled with the state in the black hole interior, and also the radiation has only an infinitesimal amount of information about the matter $|\psi\rangle_M$, which arises from the small corrections of order $\epsilon$. It is natural that to expect that a solution to the information paradox will resolve both problems at the same time. But it is useful to keep in mind the above two examples when
discussing the information paradox because the terms 'information loss' and 'mixed state' are used without distinction.

### 4.4.2 The true physical situation

The actual form of the state of the created pairs is thermal \([81, 85]\)

\[
|\Psi\rangle = \prod_i \left( C_i \sum_{n_i} e^{-\frac{\pi n_i \omega_i}{\kappa}} |n_i\rangle_{c_i} |n_i\rangle_{b_i} \right). 
\]

Notice the strong resemblance to the state (4.50), which was found in the derivation of the Unruh density matrix in section 4.2. If one wishes to take into account the fact that the surface gravity of the black hole is slowly changing during the evaporation, one can let \(\kappa\) be a slowly varying function of the index \(i\).

So just like in the toy model, there is a strong correlation between the created quanta inside and outside the black hole. This implies that the Hawking quanta on the outside are described by a density matrix representing a mixed state. So again, there exists no S-matrix connecting the initial, pure state of the matter that collapsed to form the black hole and the mixed state of the Hawking radiation after the black hole evaporation.

Another way to see this more clearly is to look at figure 4.7 where again the evolution of a nice space-like slice \(\Sigma\), or in other words, a foliation with a complete family of Cauchy surfaces, is depicted. The middle slice goes through the endpoint of the evaporation process and is divided in a piece \(\Sigma_{bh}\) on the inside of the black hole, and a piece \(\Sigma_{out}\) on the outside. As the
original derivation of the Hawking radiation and the resulting state (4.81) tells us, there are correlations between the state on $\Sigma_{\text{out}}$ and the state on $\Sigma_{\text{bh}}$. It is also clear from the figure that $I^+(\Sigma_{\text{out}}) = I^+$. So it follows that all times after the endpoint of the evaporation process, only the state on $\Sigma_{\text{out}}$ remains, which cannot be described by a pure state. The result is a thermal density matrix as follows from Hawking’s calculations [81].

Another argument which suggest that something is wrong with the evaporation process is that the total entropy contained in the Hawking radiation is calculated to be some 30% bigger than the original entropy of the black hole [86]. So the fact that the thermal radiation has more entropy than the black hole indicates that the evaporation is non-unitary.

4.5 Implications of non-unitary evolution

In the previous section, we saw that in the semiclassical approach unitary black hole evaporation is far from evident. One can now ask the following question: starting from a pure state of collapsing matter, is the final state of black hole evaporation a mixed state, even when the gravitational field of the black hole has been treated as a part of the quantum mechanical process? In other words, can a microscopic theory of gravity be constructed within the conventional framework of quantum mechanics? Originally, Hawking argued that this cannot be done, and he proposed a modified set of axioms for quantum field theory to accomodate quantum gravity [87].

The connection between systems in background gravitational fields and systems at finite temperature makes it actually intuitively quite reasonable that pure states might evolve into mixed states in quantum gravity. But if 'real' black holes can form and evaporate in quantum gravity, one might expect that ‘virtual’ black holes should have a nonzero amplitude to mediate processes in which a pure state evolves to a mixed state. In that case, the effective, 'macroscopic' (compared to the Planck scale), local dynamical laws for a quantum field might well yield a nonzero probability for evolution from pure to mixed states.

In this section, the effects of such violations of quantum mechanics on ordinary quantum field theory are analyzed. This will be done by following the arguments of Banks, Peskin and Susskind in [88]. It will appear that non-unitarity results in alarming pathological behaviour.

4.5.1 The superscattering operator

We will study the evolution equation for the quantum mechanical density matrix

$$\rho_{\text{out}} = \mathcal{S} \cdot \rho_{\text{in}},$$

(4.82)

where $\mathcal{S}$ is a linear operator which preserves the hermiticity, positivity and normalization

$$\text{tr} \rho = 1$$

(4.83)
of the density matrix. The operator $\mathcal{S}$ is called the superscattering operator and was first introduced by Hawking [81]. In normal quantum mechanics, it is derivable from the scattering matrix $S$ via the relation

$$\mathcal{S} \cdot \rho = S \rho S^\dagger.$$  \hfill (4.84)

The factorization on the right hand side is justified by the completeness of the asymptotic states at future infinity. It is this argument that was rejected by Hawking. Instead, he considered (4.82), supplemented with the requirement of overall energy-momentum conservation, as the basis of quantum dynamics. Thus, he considered a structure in which the usual quantum mechanical connection between $\rho$ and the results of measurements is retained, but where there exists no pure state limit in which $\rho$ represents the evolution of a single wave function.

As noted in [89], one could argue against the non-unitary evolution of (4.82) on the grounds that it is not CPT invariant, since it takes pure states to mixed states, but it does not give the CPT-reversed process of mixed states going to pure states. However, it would be enough to have CPT in the weak form of CPT-invariant transition probabilities

$$p(c \rightarrow a) = S_{ac}^{a} = p(\Theta a \rightarrow \Theta c),$$  \hfill (4.85)

between an initial pure state $c$ and a final pure state $a$ (note there is no sum over repeated indices here), by using (4.82) as an intermediate tool but not interpreting the final density matrix given there as literally the actual final state of the system. Hawking argued in [87] that one should interpret (4.82) as merely an intermediate tool for calculating conditional probabilities: given a measurement of a particular initial pure state, what is the conditional probability of measuring a particular final pure state? In this case the asymmetry may indeed be more in the conditional nature of the probability than in any time asymmetry. This viewpoint refutes the idea of density matrices as being the more basic objects, and probabilities as being derived from them, and puts it the other way around. It will be shown below that the problems with (4.82) are of other nature.

### 4.5.2 A general evolution equation

If the dynamics which give rise to $\mathcal{S}$ are local in time, one can write the infinitesimal version of (4.82) as

$$\frac{d}{dt} \rho = \hat{H} \cdot \rho.$$  \hfill (4.86)

In this equation, $\hat{H}$ represents an arbitrary linear operator, constrained to preserve the hermiticity, positivity and normalization of $\rho$, just like $\mathcal{S}$. The strategy will be to write a convenient canonical form for $\hat{H}$ and then use it to study the properties of (4.86).

Before continuing, a few remarks on the approach here are given. Quantum mechanics is a well-tested theory only on time scales long compared to the Planck time and in regions of spacetime which are, on average, nearly flat. One needs only assume then, that (4.86) can be derived from (4.82) in such a situation by performing a coarse-grained averaging over fluctuations of spacetime. We thus will not worry about possible effects nonlocal in time over a few million Planck times. Equation (4.86) contains the possibility of describing effects nonlocal
in space. Such effects will not be considered unless the nonlocality is of nuclear, rather than Planck, size.

Now, let us try to simplify (4.86). First consider the case of a finite-dimensional Hilbert space. Rewrite (4.86) with indices as

\[
\dot{\rho}_{ab} = \mathbb{H}_{bc}^{a} \rho_{dc}^{c}.
\]  

(4.87)

For fixed values of \(b\) and \(d\), the matrix \(\mathbb{H}_{a}^{b} c\) can be expanded in terms of a complete orthogonal set of hermitian matrices \(Q^{a}\), with \(Q^{0}\) the identity matrix. The expansion coefficients \(\mathbb{H}_{a}^{b} d\) which are, in general, complex, may now also be expanded in terms of the \(Q^{a}\). This allows one to write (4.87) in the form

\[
\dot{\rho} = \sum_{\alpha\beta} h_{\alpha\beta} Q^{\alpha} \rho Q^{\beta}.
\]  

(4.88)

Hermiticity of \(\dot{\rho}\), given the hermiticity of \(\rho\) and the \(Q^{a}\), requires that \(h_{\alpha\beta}\) is a hermitian matrix.

The condition that the normalization is preserved gives

\[
\text{tr} \dot{\rho} = 0 = \text{tr} \left( h_{00} \rho + \sum_{\alpha \neq 0} (h_{0\alpha} + h_{\alpha 0}) Q^{\alpha} \rho + \sum_{\alpha, \beta \neq 0} h_{\alpha\beta} Q^{\beta} \rho Q^{\alpha} \right).
\]  

(4.89)

Because the matrices \(Q^{\alpha}\) and \(\rho\) are assumed to be known, this expression allows us to determine

\[
- \sum_{\alpha \neq 0} h_{0\alpha} Q^{\alpha} \rho = h_{00} \rho + \sum_{\alpha \neq 0} h_{\alpha 0} Q^{\alpha} \rho + \sum_{\alpha, \beta \neq 0} h_{\alpha\beta} Q^{\beta} \rho Q^{\alpha}.
\]  

(4.90)

And similarly, using the cyclic invariance of the trace

\[
- \sum_{\alpha \neq 0} h_{\alpha 0} \rho Q^{\alpha} = h_{00} \rho + \sum_{\alpha \neq 0} h_{0\alpha} \rho Q^{\alpha} + \sum_{\alpha, \beta \neq 0} h_{\alpha\beta} \rho Q^{\beta} Q^{\alpha}.
\]  

(4.91)

Now write (4.88) as

\[
\dot{\rho} = h_{00} \rho + \sum_{\alpha \neq 0} h_{0\alpha} \rho Q^{\alpha} + h_{\alpha 0} \rho Q^{\alpha} + \sum_{\alpha, \beta \neq 0} h_{\alpha\beta} Q^{\beta} \rho Q^{\alpha}
\]  

\[
= - \frac{1}{2} \left( h_{00} \rho + \sum_{\alpha \neq 0} h_{\alpha 0} Q^{\alpha} + \sum_{\alpha, \beta \neq 0} h_{\alpha\beta} Q^{\beta} Q^{\alpha} \right)
\]  

\[+ \frac{1}{2} \left( h_{00} \rho + \sum_{\alpha \neq 0} h_{\alpha 0} \rho Q^{\alpha} + \sum_{\alpha, \beta \neq 0} h_{\alpha\beta} \rho Q^{\beta} Q^{\alpha} \right)
\]  

\[+ \frac{1}{2} \sum_{\alpha \neq 0} h_{\alpha 0} Q^{\alpha} \rho + \frac{1}{2} \sum_{\alpha \neq 0} h_{0\alpha} \rho Q^{\alpha}
\]  

\[- \frac{1}{2} \sum_{\alpha, \beta \neq 0} h_{\alpha\beta} Q^{\beta} Q^{\alpha} - \frac{1}{2} \sum_{\alpha, \beta \neq 0} h_{\alpha\beta} \rho Q^{\beta} Q^{\alpha}
\]  

\[+ \sum_{\alpha, \beta \neq 0} h_{\alpha\beta} Q^{\beta} \rho Q^{\alpha}.
\]  

(4.92)
Using (4.90) and (4.91), this becomes
\[
\dot{\rho} = -\frac{1}{2} \sum_{\alpha \neq 0} h_{0\alpha} Q^\alpha \rho - \frac{1}{2} \sum_{\alpha \neq 0} h_{\alpha 0} \rho Q^\alpha \\
+ \frac{1}{2} \sum_{\alpha \neq 0} h_{\alpha 0} Q^\alpha \rho + \frac{1}{2} \sum_{\alpha \neq 0} h_{0\alpha} \rho Q^\alpha \\
- \frac{1}{2} \sum_{\alpha, \beta \neq 0} h_{\alpha\beta} Q^\beta \rho Q^\alpha - \frac{1}{2} \sum_{\alpha, \beta \neq 0} h_{\alpha\beta} \rho Q^\beta Q^\alpha \\
+ \sum_{\alpha, \beta \neq 0} h_{\alpha\beta} Q^\beta \rho Q^\alpha.
\] (4.93)

Introduce the operator
\[
H_0 = \sum_{\alpha \neq 0} (h_{0\alpha} - h_{\alpha 0}) Q^\alpha = 2iH_0,
\] (4.94)

which is hermitian because $Q^\alpha$ is hermitian and $h_{\alpha 0}^* = h_{0\alpha}$. With $H_0$, equation (4.93) takes the form
\[
\dot{\rho} = -i [H_0, \rho] - \frac{1}{2} \sum_{\alpha, \beta \neq 0} h_{\alpha\beta} \left( Q^\beta Q^\alpha \rho + \rho Q^\beta Q^\alpha - 2Q^\alpha \rho Q^\beta \right).
\] (4.95)

The right hand side is now explicitly traceless. Equation (4.95) is called the Lindblad equation [90].

One still needs to implement the requirement that $\rho$ remains positive. This is the case if $h_{\alpha\beta}$ is a positive matrix. To see this, diagonalize $h_{\alpha\beta}$ with the unitary matrix $U$
\[
D = U^\dagger HU,
\] (4.96)

where $D$ is a diagonal matrix. This implies
\[
H = UDU^\dagger.
\] (4.97)

So one gets
\[
h_{\alpha\beta} Q^\alpha Q^\beta = u_{\lambda\alpha} u_{\lambda\beta} Q^\alpha Q^\beta
\]
\[
= h_{\lambda} (u_{\lambda\alpha} Q^\beta) (u_{\lambda\beta} Q^\beta)
\]
\[
= h_{\lambda} Q^\lambda Q^{\lambda},
\] (4.98)

where summation over $\alpha$, $\beta$ and $\lambda$ is understood. The $Q^\lambda$ are not necessarily hermitian, but are orthogonal in the sense that
\[
\text{tr} Q^\lambda Q^{\mu} = u_{\lambda\alpha} u_{\mu\beta} \text{tr} (Q^\alpha Q^\beta)
\]
\[
= u_{\lambda\alpha} u_{\mu\beta} \delta_{\alpha\beta}
\]
\[
= \delta^\mu\nu,
\] (4.99)
Because the $Q^\alpha$ were taken to be orthogonal and $U$ is unitary. Now diagonalize $\rho$, calling its eigenvalues $p_i$, and consider the situation in which one eigenvalue, say $p_1$, becomes zero. Then
\[
\frac{d}{dt} p_1 = \dot{\rho}_{11} |_{p_1=0} = \sum_\lambda h_\lambda |Q_\lambda^1|^2 p_1 ,
\]
so that $\rho$ remains positive if $h_\lambda \geq 0$. It should be noted that the condition that $h$ is positive is a sufficient condition for $\rho$ to remain positive during the evolution, but examples can be found that this is not strictly necessary.

Thus, it is shown that a linear evolution equation for $\rho$ can generally be written in the form (4.95). Assuming that $h$ is positive ensures that $\rho$ remains positive.

### 4.5.3 A subclass of solutions

The case in which $h$ in (4.95) is real and positive possesses a simple physical interpretation. Here, this interpretation is presented and used to expose problems with writing (4.95) as the fundamental equation.

Consider a system described by quantum mechanics evolving under the action of the following Hamiltonian
\[
H(t) = H_0 + \sum_\alpha j_\alpha(t) Q^\alpha ,
\]
where the $Q^\alpha$ are a set of hermitian operators and the source terms $j_\alpha(t)$ are complex numbers. Let the $j_\alpha$ vary randomly in time, according to a Gaussian distribution with covariance
\[
\langle j_\alpha(t) j_\beta(t') \rangle = h_{\alpha\beta} \delta(t - t') .
\]
In (4.102), $h_{\alpha\beta}$ is real, symmetric and positive.

In ordinary quantum mechanics, the evolution of the density matrix is determined by the Liouville-Von Neumann equation
\[
\frac{\partial \rho}{\partial t} = -i[H(t), \rho] .
\]
Integrating both sides from 0 to $t$ gives
\[
\rho(t) = \rho(0) - i \int_0^t dt' [H(t'), \rho(t')] .
\]
This equation can be solved recursively, leading to the series
\[
\rho(t) = \rho(0) - i \int_0^t dt' [H(t'), \rho(0)] + (-i)^2 \int_0^t dt' \int_0^{t'} dt'' [H(t'), [H(t''), \rho(0)]] + ...
\]
So if $\rho(0)$ is the density matrix at time $t = 0$ of the system with the 'random noise'-Hamiltonian (4.101), the density matrix after a small time $t = \epsilon$ is given by

$$\rho(\epsilon) = \rho(0) + i \int_0^\epsilon dt' [H_0 + j_\alpha(t')Q^\alpha, \rho(0)] - \int_0^\epsilon dt' \int_0^{t'} dt'' [H_0 + j_\alpha(t')Q^\alpha, [H_0 + j_\beta(t'')Q^\beta, \rho(0)]] + ... \quad (4.106)$$

Where summation over $\alpha$ and $\beta$ is understood. Averaging over the $j_\alpha(t)$ and using (4.102) together with $\langle j_\alpha(t) \rangle = 0$, one finds

$$\rho(\epsilon) = \rho(0) - i \epsilon \left[ H_0, \rho(0) \right] - \frac{1}{2} \epsilon h_{\alpha\beta} [Q^\alpha, [Q^\beta, \rho(0)]] - \frac{1}{4} \epsilon^2 [H_0, [H_0, \rho(0)]] \quad (4.107)$$

So, working up to first order in $\epsilon$, this gives

$$\rho(\epsilon) - \rho(0) = -i \epsilon [H_0, \rho(0)] - \frac{1}{2} \epsilon h_{\alpha\beta} [Q^\alpha, [Q^\beta, \rho(0)]] + \mathcal{O}(\epsilon^2),$$

$$= -i \epsilon [H_0, \rho(0)] - \frac{1}{2} \epsilon h_{\alpha\beta} (Q^\alpha Q^\beta \rho(0) - Q^\alpha \rho(0) Q^\beta - Q^\beta \rho(0) Q^\alpha + \rho(0) Q^\beta Q^\alpha) + \mathcal{O}(\epsilon^2), \quad (4.108)$$

which is equal to (4.95) since $h$ is symmetric. Thus, in this special case, (4.95) is simply equivalent to ordinary quantum mechanics in the presence of a random source term.

However, quantum mechanics with a random source differs from the observed behavior of elementary particles in two important respects. First, energy is not conserved. In each realization of the random source, the nontrivial time dependence of the source allows energy to be added or removed. Secondly, in the case of a field theory, there is an irreconcilable conflict between momentum conservation and locality. For a field theory, (4.102) must be generalized to

$$H(t) = H_0 + \int d^3 x j_\alpha(t, x)Q^\alpha(x). \quad (4.109)$$

If the sources $j_\alpha(x)$ fluctuate randomly as a function of spatial position, then, in each given realization, the sources will break translational invariance and add momentum to the system. On the other hand, if the fluctuations of the sources are translationally invariant, the sources must go through the same random fluctuations at widely separated points on the same space-like surface. This will introduce correlations between fields at space-like separated points. So locality is violated. In general, the range of the spatial correlations of $\langle j_\alpha(x)j_\beta(y) \rangle$ will be just the reciprocal of the size of the typical momenta added or subtracted.

The violation of energy conservation and the conflict between locality and momentum conservation observed here for a particular class of solutions can be shown to also hold for the
general evolution equation (4.95) [88].

The failure of energy conservation can be seen from the following observation. What if the theory did possess some hermitian operator $H$, not necessarily equal to $H_0$, which was conserved by the dynamics? Then any $\rho$ which was a function only of $H$ could not change under the action of (4.95). However, this is possible only if (4.95) contains only operators $Q$ which are simultaneously diagonalizable with $H$. Unless $H$ has highly degenerate eigenvalues, a property which would exclude it as a good candidate for the energy, this is a serious restriction on $h_{\alpha\beta}$, especially if $Q_\alpha$ must be a local operator rather than a global charge.

From the arguments in this section it is clear that unitarity is not something one simply ’gives up’. The consequences on effective field theory would be dramatic, leading to a non-conservation of energy and a non-compatibility of momentum conservation and locality. It would lead to a major rethinking on some of the most profound principles of physics. For this reason, we will examine the possibilities to retain unitarity in the evaporation of black holes.

### 4.6 Possible ways to unitarity

In the previous section it was shown that the conclusion, based upon the loss of information in the semiclassical approach to black hole evaporation, that quantum mechanics should be altered to describe non-unitary processes seems to open Pandora’s box. In this section, it will be investigated what the possibilities are to preserve unitarity and what the implications would be. The arguments are taken from [82, 89, 91].

#### 4.6.1 Information in the Hawking radiation

In this section we’ll be stubborn and assume that a black hole, even with all the arguments of section 4.6, is nevertheless capable of returning all the information about the collapsed matter in the Hawking radiation. We will present two possible scenarios for the Hawking radiation to contain the desired information.

##### 4.6.1.1 Backreaction and small corrections

The first scenario states that the Hawking radiation contains subtle correlations which make it not exactly thermal. The thermality as found in the original derivation is seen as a ’leading order’ result, not capturing enough of the physics to provide a unitary description of black hole formation and evaporation. If this model were true, then the information paradox could be solved entirely within the semiclassical approach.

The reason to assume that the thermality of the Hawking radiation might not be exact is that its derivation does not take into account backreaction effects. Backreaction is the influence of the evaporation process on the metric, or in other words, the shrinking of the black hole due to energy conservation. How beautiful the Hawking derivation might be, it actually does
Chapter 4. Entanglement and information

not contain energy conservation. This has to be imposed as an extra condition. Therefore, the original derivation of the Hawking radiation in section 2.3.1 could be seen as incomplete. So a natural question is whether this backreaction effect could resolve the information paradox.

An approach to black hole radiation that successfully takes into account the effect of backreaction was put forth in [92, 93]. It is based on a viewpoint about the radiation process that was already mentioned in section 2.3.2.2, namely that one can think of if as a tunneling process. The black hole particle creation process happens in complete analogy to the Schwinger process [94]. There, a virtual particle pair becomes real through energy extraction out of an electric field. The Schwinger also possesses following alternative interpretation. A particle moves backwards in Lorentzian time and is slowed down by the field until it becomes momentarily at rest. At that point one can make the transition to imaginary time so that the particle undergoes a tunneling process where it extracts an energy $2m$, with $m$ the particle’s mass, from the field. After that, the particle’s orbit re-enters real time and it continues its existence according to the conventional picture of a particle being accelerated by the field. The tunneling approach incorporates the idea of energy conservation rather naturally since the tunneling process takes place between states with the same energy.

The tunneling process is described in the WKB approximation [95]. In this approximation, the ansatz $\psi = e^{iW}$ for the time-independent Schrödinger equation at a fixed energy yields the Hamilton-Jacobi equation for $W$. This identifies $W$ as the Jacobi action

$$W = \int (L + H)dt,$$

where $L$ and $H$ are the Lagrangian and the Hamiltonian, respectively. $W$ is minimized by the classical orbit. For configurations separated by a barrier, but connectable by a classical tunneling orbit in imaginary time, $W$ becomes complex and gives a tunneling probability

$$P = \psi\psi^* = e^{-2imW}$$

in lowest order approximation. This expression has a rather general validity and is readily applicable in situations where a tunneling path can be clearly defined, as is the case for a particle. For a field, a reduction procedure is needed to describe the tunneling of a one-particle excitation. This is done in [93].

Now consider the following view on black hole radiation. A virtual particle, which will be taken to be an s-wave since they are by far the most dominant in the Hawking radiation, is represented by a spherical shell and formed just beneath the future horizon $H^+$. The outer surface of the shell is denote by $\Sigma_+$ and the inner surface by $\Sigma_-$. It then tunnels through the horizon and escapes to infinity. The loss of mass causes the horizon to shrink from $H^+$ to $H^-$ (note that $H^-$ does not represent the past event horizon here). For an optimum tunneling probability, the outward velocity before the tunneling should be as large as possible and the velocity after the tunneling should be as small as possible. Because time and space switch roles at the horizon, this requires that both pre- and post-tunneling histories should be very nearly surfaces of constant retarded time $v = t + r^*$. Tunneling thus takes place between two given
values of $u$, i.e. $u_+$ and $u_-$, where

$$u_\pm = -\frac{1}{\kappa_\pm} \ln U_\pm ,$$  \hspace{1cm} (4.112)$$

with $U$ the usual Kruskal-Szekeres coordinate introduced in section 1.5, and '+' denotes 'before the passage of $\Sigma_+' and '-' denotes 'after the passage of $\Sigma_-'.

The horizon $H^+$ is given by $U_+ = 0$ and the horizon $H^-$ is given by $U_- = 0$. The precise difference in the magnitude of $U_+$ and $U_-$ is not so relevant, the only thing which matters is that the particle goes from a negative $U$-value to a positive one. Just as in the Schwinger process, where time is made imaginary so as to create a route in the complex plane 'around' the $t$-axis, we will make $U$ imaginary so that the particle has a route in the complex $U$-plane 'around' the horizon at $U = 0$. So making $U$ imaginary during the tunneling and using the general identity of complex logarithms

$$\ln x = \ln|x| + \text{sgn}(x)i\frac{\pi}{2} ,$$  \hspace{1cm} (4.113)$$

leads, together with (4.112), to the conclusion that the imaginary part of $u$ jumps from $\pi/2\kappa_+$ before the horizon shrinks to $-\pi/2\kappa_-$ afterwards.

It is shown in [95], based on so called 'shell dynamics', that the appropriate Jacobi action for the tunneling process is

$$W = -\int Md\tau + \int dE \int dt(E) ,$$  \hspace{1cm} (4.114)$$

where $M$ is the mass of the black hole, $\tau$ is the proper time of the particle, $t$ is the Schwarzschild time and $E$ is the conjugate variable, i.e. the energy measured at infinity. In the same paper, it is also shown that the expression for the Jacobi action does not change when the Schwarzschild time is subjected to arbitrary space-dependent translations $t \rightarrow t_{\text{gen}} = t + f(r)$. So for the tunneling process described above, the natural choice of parameter is $t_{\text{gen}} = u$. Using that the imaginary part of $u$ jumps from $\pi/2\kappa_+$ to $-\pi/2\kappa_-$, the imaginary part of the Jacobi action (4.114) becomes

$$\text{Im}W = \int dE \frac{\pi}{\kappa(E)} ,$$  \hspace{1cm} (4.115)$$

where $\kappa(E)$ is the surface gravity after a mass $E$ has been lost by the hole. If we now assume that the evaporation is quasi-stationary, which is almost exactly true for large black holes, then it follows from the first law of black hole mechanics of section 1.11.3 that

$$-\frac{2\pi}{\kappa(E)}dE = \frac{1}{4G}dA = dS .$$  \hspace{1cm} (4.116)$$

So we are lead to the very important result for the tunneling amplitude

$$P = e^{\Delta S} .$$  \hspace{1cm} (4.117)$$

Alternatively and along the same lines, Hawking radiation can also be regarded as a pair creation outside the horizon, with the negative energy particle tunneling into the black hole. Since
such a particle propagates backwards in time, one has to reverse time in the equations of motion. Both channels, particle or anti-particle tunneling, contribute to the rate for the Hawking process. So, in a more detailed calculation, one would have to add their amplitudes before squaring in order to obtain the semiclassical tunneling rate. Such considerations, however, only concern a prefactor. In either treatment, the exponential part of the semiclassical emission rate is the same \[93\].

Now consider a virtual s-wave quantum with energy \(-\omega\), which escapes from a Schwarzschild black hole. Then, the tunneling probability becomes

\[
P \sim e^{4\pi G((M-\omega)^2-M^2)}
\sim e^{-8\pi G\omega(M-\frac{\omega}{2})}
\sim e^{-2\pi\omega(M-\frac{\omega}{2})}.
\] (4.118)

So the important conclusion is that energy conservation makes the radiation spectrum not exactly thermal, there is an order \(O(\omega^2)\) correction to the usual Boltzmann factor. (It should be noted that the proof for the generalized second law of black hole thermodynamics in section 4.3 used the thermality of the Hawking radiation. However, the proof can be modified to show that small corrections do not invalidate the second law \[80\].)

In \[96\], it was argued that these non-thermal deviations have the power of carrying away all the information about the black hole’s initial state. One defines the correlation coefficient \(C(a, b)\) between two events \(a\) and \(b\) by

\[
C(a, b) = \ln \left( \frac{P(a,b)}{P(a)P(b)} \right),
\] (4.119)

where \(P(a,b)\) is the probability of both \(a\) and \(b\), and \(P(b) = \sum_a P(a,b)\) the probability of \(b\). The conditional probability of \(b\) is

\[
P(b|a) = \frac{P(a,b)}{P(a)}.
\] (4.120)

With (4.118), one then gets a non-trivial correlation between the emission of two radiation quanta \(\omega_1\) and \(\omega_2\)

\[
C(\omega_1, \omega_2) = \ln \left( \frac{P(\omega_1, \omega_2)}{P(\omega_1)P(\omega_2)} \right) = 8\pi\omega_1\omega_2.
\] (4.121)

It it important to note that one should not replace \(M\) by \(M-\omega_1\) in \(P(\omega_2)\) to take into account the loss of mass from the first emission. This would come down to replacing \(P(b)\) by \(P(b|a)\) in (4.119). But one sees from (4.120) that this would give a correlation identically zero between any two events. The argument is circular: it absorbs the correlations themselves into the test for their existence.

The conditional probability

\[
P(E_i|E_f) = e^{-8\pi GE_i(M-E_f-E_i^2)}
\] (4.122)
corresponds to the tunneling probability of a particle with energy $E_i$, conditional on a total energy $E_f$ having left the black hole. The entropy taken away by the tunneling particle with energy $E_i$ is then given by

$$S(E_i|E_f) = -\ln P(E_i|E_f).$$

(4.123)

In quantum information theory, $S(E_i|E_f)$ denotes the conditional entropy. Quantitatively, it is equal to the decrease of the entropy of a black hole with mass $M - E_f$ upon the emission of a particle with energy $E_i$. Such a result is consistent with the second law of black hole thermodynamics: the emitted particles must carry entropy in order to balance the total entropy of the black hole and the radiation.

We now count the entropy carried away by the Hawking radiation. The entropy of the first emission with an energy $E_1$ from a black hole of mass $M$ is

$$S(E_1) = -\ln P(E_1).$$

(4.124)

the conditional entropy of a second emission with an energy $E_2$ after the $E_1$ emission is

$$S(E_2|E_1) = -\ln P(E_2|E_1).$$

(4.125)

The total entropy for the two emissions then becomes

$$S(E_1, E_2) = S(E_1) + S(E_2|E_1),$$

(4.126)

and the mass of the black hole reduces to $M - E_1 - E_2$ while it proceeds with the emission of a third particle with energy $E_3$ and entropy $S(E_3|E_1, E_2)$. Continuing this reasoning, the total entropy emitted by the Hawking radiation is

$$S(E_1, E_2, ..., E_n) = \sum_{i=1}^{n} S(E_i|E_1, E_2, ..., E_{i-1}),$$

(4.127)

with $M = \sum_{i=1}^{n} E_i$ due to energy conservation. Since the conditional probability can be rewritten as

$$\ln P(E_i|E_1, E_2, ..., E_{i-1}) = 8\pi G E_i \left( M - \sum_{j=1}^{i-1} E_j - \frac{E_i}{2} \right) = -4\pi G \left( \left( M - \sum_{j=1}^{i} E_j \right)^2 - \left( M - \sum_{j=1}^{i-1} E_j \right)^2 \right) = -\Delta S,$$

(4.128)

it readily follows that

$$S(E_1, E_2, ..., E_n) = 4\pi G M^2 = S_{\text{black hole}}.$$  

(4.129)

So the entropy carried away by the Hawking radiation is now equal to the initial black hole entropy. As mentioned before, when the Hawking radiation is exactly thermal, its entropy is some 30% bigger than the black hole entropy. So this definitely indicates an improvement towards solving the information paradox. To recapitulate, backreaction effects cause a deviation.
of thermality in the emission spectrum which is shown to contain correlations that have the
capacity to carry off the maximum information content of the hole. This viewpoint leads to a
possible interpretation of black hole entropy as the uncertainty about the information of the
black hole forming matter precollapsed configurations [97].

In a very recent paper [98], the collapsing shell in the semiclassical approach was also treated
quantum mechanically. This produces small off-diagonal components in the density matrix of
the Hawking radiation with magnitude of order $S^{-1/2}$. These off-diagonal elements seem to
store the correlations between the collapsing shell and the emitted radiation and allow informa-
tion to continuously leak from the collapsed body. These results again favor the idea that small
corrections restore unitary evolution.

4.6.1.2 Quantum hair and fuzzballs

The reasoning behind the information paradox fails if a black hole did not have an 'information-
free' horizon as mentioned in section 4.6.1.4. But in section 1.8, it was argued that a classical
black hole has no hair, implying that it does not posses any degrees of freedom to store the
information about the collapsed matter such that it is available to an outside observer. So the
only possibility for a black hole to have information about the collapsed state at its horizon is
that is has 'quantum hair'. With quantum hair, black holes do burn up like an ordinary piece
of coal, releasing its information during the evaporation process. There are claims based on
fuzzball models, black hole models from string theory, that this required quantum hair is found
[99, 100]. If these models were correct, then it would imply that it is impossible to resolve the
information paradox within the semiclassical approach since the mechanism needed to create
the correlations lies at the Planck scale, in string theory.

It should be noted that although the two models described above provide an acceptable res-
olution of the information paradox, it is not yet settled. The resolutions from both models
obviously have very different origins, and people are still debating on which one holds the true
key to resolving the paradox. Basically, the debate is centered around the question whether
or not small correlations to the leading order state (4.81) are sufficient to make the final state
of the radiation pure. Proponents of the fuzzball model claim that small correlations do not
change the basic conclusion of the outline in section 4.6, while proponents of backreaction state
they do [82, 83, 98, 99].

(The shorter treatment of the fuzzball model is due to the incompetence of the author on
string theory, it should not be regarded as a biased point of view towards the resolution of the
information paradox.)

4.6.2 Stable remnants

Perhaps quantum gravity effects halt the evaporation process, so that a stable black hole rem-
nant is left behind. At first sight, this seems to resolve the information paradox because all
of the information about the initial collapsing object can in principle reside inside the remnant. It should be noted that remnants are not ruled out by CPT invariance as Hawking once claimed. He said that because black holes can form when there was no black hole present beforehand, CPT implies that they must also be able to evaporate completely. However, the only requirement from CPT is that a CPT-reversed remnant should be able to combine with a CPT-reversed Hawking radiation to form a large CPT-reversed black hole, i.e. a white hole, which can convert into the CPT-reversed of whatever collapsed to form the black hole. If there is no CPT-reversed Hawking radiation impinging on the CPT-reversed remnant, it can be absolutely stable and yet be consistent with CPT-invariance.

But upon further reflection on the stable remnant solution to the information paradox, the cure may appear worse than the disease. Since the initial black hole could have been arbitrarily massive, the remnant must be capable of carrying an arbitrarily large amount of information, about $M^2/M_p^2$ bits, if the initial mass was $M$. This means that there must be an infinite number of species of stable remnants, all with mass comparable to $M_p$.

It seems hard to reconcile this sort of infinite degeneracy with the fundamentals of quantum field theory, that is, with causality and unitarity [101]. The coupling of the remnants to hard quanta might be surpressed by form factors, but the coupling to soft quanta, i.e. wavelength $\gg l_p$, should be well-described by an effective field theory in which the remnant is reagarded as a pointlike object. Then the coupling to soft gravitons, say, should be determined only by the mass of the remnant, and should be independent of its internal structure, including its information content. It should be possible to use this effective field theory to analyze, for example, the emission of Planck-size remnants in the evaporation of a large black hole. For each species, the emission is suppressed by a tiny Boltzman factor $\exp(-\beta_{\text{Hawking}} M_{\text{remnant}})$. But if there are an infinite number of species, the luminosity is nonetheless infinite.

The emission of Planck-size remnants in the evaporation of a large black hole is merely an example of a soft process in which heavy particles can be produced, a process that is expected to admit an effective field theory description. If such processes really have infinite rates, as would be expected if there are an infinite number of Planck-mass species, then these infinities will inevitably infect other calculated processes, as a consequence of unitarity. These infinities would destroy the consistency of the theory. So if stable remnants really are the answer, an effective field theory description of the coupling of the remnants to soft quanta cannot be valid. The coupling must depend on the hidden information content of the remnant.

A suggested variation on the stable remnant idea is that a black hole which harbors a lot of information actually stops evaporating when it is still large compared to the Planck length $l_p$ [102]. The more information, the larger the remnant. So the number of species less than a specified mass $M$ is always finite, and the contributions of remnants to soft processes can be heavily suppressed. But the odd thing about this idea is that there must be arbitrarily large black holes that emit no Hawking radiation, contrary to the semiclassical theory. This failure of the semiclassical theory must occur even though the curvature at the horizon is arbitrarily small.

Another displeasing feature of the remnant idea is that it leaves us without a reasonable interpretation for the black hole entropy. If information is really encoded in the Hawking radiation,
then it seems to make sense to say that $e^{S(M)}$ counts the number of accessible black hole internal states for a black hole of mass $M$. But if the information stays inside the black hole, then the number of internal states has nothing to do with the mass of the black hole. Indeed, we can prepare a black hole of mass $M$ that holds for an arbitrarily large amount of information by initially making a much larger hole, and then letting it evaporate for a long time. Thus, the number of possible internal states for a black hole of mass $M$ must really be infinite. The beautiful framework of black hole thermodynamics then seems like an inexplicable accident. If a black hole really destroys information, then the interpretation of the intrinsic entropy must be somewhat different, but perhaps still sensible. The black hole entropy can be seen as the amount of inaccessible information. As the black hole evaporates, the entropy is transferred to the outgoing radiation. The entropy of the radiation does not result from coarse graining, the mixed density matrix characterizing the radiation is really an exact description of its state.

Note that if the idea of stable black hole remnants is rejected, there is a very important consequence: there can be no exact continuous global symmetries in nature. Suppose that $Q$ is a conserved charge, and that $m > 0$ is the mass of the particle with the smallest mass-to-charge ratio and take it’s charge to be one. By assembling $N$ of these particles, one can create a black hole with charge $Q = N$ and mass $M$ of order $Nm$. If $N$ is large enough, one has $M \gg M_{\text{Planck}}$, so that the semiclassical theory can be safely applied to this black hole. In fact, one can make $M$ so large that the Hawking temperature is small compared to the masses of all charged particles. Then the black hole will radiate away most of its mass in the form of light uncharged particles, without radiating away much of its charge. At this point, there is no way for the evaporation process to proceed to completion without violating conservation of $Q$. There is no available decay channel with charge $Q = N$ and a sufficiently small mass. The only way to rescue the conservation law is for the black hole to stop evaporating, and settle down to a stable remnant that carries the conserved charge. And there would be an infinite number of species because $N$ could take any value. If one accepts the objections to the existence of an infinite number of remnant species, then, one must accept the consequence that the conservation law is violated.

This is an unusual kind of anomaly. There is a conservation law that is exact at the quantum level, but is spoiled by classical effects! Note that this argument for nonconservation breaks down if there are massless particles that carry the conserved charge. But it is easy to think of examples where this is not the case, like for baryon number. Since by the no-hair conjecture, the black hole ‘forgets’ the value of the charge that it consumes, one may wonder whether loss of information isn’t unavoidable in theories that suffer from this anomaly, theories in which the conservation law is violated only by processes involving black holes. However, in the next chapter we will resolve this problem in the framework of black hole complementarity.

### 4.6.3 Information release at the end

In section 4.8.1 the situation was considered where after most of the mass of the black hole is radiated away, the state of the radiation that has been emitted is not thermal, but instead is nearly pure. Another logical possibility is that the radiation remains truly thermal until much later, just as the semiclassical theory indicates. Finally, when the black hole evaporates down to the Planck size, and the semiclassical theory breaks down, information starts to leak out: it
is encoded in correlations between the thermal quanta emitted earlier and the quanta emitted 'at the end'.

But if the black hole was initially very big, so that the amount of information is very large, then the information can not come out suddenly. The final stage of the evaporation process must take a very long time [103, 104]. To get an idea of how long it must take, one should count the number of quantum states that are available to the Planck-energy’s worth of radiation that is emitted in the last stage. These quanta all have wavelengths that are much larger than the size of the evaporating object, so it is an excellent approximation to suppose that they all occupy the lowest partial wave. Thus, for the purpose of counting states, the problem reduces to a one-dimensional (radial) ideal gas.

Actually, the same is true to a reasonable approximation for a big black hole, as was shown in section 2.4. It can also be seen intuitively because the emitted quanta have a wavelength comparable to the size of the hole. First, let’s consider the case of a big black hole, and check if the black hole entropy counts the number of radiation states from which the black hole can be assembled. If the mass of the black hole is \( M \), then the radiation state from which it formed must contain energy \( M \) inside a sphere with radius comparable to the Hawking evaporation time, which can be found by using the expression for the black hole radiation luminosity [105]

\[
L = \frac{C}{M^2},
\]

(4.130)

where \( C \) is a positive constant that depends on the number of quantized matter fields that couple to gravity [40]. It now follows from energy conservation that the rate of loss of mass is proportional to the luminosity

\[
\frac{dM}{dt} = -\frac{C}{M^2},
\]

(4.131)

So it follows that

\[
M(t) = (M^3 - 3Ct)^{1/3},
\]

(4.132)

implying that the Hawking evaporation time is \( t_{\text{Hawking}} \approx M^3 \), where units \( \text{M}_{\text{Planck}} = 1 \) are used. The entropy of a one-dimensional ideal gas with energy \( E \) and volume \( L \) is, in order of magnitude,

\[
S \sim \sqrt{EL}.
\]

(4.133)

So for \( E \approx M \) and \( L \approx M^3 \), one finds the usual relation for the black hole entropy \( S \sim M^2 \).

It is interesting to ask how the above analysis is modified if there are \( n \) different species of massless radiation, with \( n \gg 1 \). Then the entropy scales like \( S \sim \sqrt{nEL} \), but the Hawking time decreases like \( L \sim M^3/n \). So we see that \( n \) drops out of the entropy, and one can begin to understand how the black hole entropy can be a universal quantity, independent of the details of the matter Lagrangian.

Now let’s ask what the volume of a one-dimensional ideal gas would have to be, if the gas has the same entropy as above, but energy \( E \approx 1 \). Or in other words, how much would the gas have to expand adiabatically to cool down to \( E \approx 1 \). Evidently, it would need to expand by the factor \( M \), so that \( L \approx M^4 \). If it takes a time \( t_{\text{remnant}} \) before the long-lived remnant finally
disappears, then the radiation emitted during this time occupies a sphere of radius $L \sim t_{\text{remnant}}$.

Thus, one obtains a lower bound

$$t_{\text{remnant}} \geq M^4.$$  \hfill (4.134)

This bound is saturated if the final radiation is equilibrated, that is, if it is able to occupy nearly all of the states that are available in the allotted time. Of course, the decay of the remnant might actually take much longer, but it has to take at least this long.

Another way to say what is going on is that the remnant must emit about $S \sim M^2$ quanta to reinstate the information. Since the total energy is of order one, a typical quantum has energy $M^{-2}$ and wavelength $M^2$. Further, to carry the required information, these quanta must be only weakly correlated with one another. This means, roughly speaking, that they must come out one at a time, as non-overlapping wave packets. Since the time for the emission of each quantum is $M^2$, and there are $M^2$ quanta, the total time is $M^4$.

If the information comes out at the end, then the scenario is that a black hole with initial mass $M$ evaporates down to Planck size in time $M^3$, but the time for the Planck-size remnant to disappear is much longer, at least $M^4$. The trouble is that, since $M$ can be arbitrarily large, there must be Planck-size black hole remnants that are arbitrarily long lived, even if no species is absolutely stable. If there are an infinite number of species with mass of order the Planck mass, all with an enormous lifetime, then one has all the same problems as if the remnants were absolutely stable.

### 4.6.4 Baby universes

It could also be that the disappearance of black holes results in mixed states that are simply unpredictable. This could occur for a $CPT$-invariant model in which our universe is an open system, and information can both leave and enter. An analogue would be a room with a window: from the density matrix of the inside of the room alone at one time, one cannot know what light might come in from the outside, and hence one cannot predict even the density matrix inside the room at a later time. Unlike the case of deterministic evolution of the density matrix by a superscattering matrix, in the case of an open system one generally cannot extrapolate backward from the later density matrix to a unique earlier one, so information would be truly lost in an even more fundamental way.

A concrete mechanism for this open universe-view was offered in [106–109]. The picture is that quantum gravity effects prevent the collapsing body from producing a true singularity inside the black hole. Instead the collapse induces the nucleation of a closed 'baby universe'. This new universe carries away the collapsing matter, and hence all detailed information about its quantum state. The baby universe is causally disconnected from our own, and so completely inaccessible to us, there is no hope of recovering the lost information. Yet there is a larger sense in which information is retained. The proper setting for quantum theory, in this picture, is a 'multiverse' which encompasses the quantum-mechanical interactions of all of the universes that are causally disconnected at the classical level. To the 'superobserver' who is capable of perceiving the state of the whole multiverse, no information is lost. It is merely transferred from
one universe to another. In a more correct quantum-mechanical language, black holes produce correlations between the state of the parent universe and the state of the baby universe, and it is because of these correlations that both the parent and the baby are described as mixed quantum states.

To obtain a CPT-invariant version of the mechanism above, one could postulate that there is an $S$ matrix for the superobserver, from the product Hilbert space of our past universe and the past baby universes, to the product Hilbert space of our future universe and the future baby universes. Now the reasoning above implies that quantum gravity can allow connections to baby universes that can branch off or join on. However, it raises some questions that are not very clear. For example, if the dimension of the Hilbert space of our universe stays the same from past to future, then the two hidden Hilbert spaces should also have the same dimension in order that there can be an $S$ matrix between the two Hilbert spaces, at least the argument would be valid if all these dimensions were finite. That would mean that there would be in principle as many ways for information to enter our universe as to leave it. And yet the semiclassical approximation seems to show many ways for old information to leave our universe, but the only place it seems to allow for new information to enter is at a possible naked singularity at the end of the black hole evaporation, where the semiclassical approximation breaks down. One might even expect quantum gravity to heal the naked singularity so that no new information enters the universe from it, a possibility called the Quantum Cosmic Censorship hypothesis. In other words, the semiclassical approximation suggests that the dimension of the past hidden Hilbert space is small or perhaps even zero. If the dimension of past and future hidden Hilbert spaces are actually equal, as one should expect from the reasoning above, which suggestion is then correct? Taking the large dimension supports the view that pure states go to mixed states, but taking the small dimension suggests that little or no information is lost, and that pure states may stay pure.

On the other hand, it could turn out that even if the dimensions of the two hidden Hilbert spaces are identical and nontrivial, some principle influencing the states on those two spaces might make it so that in actuality more information leaves our universe than enters it. As an analog, take again the room with a window. When it is dark outside, little information in the visual band of photon modes is coming in, whereas there is much more information going out from the light inside. From the inside, one can more easily predict the light one sees reflected in the window, whereas in the daytime, one cannot predict the light entering from the clouds outside that are floating by. So in this language, the question would be, why do past baby universes seem to be so dark? Perhaps the answer is that something like the Linde inflationary proposal [110] makes the state of small past baby universes simple, just as the state of our past universe seems to have been simple when it was small. Now our universe has grown to be large and complicated, and so it it connects to the Hilbert space of small baby universes in initially simple states, information would naturally tend to go from our universe into the baby universes rather than the other way around.

Still, it provides us little solace that only the superobserver can understand what is going on. One would like to know how to describe physics in the universe that we have access to. In this regard, it is quite important to observe that, since the baby universe is closed, the energy that it carries away is precisely zero because of the lack of time and space translation symmetry. Its energy and momentum being precisely known, its position in spacetime is completely
undetermined. Thus, the baby universe wave function is really a global quantity in our universe, with no spacetime dependence. As was shown in [111, 112], this means that the baby universe Hilbert space has a natural basis, such that different elements of the basis correspond to different superselection sectors from the perspective of our universe.

(A large physical system with infinitely many degrees of freedom does not always visit every possible state, even if it has enough energy. For example, if a magnet is magnetized in a certain direction, each spin will fluctuate at any temperature, but the net magnetization will never change. The reason is that it is infinitely improbable that all the infinitely many spins at each different position will all fluctuate together in the same way. Most big systems have superselection sectors. In a solid, different rotations and translations which are not lattice symmetries define superselection sectors. In general, a superselection rule is a quantity that can never change through local fluctuations.)

In each superselection sector for the baby universe, it is in a unique pure quantum state and it follows that our universe is also described by a pure state. Mixed states arise only if one commits the unphysical act of superposing the different superselection sectors. The baby universe idea, then seems to lead us to the following picture: when a pure state collapses to form a black hole and then evaporates, it evolves to a pure state. This state is predictable in the sense that if we perform the experiment many times with the same initial state, we always get the same final state. But the result of the experiment might not be predictable from the fundamental laws of physics, it might depend on what superselection sector we happen to reside in. There may be many, many phenomenological parameters that we need to measure before we can predict unambiguously how a black hole with initial mass $M$ will evaporate, conceivably as many as $e^{S(M)}$.

Not only is this a disappointing conclusion, but we are still left without a satisfactory resolution of the information paradox. Once we have measured all of the relevant parameters, and can make predictions, we still long to learn the mechanism by which the black hole remembers the initial state so that it knows how to evaporate.

4.6.5 Other modifications of conventional theories

Another attempt to avoid the loss of information in black holes is to postulate that black holes never really form. For example, it was conjectured in [113, 114] that gravitational collapse might lead to no singularities or event horizons, only apparent horizons, and so no true black holes. Nevertheless, there would be a very large time delay before ingoing null rays become outgoing null rays, and there would be Hawking radiation. So the quantum-corrected system would appear much like a true semiclassical black hole, thus fulfilling the correspondence principle. Unfortunately, the present understanding of the principles of quantum gravity is too meagre to confirm or refute this conjecture. However, the reasoning in section 1.4.1, which states that the conditions for matter to go through its Schwarzschild radius need not to be in any way extreme, suggests that a quantized theory of gravity will not halt black hole formation.

An even more direct way to try to eliminate black holes is to assume a different classical theory of gravity. For example, it was postulated in [115, 116] that if the correct theory of gravity were
NGT (nonsymmetric gravity theory), the NGT charge could prevent black holes from forming. But even if NGT were a consistent theory of gravity, it would allow black holes to be formed from pure radiation without NGT charge, and so it would not really succeed in circumventing the problem. It would probably be very difficult for any simple consistent classical theory of gravity, which agrees with Newtonian gravity and with special relativity in the appropriate limits, to avoid producing black holes in all circumstances.

Other possibilities to resolve the information paradox could be that density matrices evolve deterministically but nonlinearly, or that density matrices have to be replaced by something more fundamental. But it is clear that one would like to avoid going down these roads unless all other possibilities are ruled out since they deviate so drastically from the conceptions we presently have of nature.

To conclude this section, it is fair to say that all of the possibilities listed here seem to require a rather drastic revision of cherished ideas about physics. The possibility that we are just overlooking something can practically be removed from the table and it appears that to resolve the information paradox we will have to take our understanding of nature’s ways to a deeper level. It seems increasing likely that it is as hopeless to reconcile relativistic quantum mechanics with black hole evaporation as it would have been to understand the spectrum of black body radiation using classical physics.

4.7 AdS/CFT and the information paradox

One could argue that the information paradox is solved by the discovery of the AdS/CFT duality, conjecturing the duality between string theory in anti-de Sitter spacetime and a conformal field theory on the boundary of anti-de Sitter [117]. Because gravity appears to be dual to a CFT, and the CFT is unitary, there cannot be any information loss and so there is no problem.

To see why this argument not holds, first look at what the information loss exactly tells us about quantum mechanics. It does not imply that quantum mechanics is no longer valid in laboratory situations, all that is states is that quantum mechanics is violated once a black hole is involved. So one cannot use tests of quantum mechanics in the everyday world to argue that there will be no problem when black holes are formed.

The same argument holds for the AdS/CFT correspondence. The known agreements between AdS gravity and the CFT involves comparison of scaling dimensions, \(n\)-point correlation functions, etc. But the information paradox does not say that any loss of unitarity occurs in normal \(n\)-particle scattering. It is only when a black hole is formed that a disagreement with unitarity shows up. The correspondences between AdS gravity and the CFT do not involve black hole formation and therefore do not adress the information paradox.

The arguments of section 4.4 equally apply to the AdS-Schwarzschild black hole for \(AdS_5 \otimes S^5\)

\[
ds^2 = \left( r^2 + 1 - \frac{C}{r^2} \right) dt^2 - \frac{dr^2}{r^2 + 1 - \frac{C}{r^2}} - r^2 d\Omega^2_5 \otimes d\Omega^2_5, \tag{4.135}
\]
which is similar to the usual Schwarzschild black hole in its essential respects. So a person who states that AdS/CFT resolves the information paradox has to give an explanation why local Hamiltonian evolution breaks down under the niceness conditions or has to provide a mechanism by which small corrections to the thermality of the Hawking radiation arise which encapture the necessary information. Either that or he has to accept stable remnants or the evolution of pure into mixed states, in which case he loses AdS/CFT and string theory as well since these are built on a foundation of usual quantum theory. So it appears that it is as hard to solve the information paradox in AdS as it is in usual asymptotically flat spacetime.

A different argument to evade the problem could be to use the CFT to define the gravity theory. Then the gravity theory has the expected weak field behavior and it will never violate quantum mechanics, so by construction there will never be a mixed state resulting from a pure state. But in that case, the arguments of section 4.4 also imply that one has to choose between the following options: (1) There are no traditional black holes in the theory, (2) the black hole horizon forms, in which case one should ask what extra conditions are necessary to get the right low-energy physics (e.g. quantum hair) or what the deficiencies are of the present low-energy model (e.g. the negligence of energy conservation), or (3) the theory contains stable remnants. So one can make no further claims on what exactly happens without studying the black hole formation/evaporation process in detail in either the CFT or the gravity theory.

So to solve the information paradox, one will have to provide a mechanism to get the information out of the black hole. One cannot do it with any abstract arguments like ‘AdS/CFT removes the paradox’. Solving the information paradox implies that one can tell what exactly happens in the information evaporation process.

### 4.8 Euclidean gravity and unitarity

As argued in the previous section, the discovery of the AdS/CFT duality does not solve the information paradox. But it is fair to say that it greatly favors the idea of information conservation. The discovery of AdS/CFT even persuaded Hawking, one of the greatest opponents of information conservation, to say that quantum gravity has to be unitary. Here, Hawking’s argument in favor of information conservation is given [118]. The outline of the argument is given here because it is an interesting idea to consider, meant to broaden the mind.

Black hole formation and evaporation can be thought of as a scattering process. One sends in particles and radiation from infinity and measures what comes back out to infinity. All measurements are made at infinity where the fields are weak, one never probes the strong field region in the middle. So one can’t be sure if a black hole forms or not, no matter how certain it might be in the classical theory. It will appear that this provides a possibility for information to be preserved and to be returned to infinity. Hawking uses the Euclidean path integral approach introduced in section 2.6 to study this phenomenon.

One might think that one should calculate the time evolution of the initial state by doing a path integral over all positive definite metrics that go between two space-like surfaces that are
a distance $T$ apart at infinity. One would then Wick rotate this interval $T$ to the Lorentzian time interval. However, the problem with this is that the quantum state for the gravitational field on an initial or final space-like surface is described by a wave function which is a functional of the geometries of the space-like surfaces and the matter fields on it

$$\Psi[h_{ij}, \phi, t],$$

where $h_{ij}$ is the three-metric of the surface, $\phi$ stands for the matter fields and $t$ is the time at infinity. But there is no gauge invariant way in which one can specify the time position of the surface in the interior.

One can measure the weak gravitational fields on a time-like tube around the system but not on the caps at the top and bottom which go through the interior of the system where the fields may be strong. This is shown on figure 4.8.

One way of getting rid of the difficulties of the caps would be to join the final surface back to the initial surface and integrate over all spatial geometries of the join. If this was an identification under a Lorentzian time interval $T$ at infinity, it would introduce closed time-like curves. But if the interval at infinity is the Euclidean distance $\beta$, the path integral gives the partition function for gravity at temperature $\beta^{-1}$

$$Z(\beta) = \int \mathcal{D}g \mathcal{D}\phi e^{-S[g, \phi]} = \text{tr}(e^{-\beta H}).$$

There is an infrared problem with this idea for an asymptotically flat space. The partition function is infinite because the volume of space is infinite. This problem can be solved by adding a small negative cosmological constant $\Lambda$ which makes the effective volume of the space of the order $\Lambda^{-3/2}$. It will not affect the evaporation of a small black hole but it will change infinity to

---

**Figure 4.8:** The time-like tube around the black hole formation-evaporation scattering process.
anti-de Sitter space and make the thermal partition function finite. It seems that asymptotically anti-de Sitter space is the only arena in which particle scattering in quantum gravity is well formulated.

The boundary at infinity has topology $S^1 \otimes S^2$. The path integral (4.137) that gives the partition function is takes over metrics of all topologies that fit inside this boundary. The simplest topology is the trivial topology $S^1 \otimes D^3$, where $D^3$ is the three disk. The next simplest topology and the first non-trivial topology is $S^2 \otimes D^2$. This is the topology of the Schwarzschild anti-de Sitter metric. There are other possible topologies that fit inside the boundary but these two are the important cases. The black hole here is eternal, i.e. it can not become topologically trivial at late times.

As already mentioned in section 2.6.2, the trivial topology can be foliated by a family of surfaces of constant time. The path integral over all metrics with trivial topology can be treated canonically by time slicing. The argument is the same as for the path integral of quantum fields in flat spacetime. One divides the time interval $T$ into time steps $\Delta t$. In each time step, one makes a linear interpolation of the fields and their conjugate momenta between their values on successive time steps. This method applies equally well to topologically trivial quantum gravity and shows that the time evolution, including gravity, will be generated by a Hamiltonian. This will give a unitary mapping between quantum states on surfaces separated by a time interval $T$ at infinity.

This argument can not be applied to the non-trivial black hole topologies. They can not be foliated by a family of surfaces of constant time because they don’t have any spatial cross-sections that are a three-cycle modulo the boundary at infinity. Any global symmetry would lead to conserved global charges on such a three cycle. These conserved charges would prevent correlation functions from decaying in topologically trivial metrics. Indeed, one can regard the unitary Hamiltonian evolution of a topologically trivial metric as a global conservation of information flowing through a three cycle under a global time translation. On the other hand, non-trivial black hole topologies won’t have any conserved quantity that will prevent correlation functions from decaying. It is therefore very plausible that the path integral over a topologically non-trivial metric gives correlation functions that decay to zero at late Lorentzian times. A way to look at this is that the correlation function decays more as more of the wave falls through the horizon into the black hole.

In this scattering approach, one can not just set up a small black hole, and watch it evaporate. All one can do, is to consider correlation functions of operators at infinity. One can apply a large number of operators at infinity, weighted with time functions, that in the classical limit would create a spherical ingoing wave from infinity, and in the classical theory would form a black hole. This would presumably then evaporate away. As described above, the path integral over metrics with trivial topology is unitary and information preserving. However, the information is lost in topologically non-trivial metrics. But in the case of those metrics, the correlation functions are rapidly decaying at late Lorentzian times. Maldacena even showed in the Ads/CFT context that the vacuum expectation value $\langle O(x)O(y) \rangle$ in dominant giant black hole solutions in anti-de Sitter decays exponentially as $y$ goes to late times and most of the effect of the disturbance at $x$ falls through the horizon of the black hole [119].
So in this viewpoint, everyone was right in a way. The confusion and paradox arose because people thought classically in terms of a single topology for spacetime. It was either $R^4$ or a black hole. But the Feynman sum over histories allows it to be both at once. One can not tell which topology contributed to the observation, any more than one can tell which slit the electron went through in the two slits experiment. All that observations at infinity can determine is that there is a unitary mapping from initial states to final states and that information is not lost. Quantum mechanics is safe.

Now how does information get out of a black hole? In section 4.6.1.1 it was shown that that particle creation by black holes can be thought of as tunnelling out from inside the black hole and that this process could carry information out of the black hole. But in the current viewpoint there is a problem with this description. Because strictly speaking, here, the only observables in quantum gravity are the values of the field at infinity. One can not define the field at some point in the middle because there is quantum uncertainty in where the measurement is done. In the semi-classical approximation one assumes that there is a large number $N$ of light matter fields coupled to gravity and that one can neglect the gravitational fluctuations because they are only one among $N$ quantum loops. However, in ignoring quantum loops, one throws away unitarity. A semi-classical metric is in a mixed state already. The information loss corresponds to the classical relaxation of black holes according to the no hair conjecture. One can not ask when the information gets out of a black hole because that would require the use of a semi-classical metric which has already lost the information.

This line of reasoning is very intriguing, but of course it needs to be supported by some detailed mathematical calculations before it can claim to resolve the information paradox. However, the arguments presented here are worth considering because in the light of the search for new principles, the information paradox should be approached with an open mind. Only in the process of exploring new and creative ideas one can expect progress towards a theory of quantum gravity.

On a sidenote, Hawking concluded his paper with the words: "In 1997, Kip Thorne and I, bet John Preskill that information was lost in black holes. The loser or losers of the bet were to provide the winner or winners with an encyclopedia of their own choice, from which information can be recovered with great ease. I gave John an encyclopedia of baseball, but maybe I should just have given him the ashes."

### 4.9 Thermodynamics of horizons

At this point, we’ve established the semiclassical framework of black hole formation and evaporation. It has many beautiful and inspiring features, but also some defects which most likely require extensions of the postulates and axioms which are the foundations of our current theories about nature. In the next chapters, we will discuss a specific set of postulates that might be required to incorporate the black hole formation and evaporation process in these theories. Before taking this next step, it is instructive to take a small detour and generalize some of the main aspects of the semiclassical results in black hole physics since it is still the primary goal of this subject to gain insights that might take our general understanding of nature to a deeper
level. In particular, we will show that horizons have a very general and natural relation to thermodynamics. The analysis below is based on [120].

In a certain spacetime, consider a time-like curve \( X^\mu(t) \), parametrised by the proper time of the clock moving along that curve. One can construct the past light cone for each event on this trajectory. The union \( U \) of all these past light cones determines whether an observer on this trajectory can receive information from all events in the spacetime or not. If \( U \) has a nontrivial boundary, there will be regions in the spacetime from which this observer cannot receive signals. In fact, one can extend this notion to a family of time-like curves which fill a region of spacetime, previously called a congruence. Given a congruence of time-like curves, i.e. a family of observers, the boundary of the union of their causal pasts will define a horizon for this set of observers. It will be assumed that each of the time-like curves has been extended to the maximum possible value for the proper time parametrising the curve. If the curves do not hit any spacetime singularity, this requires extending the proper time to infinite values. This horizon is dependent on the family of observers that is chosen, but is coordinate independent.

Given any family of observers in a spacetime, it is most convenient to interpret the results of observations performed by these observers in a frame in which these observers are at rest. So the natural coordinate system \( (t, x) \) attached to any time-like congruence is the one in which each trajectory of the congruence corresponds to \( x = \text{constant} \). This means the observers move on orbits of \( \partial/\partial t \). We will also assume that the spacetime has at least one Killing vector field and that we have chosen the coordinates \( (t, x) \) such that \( \partial g_{\mu\nu}/\partial t = 0 \). This means we define our family of observers as moving on time-like orbits of the Killing vector field \( \partial/\partial t \).

So let us now consider a general class of metrics which are

1) static in the \( (t, x) \) coordinate system, i.e. \( g_{00} = 0 \) and \( g_{ij}(t, x) = g_{ij}(x) \);
2) \( g_{00}(x) \equiv N^2(x) \) vanishes on some 2-surface \( H \) defined by the equation \( N^2 = 0 \);
3) \( \partial_i N \) is finite and non zero on \( H \);
4) all other metric components and the curvature remain finite and regular on \( H \).

The line element will now take the form

\[
ds^2 = N^2(t)dt^2 - \gamma_{ij}(x)dx^i dx^j. \tag{4.138}
\]

The comoving observers in this frame have trajectories \( x = \text{constant} \), four-velocity \( u_\mu = N\delta^0_\mu \) and four-acceleration \( a^\mu = u^\nu \nabla_\nu u^\mu = (0, a) \) which has the purely spatial components \( a_i = -(\partial_i N)/N \). The unit normal \((0, n)\) to the \( N = \text{constant} \) surface is given by

\[
n_i = -\partial_i N (g^{\mu\nu}\partial_\mu N\partial_\nu N)^{-1/2} = a_i(a_\mu a^\mu)^{-1/2}. \tag{4.139}
\]

The normal component of the acceleration \( a^\mu n_\mu \), ‘redshifted’ by a factor \( N \), has the value

\[
N n_\mu a^\mu = Na = (g^{\mu\nu}\partial_\mu N\partial_\nu N)^{1/2}. \tag{4.140}
\]
From the assumptions above about the metric, it follows that on the horizon \( N = 0 \), this quantity is finite. According to (1.83), this quantity is called the surface gravity \( \kappa = Na|_H \).

These static spacetimes, however, have a more natural coordinate system defined in terms of the level surfaces of \( N \). That is, one transforms from the original space coordinates \( x^i \) to the set \( (N, y^2, y^3) \) by treating \( N \) as one of the spatial coordinates. The \( y^i \) denote the two transverse coordinates on the \( N = \) constant surface. This can always be done locally, by possibly not globally since \( N \) could be multiple valued etc. However, we need this description only locally. The components of the four-acceleration in the \( (N, y^b) \) coordinates are

\[
\begin{align*}
a^N &= a^\mu \partial_\mu N = a^i a_i N = Na^2 \\
a^b &= a^\mu \frac{\partial y^b}{\partial x^\mu} \\
a_N &= a_i \frac{\partial x^i}{\partial N} = -\frac{1}{N} \frac{\partial N}{\partial x^i} \frac{\partial x^i}{\partial N} = -\frac{1}{N} \\
a_b &= a_i \frac{\partial x^i}{\partial y^b} = -\frac{1}{N} \frac{\partial N}{\partial x^i} \frac{\partial x^i}{\partial y^b} = 0.
\end{align*}
\]

Using these expressions, one can express the metric in the new coordinates as

\[
\begin{align*}
g^{NN} &= -N^2 a^2 = -\gamma^{\mu\nu} \partial_\mu N \partial_\nu N \\
g^{Nb} &= -Na^b.
\end{align*}
\]

The line element now becomes

\[
ds^2 = N^2 dt^2 - \frac{dN^2}{(Na)^2} - \sigma_{bc}(dy^b - \frac{a^b dN}{Na^2})(dy^c - \frac{a^c dN}{Na^2}).
\]

This metric describes the spacetime in terms of the magnitude of acceleration \( a \), the transverse components \( a^b \) and the metric \( \sigma_{bc} \) on the two surface and it maintains the \( t \)-independence. The \( N \) is now merely a coordinate and the spacetime geometry is described in terms of \( (a, a^b, \sigma_{bc}) \), all of which are, in general, functions of \( (N, y^b) \). In spherically symmetric spacetimes with horizon, one has \( a = a(N) \), \( a^b = 0 \) by choosing \( y^b = (\theta, \phi) \). Important features of the dynamics are usually encoded in the function \( a(N, y^b) \).

Near the \( N = 0 \) surface, \( Na \to \kappa \) and the metric reduces to the Rindler form

\[
ds^2 = N^2 dt^2 - \frac{dN^2}{(Na)^2} - dL^2 \approx N^2 dt^2 - \frac{dN^2}{\kappa^2} - dL^2.
\]

So this metric is a good approximation to a large class of static metrics with \( g_{00} \) vanishing on a surface.

To make the connection with black hole spacetimes, change the variable \( N \) to \( l \) according to

\[
dl = \frac{dN}{a}.
\]
Near the horizon, with \( Na \approx \kappa \), this can be integrated to \( l \approx N^2/2\kappa \). With the new coordinate \( l \), one can write (4.148) as
\[
ds^2 = f(l)dt^2 - \frac{dl^2}{f(l)} - dL^2.
\] (4.150)

Taking \( l = r \), \((y^2, y^3) = (\theta, \phi)\) and \( f(l) = (1 - 2GM/r) \), one finds the Schwarzschild black hole. Near the horizon, (4.150) becomes
\[
ds^2 \approx 2\kappa l dt^2 + \frac{dl^2}{2\kappa l} - dL^2.
\] (4.151)

Now with
\[
\frac{dl^2}{2\kappa l} = d\rho^2
\] (4.152)
equation (4.151) becomes
\[
ds^2 \approx \rho^2 d(\kappa t)^2 - d\rho^2 - dL^2,
\] (4.153)
which is identical to the previously found expression in section 2.6.1 when \((y^2, y^3) = (\theta, \phi)\).

In the metrics of the form in (4.148), the surface \( N = 0 \) acts as a horizon and the coordinates \((t, N)\) and \((t, l)\) are badly behaved near this surface. This is most easily seen by considering the light rays traveling along the \( N \)-direction in equation (4.148) with \( y^b = \text{constant} \). These light rays are determined by the equation
\[
\frac{dt}{dN} = \pm \frac{1}{N^2 a}.
\] (4.154)
So as \( N \to 0 \), one gets
\[
\frac{dt}{dN} \approx \pm \frac{1}{N \kappa}.
\] (4.155)
The slopes of the light cones diverge making the \( N = 0 \) surface act as a barrier dividing the spacetime into two causally disconnected regions in the \((t, N)\) coordinates and as a one-way membrane in the \((t, l)\) coordinates. This difference arises because the light cone \( T = X \) on figure 4.9 separates \( R \) from \( F \) and both regions are covered by the \((t, l)\) coordinates, the regions \( F \) and \( P \), however, are not covered in the \((t, N)\) coordinates. The following difference between the \((t, N)\) and \((t, l)\) coordinates needs to be stressed: In the \((t, N)\) coordinates, \( t \) is time-like everywhere (see (4.148)) and the two regions \( N < 0 \) and \( N > 0 \) are completely disconnected. In the \((t, l)\) coordinates, \( t \) is time-like where \( l > 0 \) and space-like where \( l < 0 \) (see (4.151) and the surface \( l = 0 \) acts as a one-way membrane. When we talk of \( l = 0 \) as a horizon, we often have the interpretation based on this feature.

The bad behaviour of the metric near \( N = 0 \) is connected with the fact that the observers at constant-\(x\) perceive a horizon at \( N = 0 \). Given a congruence of timelike curves, with a non-trivial boundary for their union of past light cones, there will be trajectories in this congruence which are arbitrarily close to the boundary. Since each trajectory is labelled by a \( x = \text{constant} \) curve in the comoving coordinate system, it follows that the metric in this coordinate system will behave badly at the boundary. But this bad behaviour can be removed by going to a local inertial frame near the horizon. The observers in this frame, i.e. freely falling observers, will have regular trajectories that cross the horizon. In a coordinate system where such freely falling observers are at rest and use their clocks to measure time, there will be no pathology at the
To construct the inertial coordinate system, introduce the tortoise coordinate $r^*$ to rewrite (4.148) as
\[ ds^2 = N^2(r^*)(dt^2 - dr^2) + dL^2 \] (4.156)

Introducing the null coordinates $u = t - r^*$ and $v = t + r^*$, one sees that near the horizon
\[ N \approx e^{\kappa r^*} = e^{\frac{\kappa}{2}(v-u)} \] (4.157)

where the $N > 0$ region was selected. So the horizon lies at $r^* \to -\infty$. This suggests the transformations to two new null coordinates $(U, V)$ with
\[ \kappa U = -e^{-\kappa u} \] (4.158)
\[ \kappa V = e^{\kappa v} \] (4.159)

which are regular at the horizon. The coordinates $(U, V)$ clearly are the generalization of the Kruskal-Szekeres coordinates of section 1.5. The corresponding inertial coordinates $(T, X)$ are then given by $U = T - X$ and $V = T + X$. Putting it all together, the transformation from the $(t, N)$ coordinate system to the $(T, X)$ coordinate system is given by
\[ \kappa X = e^{\kappa r^*} \cosh \kappa t \] (4.160)
\[ \kappa T = e^{\kappa r^*} \sinh \kappa t. \] (4.161)
Now we want consider quantum fields in a spacetime with a $N = 0$ surface. In the $(t, N)$ coordinate system, all physically relevant results in the spacetime will depend on the combination $N dt$ rather than on the coordinate time $dt$. As seen in the previous chapter, many interesting features of quantum fields in curved backgrounds can be investigated by using Euclidean metrics. The Euclidean rotation $t \rightarrow e^{i\pi/2}t$ can equivalently be thought of as the rotation $N \rightarrow Ne^{i\pi/2}$. However, this procedure becomes ambiguous on the horizon at which $N = 0$. But the family of observers with a horizon will be using a comoving coordinate system in which $N \rightarrow 0$ on the horizon. This ambiguity is solved rather naturally when one analytically continues in the time coordinate $t$ to the Euclidean sector. If we take $t_E = it$, then the metric near the horizon (4.148) becomes

$$ds^2 \approx N^2 dt_E^2 + \frac{1}{\kappa^2} dN^2 + dL^2,$$

after a redefinition to positive metric components. As already mentioned in section 2.6.1, one needs to interpret $t_E$ as an angular coordinate with $0 \leq t_E \leq 2\pi/\kappa$ in order to avoid the conical singularity at the origin. When we analytically continue in $t$ and map the $N = 0$ surface to the origin of the Euclidean plane, the ambiguity in defining $N dt$ on the horizon becomes similar to the ambiguity in defining the $\theta$ direction of the polar coordinates at the origin of the plane. This is resolved by imposing the periodicity in the angular coordinate.

The formulas (4.160) and (4.161) relating the $(t, N)$ coordinates to the $(T, X)$ coordinates now become

$$\kappa X = e^{\kappa r^*} \cos \kappa t_E$$

$$\kappa T_E = e^{\kappa r^*} \sin \kappa t_E.$$  

(4.163)

(4.164)

Where $T_E = iT$. Thus, the hyperbolic trajectories of constant $N$ now become cirkels, covering the entire $T_E - X$ plane. The horizon $N = 0$ lies at the origin. The complex plane probes the region which is classically inaccessible to the family of observers on $N = constant$ trajectories. A way to see this is to replace $\kappa t$ by $\kappa t - i\pi$ in (4.160), which changes $X$ to $-X$. So the complex plane contains information about the physics beyond the horizons through imaginary values of $t$. Thus, the 'forbidden region behind the horizon' simply disappears in the Euclidean sector.

This procedure of mapping the $N = 0$ surface to the origin of the Euclidean plane plays an important role. To see this role in a broader context, consider a class of observers who have a horizon. A natural interpretation of general covariance will require that these observers will be able to formulate quantum field theory entirely in terms of an 'effective' spacetime manifold made of regions which are accessible to them. Further, since the quantum field theory is well defined only in the Euclidean sector via the $i\epsilon$ prescription, it is necessary to construct an effective spacetime manifold in the Euclidean sector by removing the part of the manifold which is hidden by the horizon. As was shown above, for a wide class of metrics with horizon, the metric close to the horizon takes the Rindler form (4.162) in which the region inside the horizon is reduced to a point which we take to be the origin. The region close to the origin can be described in Cartesian coordinates, which correspond to the freely falling observer, or in polar coordinates, which would correspond to observers at rest in a Schwarzschild-type coordinates, in the Euclidean space. The effective manifold for the observers with horizon can now be thought to be the Euclidean manifold with the origin removed. This principle is of very broad validity
since it only uses the form of the metric very close to the horizon where it is universal.

Now one can construct a quantum field theory in the accessible region in \( N > 0 \) by integrating out the information contained in \( N < 0 \). That is, one family of observers may describe the quantum state in terms of a wave function \( \Psi(f_L, f_R) \) which depends on the field modes both on the 'left' \( (N < 0) \) and the 'right' \( (N > 0) \) sides of the horizon while another family of observers will describe the same system by a density matrix obtained by integrating out the modes \( f_L \) in the inaccessible region.

On the \( T = t = 0 \) hypersurface one can define a vacuum state \( |0\rangle \) of the theory by giving the field configuration for the whole of \( -\infty < X < +\infty \). This field configuration separates into two disjoint sectors when one uses the \((t, N)\) coordinate system. Concentrating on the \((T, X)\) plane and suppressing \( Y, Z \) for simplicity, we now need to specify the field configuration \( \psi_R(X) \) for \( X > 0 \) and \( \psi_L(X) \) for \( X < 0 \) such that it matches the initial data in the global coordinates. The vacuum state is then specified by the functional \( \langle 0 | \psi_L, \psi_R \rangle \).

\[ \langle 0 | \psi_L, \psi_R \rangle \]

Now make the transition to the Euclidean sector in the \((T_E, X)\) plane. The quantum field in this plane can be defined along standard lines. The analytic continuation in \( t \), however, is a different matter. As mentioned above, it can be seen from \((4.162)\) that the coordinates \((\kappa t_E, N)\) are like polar coordinates in the \((T, X)\) plane. This implies \( t_E \) to have a periodicity of \( 2\pi/\kappa \). Figure 4.10 makes it clear that evolving \( t_E \) from 0 to \( \pi \) will take the system from \( X < 0 \) to \( X > 0 \).

Now consider the ground state wave functional \( \langle 0 | \psi_L, \psi_R \rangle \) in the extended spacetime expressed as a path integral. The ground state wave functional can be represented as a Euclidean path integral of the form

\[ \langle 0 | \psi_L, \psi_R \rangle = C \int_{T_E=0}^{T_E=\infty} [D\psi] e^{-I_{E}}, \]

where \( C \) is a normalization constant. This equality follows from the standard procedure of computing the ground state by path integration via the Feynman-Hellman theorem. The Euclidean
action $I_E$ in (4.165) is evaluated as an integral over $T_E \geq 0$ and the integration over the field is constrained to equal $\psi = (\psi_L, \psi_R)$ on the $T_E = 0$ surface. From figure 4.10 it is clear that this path integral could also be evaluated in the polar coordinates by varying the angle $\theta = \kappa t_E$ from 0 to $\pi$. When $\theta = 0$, the field configuration corresponds to $\psi = \psi_R$ and when $\theta = \pi$, the field configuration corresponds to $\psi = \psi_L$. Therefore

$$\langle 0 | \psi_L, \psi_R \rangle = C \int_{\kappa t_E = 0; \, \psi = \psi_R}^{\kappa t_E = \pi; \, \psi = \psi_L} [D\psi] e^{-I_E}.$$  

(4.166)

In the Heisenberg picture, this path integral can be expressed as a matrix element of the Hamiltonian $H_R$ in the $(t, N)$ Rindler coordinates

$$C \int_{\kappa t_E = 0; \, \psi = \psi_R}^{\kappa t_E = \pi; \, \psi = \psi_L} [D\psi] e^{-I_E} = C \langle \psi_L | e^{-\pi H_R/\kappa} | \psi_R \rangle.$$  

(4.167)

So the path integral defining the vacuum functional is computed as a transition matrix element between the initial state $|\psi_R\rangle$ and the final state $|\psi_L\rangle$. This connection can be seen by interpreting $H_R$ as the generator of infinitesimal $t_E$ translations, i.e. infinitesimal rotations in the $T_E - X$ plane, and writing

$$e^{-\pi H_R/\kappa} = \lim_{m \to +\infty} (1 - \frac{\pi}{m} H_R)^m.$$  

(4.168)

Equation (4.167) has its origin in the fact that boost invariance in Lorentzian spacetime becomes rotational invariance in Euclidean spacetime.

Now the ground state wave functional can be normalized as follows

$$\sum_{\psi_L, \psi_R} |\langle 0 | \psi_L, \psi_R \rangle|^2 = \sum_{\psi_R \psi_L} \langle \psi_L | e^{-\pi H_R/\kappa} | \psi_R \rangle \langle \psi_R | e^{-\pi H_R/\kappa} | \psi_L \rangle$$

$$= \sum_{\psi_L} \langle \psi_L | e^{-2\pi H_R/\kappa} | \psi_L \rangle$$

$$= \text{tr}(e^{-2\pi H_R/\kappa}).$$  

(4.169)

So one gets

$$\langle 0 | \psi_L, \psi_R \rangle = \frac{\langle \psi_L | e^{-\pi H_R/\kappa} | \psi_R \rangle}{(\text{tr}(e^{-2\pi H_R/\kappa}))^{1/2}}.$$  

(4.170)
This result implies that for operators $\mathcal{O}$, made out of variables having support on $R$ ($N > 0$), the vacuum expectation value becomes thermal. This can be seen as follows

$$\langle 0 | \mathcal{O}(\psi_R) | 0 \rangle = \sum_{\psi_L} \sum_{\psi_R} \langle 0 | \psi_L, \psi_R \rangle \langle \psi_R | \mathcal{O} | \psi_R' \rangle \langle \psi_R' | \psi_L | 0 \rangle$$

$$= \sum_{\psi_L} \sum_{\psi_R} \langle \psi_L | e^{-\pi H_R/\kappa} | \psi_R \rangle \langle \psi_R | \mathcal{O} | \psi_R' \rangle \langle \psi_R' | e^{-\pi H_R/\kappa} | \psi_L \rangle$$

$$\frac{\text{tr}(e^{-2\pi H_R/\kappa})}{\text{tr}(e^{-2\pi H_R/\kappa})}$$

$$= \sum_{\psi_R} \langle \psi_R | e^{-2\pi H_R/\kappa} | \mathcal{O} | \psi_R' \rangle \frac{\text{tr}(e^{-2\pi H_R/\kappa})}{\text{tr}(e^{-2\pi H_R/\kappa})}$$

$$= \text{tr}(e^{-2\pi H_R/\kappa} \mathcal{O})$$

$$= \text{tr}(e^{-2\pi H_R/\kappa} \mathcal{O})$$

Thus, we come to the conclusion that tracing over the field configuration $\psi_L$ behind the horizon leads to a thermal density matrix $\rho \propto \exp[-2\pi H/\kappa]$ for observables in $R$. So the vacuum $|0\rangle$ can be expressed in terms of quantum states defined in $R$ and $L$ as

$$|0\rangle = \prod_i \left( \sqrt{1 - e^{-2\pi \omega_i/\kappa}} \sum_{n_i=0}^{\infty} e^{-\pi n_i \omega_i/\kappa} |n_i\rangle_R |n_i\rangle_L \right)$$

Compare with (4.50) and (4.81). This shows that when the vacuum is partitioned by the horizon at $N = 0$, it can be expressed as a highly correlated combination of states defined in $R$ and $L$. To avoid misunderstanding, it should be stressed that the temperature associated to a horizon is not directly related to the question of what a given non-inertial detector will measure. In the case of a uniformly accelerated detector in flat spacetime, it turns out that the detector results will match with the temperature of the horizon as was shown in section 2.2.2. In the case of black holes, the situation is more subtle since the discussion above holds for eternal black holes and particle creation has been shown in section 2.3.1 for gravitational collapse spacetimes. Backreaction effects also complicate the situation. But it can be shown that measurements of detectors will agree with the temperature of black holes [120]. There are, however, several other situations in which these two results do not match [121, 122].

Next to the thermality of horizons discussed above, also all the other classical thermodynamic features of black holes seem to generalize to any causal horizon. This can be seen by looking at the proofs that were given in section 1.11. The proof of the zeroth law only uses the fact that a black hole horizon is a Killing horizon and the Einstein equations, so it can readily be extended to any causal horizon. The area theorem relied on the fact that a horizon is a null surface, a property which is also satisfied by any other causal horizon. The existence of a first law for general causal horizons is less evident, but can nevertheless be shown to exist [123]. So combining all these arguments, one can conclude that any causal horizon will have a surface entropy density of $1/4G$.

In this section we have deflected our attention away from black holes and towards horizons.
It is sometimes considered a mystery how a black hole horizon could be capable of carrying so much entropy when after all it has no local significance since it is defined in terms of the future evolution of the spacetime, as was argued in section 1.4.1. Also, it is puzzling that when a star collapses and forms a black hole, the entropy suddenly rockets up to a value many orders of magnitude greater than it was in the star, ‘just because’ the horizon has formed. This becomes much less mysterious when it is realized that in essence the black hole really has nothing to do with it. As argued above, any causal horizon is endowed with a surface entropy density of $1/4G$.

The realization that horizon entropy is an intrinsically observer dependent notion raises the obvious question of what are the states that the horizon entropy counts. Surprisingly enough, the intuitive picture that it counts the number of configurations behind the horizon appears to be false [124]. A better way to look at it, is that to an outside observer, the horizon entropy somehow captures the number of ways that the world inside the horizon can affect the world outside. So a challenge to be met by any viable candidate for a microscopic theory of gravity is to explain this horizon entropy. Apart from their thermodynamics, horizons appear to have another very intriguing property that goes under the name of ‘the Holographic Principle’, which states that the entire description of the world behind any horizon can be fully done on its bounding surface [125]. However, the details of this principle are beyond the scope of this thesis.

### 4.10 Horizon entanglement entropy

As we saw in the previous section, the notion of black hole entropy can be generalized to any horizon. The origin of this entropy remains a puzzle. Especially its scaling with area makes it rather different from the usual entropy, for example the entropy of a thermal gas in a box, which is proportional to the volume.

In this section we will look at a possible quantum source for horizon entropy, entirely within the semiclassical approach. Namely, we will consider the short-distance fluctuations of quantum fields between modes on both sides of the horizon and calculate the corresponding entanglement entropy for an outside observer. This seems like a viable candidate to account for horizon entropy since it automatically has a scaling with the horizon area.

For a free massless scalar field, the two-point correlation function in $d$ spacetime dimensions has the standard form

$$\langle \psi(x)\psi(y) \rangle = \frac{\Omega_d}{|x-y|^{d-2}}, \quad (4.173)$$

where $\Omega_d = \Gamma\left(\frac{d-2}{2}\right)/4\pi^{d/2}$. This two-point function has the typical singular behavior when $x \to y$ which makes that quantum fields need renormalization in order to obtain physical results. From this observation it is intuitively clear that the typical behavior of the entanglement entropy in $d$ dimensions is

$$S \sim \frac{A(\Sigma)}{\epsilon^{d-2}}, \quad (4.174)$$

where $A(\Sigma)$ is the area of the horizon spatial cross section and $\epsilon$ is a UV-cutoff of the field theory. Below we will calculate the entanglement entropy more rigorously in flat spacetime...
and then sketch the extension to horizons in general relativity. Finally, we address the question whether horizon entropy can really be entanglement entropy. All of this will be done according to [72].

In this section we will refer to the horizon entropy derived in the previous section as the ther-
modynamical horizon entropy to make no confusion with the entanglement entropy.

4.10.1 Entanglement entropy in flat spacetime

Consider a quantum field $\psi(X)$ in a $d$-dimensional spacetime. We will work in a Euclidean spacetime with Euclidean time $t = i \tau$. Choose Cartesian coordinates $X^\mu = (\tau, x, z^i)$ where $i = 1, ..., d - 2$ such that the surface we will use to create our two subsystems is given by the condition $x = 0$ and the $z^i$ are the coordinates on $\Sigma$.

It will be convenient to use the polar coordinate system

$$\begin{align*}
\tau &= r \sin \theta \\
x &= r \cos \theta,
\end{align*}$$

where $\theta$ varies between 0 and $2\pi$. As mentioned in the previous section, boosts in Lorentzian spacetime become rotations in Euclidean spacetime, so if the field theory in question is relativistic then the field operator is invariant under the shifts $\theta \rightarrow \theta + w$, where $w$ is an arbitrary constant.

Just as in the previous section we will define the vacuum state of the quantum field by the path integral over the upper half of the Euclidean spacetime defined by $\tau \geq 0$ and impose the boundary condition $\psi(\tau = 0, x, z^i) = \psi_0(x, z^i)$

$$
\Psi[\psi_0(x, z^i)] = \int_{\tau=0}^{\tau=\infty; \psi(x, z^i) = \psi_0(x, z^i)} [D\psi] e^{-I_E}.
$$

The $d - 2$-surface $\Sigma$ separates the $\tau = 0$ surface in two parts, namely $x < 0$ and $x > 0$. These are the two subregions $L$ and $R$ that we will discuss.

The boundary data can be separated into $\psi_L = \psi_0(x, z^i)$ if $x < 0$ and $\psi_R = \psi_0(x, z^i)$ if $x > 0$. Contrary to the previous section, here we will work in the continuum case and not use discrete modes. By tracing out the modes $\psi_L$ in $L$ one defines a reduced density matrix in $R$

$$\rho(\psi_R^2, \psi_R^1) = \int [D\psi_L] \Psi(\psi_R^2, \psi_L) \Psi(\psi_R^1, \psi_L),$$

where the path integral goes over fields defined on the whole Euclidean spacetime except along the cut ($\tau = 0, x > 0$). In the path integral, the field $\psi(X)$ takes the boundary value $\psi_R^2$ above the cut and $\psi_R^1$ below the cut. The trace of the $n$-th power of the density matrix (4.178) is then given by the Euclidean path integral over fields defined on an $n$-sheeted covering of the cut spacetime. In the polar coordinates $(r, \theta)$, the cut corresponds to the values $\theta = 2\pi k, k = 1, 2, ..., n$. When passing across the cut from one sheet to another, the fields are glued together
analytically. Because the total $\theta$-angle adds up to $2\pi n$, this $n$-fold space is a flat cone $C_n$ with an angle deficit of $2\pi - 2\pi n = 2\pi(1 - n)$ at the surface $\Sigma$. To summarize, one has

$$\text{tr}\rho^n = Z[C_n], \quad (4.179)$$

where $Z[C_n]$ denotes the Euclidean path integral over the $n$-fold cover of the Euclidean space.

The trick to compute the entanglement entropy is to analytically continue $n$ to non-integer values. With this analytic continuation to real values of $\alpha$ one can compute

$$\left. (\alpha \frac{\partial}{\partial \alpha} - 1) \ln(\text{tr}\rho^\alpha) \right|_{\alpha=1} = \alpha \frac{1}{\text{tr}\rho^\alpha} \left. \frac{\partial}{\partial \alpha} \left( \sum_i \lambda_i^\alpha \right) \right|_{\alpha=1} - \ln(\text{tr}\rho)$$

$$= \alpha \frac{1}{\text{tr}\rho^\alpha} \left. \frac{\partial}{\partial \alpha} \left( \sum_i e^{\alpha \ln \lambda_i} \right) \right|_{\alpha=1} - \ln(\text{tr}\rho)$$

$$= \alpha \frac{1}{\text{tr}\rho^\alpha} \left( \sum_i \ln \lambda_i e^{\alpha \ln \lambda_i} \right) \left|_{\alpha=1} - \ln(\text{tr}\rho) \right)$$

$$= \frac{1}{\text{tr}\rho} \sum_i (\lambda_i \ln \lambda_i) - \ln(\text{tr}\rho)$$

$$= \frac{1}{\text{tr}\rho} \text{tr}(\rho \ln \rho - \rho \ln(\text{tr}\rho))$$

$$= \text{tr} \left( \frac{\rho}{\text{tr}\rho} \ln \left( \frac{\rho}{\text{tr}\rho} \right) \right)$$

$$= \text{tr}(\hat{\rho} \ln \hat{\rho}), \quad (4.180)$$

Now denote the eigenvalues of $\rho$ with $\lambda_i$. Then (4.180) can be written as

$$\left. (\alpha \frac{\partial}{\partial \alpha} - 1) \ln(\text{tr}\rho^\alpha) \right|_{\alpha=1} = \alpha \frac{1}{\text{tr}\rho^\alpha} \left( \sum_i \lambda_i^\alpha \ln \lambda_i \right) \left|_{\alpha=1} - \ln(\text{tr}\rho) \right)$$

Now introduce the effective action

$$W(\alpha) \equiv - \ln Z(\alpha), \quad (4.182)$$

where $Z(\alpha) = Z[C_\alpha]$ is the partition function of the field on a Euclidean space with conical singularity at the surface $\Sigma$ because of the angle deficit $2\pi(1 - \alpha)$. To remove the conical singularity, one has to make $\theta$ periodic with period $2\pi\alpha$, where $(\alpha - 1)$ is very small since we are only interested in the $\alpha \approx 1$-region in the derivation of (4.181). An important ingredient which makes this possible is the existence of the isometry $\theta \to \theta + w$ already noted above so that correlation functions with the required $2\pi\alpha$ periodicity can be constructed without any problem from the $2\pi$-periodic correlation functions. This allows one without any trouble to glue together pieces of Euclidean space to form a path integral over the conical space $C_\alpha$. Therefore, the analytic continuation of $\text{tr}\rho^\alpha$ to $\alpha$ different from 1 in the relativistic case is naturally defined by the path
Chapter 4. Entanglement and information

integral $Z(\alpha)$. This observation is strengthened by the fact that the analytical continuation appears to be unique [72].

So by the reasoning above, the definition (4.182) and the result (4.181) allow one to write the entanglement entropy as

$$S_{\text{ent}} = (\alpha \frac{\partial}{\partial \alpha} - 1)W(\alpha)\bigg|_{\alpha=1}.$$  \hfill (4.183)

One of the advantages of this method is that one does not need to care about the normalization of the reduced density matrix and can deal with a matrix which is not properly normalized.

Note again the important role for the conical singularity, this time at the surface $\Sigma$. It is this conical singularity that makes the entanglement entropy a surface effect in the derivation above. This is in complete analogy to section 2.6.2, where the removal of the conical singularity lead to a $S^2 \otimes R^2$ `cigar' topology which gave rise to the thermodynamic black hole entropy via the tip of the cigar which was non-linear in $\beta$. In the two cases, the conical singularity associates entropy with an area rather than a volume. But here, the conical singularity is introduced artificially as an intermediate tool to calculate the entanglement entropy while in section 2.6.2 it was naturally present.

We mentioned above that the isometry $\theta \to \theta + w$ allows one to construct $2\pi \alpha$-periodic correlation functions without any problem from the $2\pi$-periodic correlation functions. We will now illustrate this point with a bosonic field described by a field operator $\mathcal{D}$ so that the partition function is

$$Z = \frac{1}{\sqrt{2\pi}} \int [D\psi] e^{-\frac{1}{2} \int dX dX' \psi(X) \langle X|\mathcal{D}|X'\rangle \psi(X')} = (\det \mathcal{D})^{-1/2}. \hfill (4.184)$$

Now define the heat kernel $K(s, X, X') = \langle X|e^{-s\mathcal{D}}|X'\rangle$ as a solution to the heat equation

$$\left( \frac{\partial}{\partial s} + \mathcal{D} \right) K(s, X, X') = 0,$$ \hfill (4.185)

with boundary condition

$$K(s = 0, X, X') = \delta(X - X'). \hfill (4.186)$$

The effective action can be expressed as

$$W = -\ln(\det \mathcal{D})^{-1/2} = \frac{1}{2} \text{tr}(\ln \mathcal{D}) = \frac{1}{2} \int dX \langle X|\ln \mathcal{D}|X\rangle.$$

Now consider the integral

$$\int_z^{\infty} \frac{ds}{s} e^{-s} = -\gamma - \ln(z) + \sum_{k=1}^{\infty} \frac{(-1)^{k+1}z^k}{k!}, \hfill (4.187)$$
where $\gamma$ is the Euler constant. With this we can write
\[
\int_{\epsilon^2}^{\infty} \frac{ds}{s} e^{-as} = -\gamma - \ln(a) - \ln(e^2) + \sum_{k=1}^{\infty} \frac{(-1)^{k+1}(ae^2)^k}{k!}
\]
\[
= -\gamma - \ln(a) - \ln(e^2) + \sum_{k=1}^{\infty} \frac{(-1)^{k+1}(ae^2)^k}{k!}
\]
\[= -\gamma - \ln(a) - \ln(e^2) + \sum_{k=1}^{\infty} \frac{(-1)^{k+1}(ae^2)^k}{k!}
\] (4.188)

We will now use this formula with $a$ replaced by the operator $D$. The constant $\gamma$ will be ignored since it will drop out by normalization. $\epsilon$ will play the role of the regulator, exposing the divergent behavior in the regularization procedure. In taking the limit $\epsilon \to 0$, the sum over $k$ in (4.188) will disappear. To summarize, we can make the identification
\[
\ln(D) = -\int_{\epsilon^2}^{\infty} \frac{ds}{s} e^{-sD},
\]
where $\epsilon$ is a UV cutoff. So it follows that the effective action (4.187) can be expressed in terms of the heat kernel as
\[
W = -\frac{1}{2} \int_{\epsilon^2}^{\infty} \frac{ds}{s} \int dX \langle X | e^{-sD} | X \rangle
\]
\[= -\frac{1}{2} \int_{\epsilon^2}^{\infty} \frac{ds}{s} \text{tr} K(s),
\] (4.190)

The heat kernel $K(s, \theta, \theta')$ on regular spacetimes, where we omitted the coordinates other than $\theta$, will only depend on the difference $\theta - \theta'$ in the Lorentz invariant case because of the isometry $\theta \to \theta + w$. This function is $2\pi$-periodic with respect to $(\theta - \theta')$. The heat kernel $K_\alpha(s, \theta, \theta')$ on a space with a conical singularity is supposed to be $2\pi\alpha$-periodic. It is constructed from the $2\pi$-periodic version by applying the Sommerfeld formula [126]
\[
K_\alpha(s, \theta, \theta') = K(s, \theta - \theta') + \frac{i}{4\pi\alpha} \int_\Gamma \cot \left( \frac{w}{2\alpha} \right) K(s, \theta - \theta' + w) dw.
\] (4.191)

That this quantity still satisfies the heat kernel equation is a consequence of the isometry $\theta \to \theta + w$. The contour of integration $\Gamma$ consists of two vertical lines, one going from $(-\pi + i\infty)$ to $(-\pi - i\infty)$ and the other from $(+\pi + i\infty)$ to $(+\pi - i\infty)$. These lines intersect the real axis between the poles of $\cot(w/2\alpha)$: $-2\pi\alpha$, $0$ and $2\pi\alpha$ respectively. For $\alpha = 1$, the integrand in (4.191) is a $2\pi$-periodic function and the contribution from these two vertical lines cancel each other. Thus, for a small angle deficit the contribution of the integral in (4.191) is proportional to $(1 - \alpha)$.

Now we will use the methods developed above to calculate an explicit example. Consider the operator $D$ to be
\[
D = -\nabla^2.
\] (4.192)

One can use the Fourier transform to solve the heat equation (D.19). In $d$ spacetime dimensions one has
\[
K(s, X, X') = \frac{1}{(2\pi)^d} \int d^dp e^{ip_\mu(X^\mu - X'^\mu)} e^{-sF(p^2)}.
\] (4.193)
In the spherical coordinate system one has
\[ p^\mu (X^\mu - X'^\mu) = 2pr \sin \frac{w}{2} \cos \eta, \]
where \( w = \theta - \theta', \) \( p^2 = p_\mu p^\mu \) and \( \eta \) is the angle between the vectors \( p^\mu \) and \( (X^\mu - X'^\mu) \). The integration measure becomes
\[ \int d^dp = \Omega_{d-2} \int_0^\infty dp |p|^{d-2} \int_0^\pi d\eta \sin^{d-2} \eta, \]
where
\[ \Omega_{d-2} = \frac{2\pi^{(d-1)/2}}{\Gamma \left( \frac{d-1}{2} \right)} \]
is the area of a unit sphere in \( d - 1 \) dimensions. Performing the integration in (D.20) in these spherical coordinates one finds
\[ K(s, w, r) = \Omega_{d-2} \sqrt{2\pi} \left( \frac{2\pi}{d-2} \right)^{d-2} \int_0^\infty dp p^{d/2} J_{d-2}(2rp \sin \frac{w}{2}) e^{-sp^2}. \]
The trace then becomes
\[ \text{tr} K(s, w) = \frac{s}{4\pi^2 s} \frac{\pi \alpha}{(d-2)^2} A(\Sigma), \]
where \( A(\Sigma) = \int d^{d-2}z \) is the area of the surface \( \Sigma \). To obtain (4.198), one uses the integral
\[ \int_0^\infty dx x^{1-\nu} J_\nu(x) = \frac{2^{1-\nu}}{\Gamma(\nu)}. \]
The integral over the contour \( \Gamma \) in the Sommerfeld formula then gives
\[ C_2(\alpha) = \frac{i}{8\pi \alpha} \int_{\Gamma} \cot \left( \frac{w}{2\alpha} \right) \frac{dw}{\sin^2 \frac{w}{2}} = \frac{1}{6\alpha^2} (1 - \alpha^2). \]
Now collecting all the results, one finds
\[ \text{tr} K_\alpha(s) = \frac{1}{(4\pi s)^{d/2}} (\alpha V + 2\pi \alpha C_2(\alpha) s A(\Sigma)), \]
where \( V = \int d\tau d^{d-1}x \) is the volume of spacetime. So the effective action will contain two terms, the one proportional to \( V \) represents the vacuum energy. Since it is linear in \( \alpha \), it will give no contribution to the entanglement entropy. The second term proportional to the area \( A(\Sigma) \) is not linear in \( \alpha \). So applying (D.19), one gets
\[ S_{\text{ent}} = \frac{A(\Sigma)}{6(d-2)(4\pi)^{d/2} e^{d-2}}. \]
the surface will be important, but also the way it is embedded in the larger spacetime.

As a final remark, we would like to mention that in a theory where the two-point correlator behaves as

\[ \langle \psi(X)\psi(Y) \rangle \sim \frac{1}{|X-Y|^{d-2k}} \] (4.203)

the entanglement entropy scales as

\[ S \sim \frac{A(\Sigma)}{\varepsilon^{d-2/k}}. \] (4.204)

This implies that the entanglement entropy stays UV-divergent for all finite positive values of \( k \), even though the correlator becomes well behaved in the coincidence limit when \( k > d/2 \).

### 4.10.2 Entanglement entropy of Killing horizons

The definition of the entanglement entropy and the procedure for its calculation readily generalize to curved spacetime. The surface \( \Sigma \) can then be any smooth closed \( d-2 \) surface which divides the space in two subregions.

Of course, the notion of entanglement entropy is naturally applicable to horizons. Where in the previous cases we had to artificially introduce a surface that separated the space into two subsystems, general relativity now naturally provides us with such surfaces. Here, just as in the previous section, we will consider eternal horizons. In the black hole case, this means we do not consider backreaction and the corresponding shrinking effect on the horizon. We only work in the eternal black hole spacetime and its corresponding maximal extension.

In the construction from the previous section to obtain the entanglement entropy, \( \text{tr}\rho^n \) is given by the path integral over field configurations defined on the \( n \)-fold cover of the spacetime. This space was described by an angular coordinate which is periodic with period \( 2\pi n \). An important ingredient then was the isometry \( \theta \rightarrow \theta + w \) which allowed us to analytically continue \( n \) to arbitrary non-integer values \( \alpha \). The latter is not possible in a general spacetime. However, in the case we are considering, the surface \( \Sigma \) is a Killing horizon. So we know that the spacetime has a Killing vector field which can be expressed as \( \partial/\partial \theta \). More specifically, we saw in the previous section that we can rewrite the metric near almost any Killing horizon \( \Sigma \leftrightarrow N = 0 \) as

\[ ds^2 \approx N^2 dt_E^2 + \frac{1}{\kappa^2} dN^2 + dL^2. \] (4.205)

This leads to the identification \( (r, \theta) \sim (N, \kappa t_E) \). The metric (4.205) also clearly is invariant under \( \kappa t_E \rightarrow \kappa t_E + w \).

The presence of the so called rotational symmetry with respect to the Killing vector which generates rotations in the 2-plane orthogonal to the entangling surface \( \Sigma \) plays an important role in the construction to obtain the entanglement entropy. Without such a symmetry, it would be impossible to interpret \( \text{tr}\rho^n \) for an arbitrary \( \alpha \) as a partition function in some gravitational
background. Two points important for this interpretation. The first is that the spacetime possesses, at least locally near the entangling surface, a rotational symmetry such that, after the identification $\theta \to \theta + 2\pi \alpha$, we get a well defined $\alpha$-fold cover of the spacetime with no more than just a conical singularity. As explained above, this holds automatically if the surface in question is a Killing horizon. the second is that the field operator is invariant under $\theta \to \theta + \omega$. This is automatically satisfied if the field operator is a covariant operator. This allows us to use the Sommerfeld formula (4.191) in order to define the heat kernel on the $\alpha$-fold cover of the spacetime.

An interesting point is that the entanglement entropy does not depend on any gravitational field equation. Any metric containing a Killing horizon naturally provides us with a surface to which we can apply the mathematical toolbox of entanglement entropy. In this sense entanglement entropy is an off-shell quantity. This could be seen as a first indication that entanglement entropy might not be a good microscopic explanation for the thermodynamical horizon entropy since the thermodynamical framework of horizons does rely on the Einstein equations as was seen in the previous section.

Next to the off-shell nature of the entanglement entropy, this quantity also has another property which makes it a less probable candidate to explain the thermodynamical horizon entropy. Namely, it is proportional to the number of different field species which exist in nature. On the other hand, the thermodynamical horizon entropy does not seem to depend on any number of fields. This problem is known as the ‘species puzzle’.

Another apparent problem is that the entanglement entropy is a UV divergent quantity, while the thermodynamical horizon entropy is finite. This, however, does not cause much alarm. As well known, all one-loop quantities in quantum field theory are divergent if we do not apply a proper renormalization. So it is to be expected that the entanglement entropy can be made finite by the same kind of reasoning. However, it should be noted that although we have a strong feeling that the UV divergence of the entanglement entropy will disappear by renormalization, every model which explains horizon entropy as entanglement entropy will have to provide a precise mechanism for this. Moreover, after renormalization, the entanglement entropy should match the thermodynamical entropy $A/4G$. A possibility is that the renormalization of the Newton constant will make the entanglement entropy finite [127].

The model of induced gravity seems to solve all the problems above rather naturally [128]. In this approach the gravitational field is not fundamental but arises as a mean field approximation of the underlying quantum field theory of fundamental particles [129]. This is based on the fact that even if there is no gravitational interaction at tree level, it will appear at one-loop. The details of this mechanism will of course depend on the concrete model. However, because scalars and fermions are minimally coupled to gravity and gauge bosons are non-minimally coupled, it appears that although the induced Newton constant can be made finite, the entanglement entropy always remains UV divergent [72].

So in the end, we are lead to the conclusion that a more natural point of view is to consider the entanglement entropy of a horizon as the first quantum correction to the classical entropy $S = A/4G$ [130]. Indeed, the thermodynamical horizon entropy $S_{TH}$ can be considered
as classical, or tree-level entropy. If one restores the presence of $\hbar$, the thermodynamical horizon entropy is proportional to $\hbar^{-1}$ while the entanglement entropy is a $\hbar^0$ quantity. The total black hole entropy is then up to first order

$$S = S_{TH} + S_{ent}, \tag{4.206}$$

where all quantum fields that exist in nature contribute to the entanglement entropy $S_{ent}$.

It is clear that the intuitive notion of entanglement between modes of quantum fields in a classical background is far from the full story to explain the thermodynamical horizon entropy. It cannot provide us with a microscopic or statistical interpretation for this quantity. However, entanglement entropy has regained a lot of interest with the development of the holographic description of horizons which was referred to at the end of the previous section. But again, this matter is beyond the scope of this thesis.
Chapter 5

Black hole complementarity

"We have to remember that what we observe is not nature itself, but nature exposed to our method of questioning"
- Werner Heisenberg (1955)

In the previous chapter it was shown that the black hole evaporation process and unitarity have a difficult relation. In this chapter, however, we will not worry about this puzzle and simply assume that black holes are governed by entirely unitary dynamics. Leaving the information paradox for what it is, we would like to gain more insight in the structure of quantum black holes.

It will appear that there is another problem with assuming black hole evaporation is unitarity, namely there arises cloning of arbitrary quantum information, something which is not allowed by the linearity of quantum mechanics. This problem, together with the violation of baryon number discussed in the previous chapter, will be addressed here. Remarkably enough, a kind of reasoning that has already helped physicists in the past in the combination of the particle and wave properties of matter will again prove its value in this seemingly completely different context of black hole physics.

In this chapter there is also an important role for the stretched horizon of the membrane paradigm discussed in chapter 3. Its quantum variant seems to be an indispensable ingredient in the phenomenologically description of quantum black holes. More specifically, it will relate the properties of the quantum black hole to the thermodynamical behavior of the classical black hole.

5.1 A brick wall

When one considers the number of energy levels a particle can occupy in the vicinity of a black hole one finds a rather alarming divergence at the horizon. As seen in section 2.3.1, this infinity causes a black hole to be a source of an ideally random thermal radiation of particles. Therefore, the usual claim that a black hole is an infinite sink of information can be traced back to this infinity. Based on this observation, a first naive way to implement unitarity in evaporating black hole spacetimes is to simply cut off the particle wave functions around the horizon. Obviously
no information will be lost in that case. This might seem a physically unreasonable action since it only concentrates on the outside observer viewpoint and therefore violates the equivalence principle. But nevertheless, it will appear to be very instructive to see where this model takes us.

So let’s see what happens if we assume that the wave functions must all vanish within some fixed distance \( h \) from the horizon

\[
\psi(x) = 0 \quad \text{if} \quad r \leq 2GM + h .
\]  

This will be done by following the arguments of [101]. For simplicity, take \( \psi(x) \) to be a scalar wave function for a light particle, i.e. \( m \ll 1 \ll M \), with \( m \) the particle’s mass. To a freely falling observer, condition (5.1) corresponds to a uniformly accelerating mirror which will create its own energy-momentum tensor due to excitation of the vacuum [68]. So it is obvious that the introduction of this 'brick wall' will break the invariance under general coordinate transformations. But this model should be seen as an elementary exercise rather than an attempt to describe physical black holes accurately.

We also introduce an infrared regulator in the form of a box with radius \( L \)

\[
\psi(x) = 0 \quad \text{if} \quad r = L .
\]  

The quantum field \( \psi(x) \) is put in a Schwarzschild background with the usual metric

\[
ds^2 = \left(1 - \frac{2GM}{r}\right) dt^2 - \left(1 - \frac{2GM}{r}\right)^{-1} dr^2 - r^2 d\Omega^2 .
\]  

The field equation obtained by minimal coupling

\[
(g_{\mu\nu} \partial_\mu \partial_\nu + m^2)\psi = 0
\]

then becomes in spherical coordinates

\[
\left(1 - \frac{2GM}{r}\right)^{-1} \frac{\partial^2 \psi}{\partial t^2} - \frac{1}{r^2} \frac{\partial \psi}{\partial r} \left( r^2 \left(1 - \frac{2GM}{r}\right) \frac{\partial \psi}{\partial r} \right) + \frac{1}{r^2} l^2 \psi + m^2 \psi = 0 ,
\]  

which has as time-independent version

\[
- \left(1 - \frac{2GM}{r}\right)^{-1} E^2 \psi - \frac{1}{r^2} \frac{\partial \psi}{\partial r} \left( r(r - 2GM) \frac{\partial \psi}{\partial r} \right) + \frac{l(l+1)}{r^2} \psi + m^2 \psi = 0 .
\]  

As long as \( M >> 1 \) in Planck units, one can rely on a WKB approximation

\[
\left(1 - \frac{2GM}{r}\right)^{-1} E^2 \psi + \frac{r(r - 2GM)}{r^2} \frac{\partial^2 \psi}{\partial r^2} - \left( \frac{l(l+1)}{r^2} + m^2 \right) \psi \approx 0 .
\]  

Now define a radial wave number \( k(r, l, m) \) by

\[
k^2 = \frac{r^2}{r(r - 2GM)} \left( \left(1 - \frac{2GM}{r}\right)^{-1} E^2 - \frac{l(l+1)}{r^2} - m^2 \right) ,
\]
as long as the right hand side is non-negative, and \( k^2 = 0 \) otherwise. The number of radial modes \( n \) is given by

\[
n = \frac{1}{\pi} \int_{2GM+h}^{L} dr k(r, l, m).
\]  

(5.9)

The total number \( N \) of wave solutions with energy not exceeding \( E \) is then given by

\[
N = \int (2l + 1) n dl \\
= \frac{1}{\pi} \int_{2GM+h}^{L} dr \left( 1 - \frac{2GM}{r} \right)^{-1} \int dl (2l + 1) \sqrt{E^2 - \left( 1 - \frac{2GM}{r} \right) \left( m^2 + \frac{l(l+1)}{r^2} \right)} \\
\equiv g(E),
\]  

(5.10)

where the \( l \)-integration goes over those values of \( l \) for which the argument of the square root is positive.

So now we have counted the number of classical eigenmodes of a scalar field in the vicinity of a black hole. Now we would like to find the thermodynamic properties of this system. Every wave solution may be occupied by any integer number of quanta. Thus, the free energy \( F \) at some inverse temperature \( \beta \) is

\[
e^{-\beta F} = \sum_i e^{-\beta E_i} = \prod_{n,l,m} \frac{1}{1 - e^{-\beta E}} ,
\]  

(5.11)

or

\[
\beta F = \sum_N \ln(1 - e^{-\beta E}) .
\]  

(5.12)

So one gets, using (5.10),

\[
\beta F = \int dE (g(E) \ln(1 - e^{\beta E})) \\
= - \int_0^\infty dE \frac{\beta g(E)}{e^{\beta E} - 1} \\
= \frac{\beta}{\pi} \int_0^\infty dE \int_{2GM+h}^{L} dr \left( 1 - \frac{2GM}{r} \right)^{-1} \int dl (2l + 1) \\
\times \left( e^{\beta E} - 1 \right)^{-1} \sqrt{E^2 - \left( 1 - \frac{2GM}{r} \right) \left( m^2 + \frac{l(l+1)}{r^2} \right)} ,
\]  

(5.13)

where again the integral is taken only over those values for which the square root exists. In the approximation

\[
m^2 \ll \frac{2GM}{\beta^2 \hbar} , \quad L \gg 2GM
\]  

(5.14)

one finds that the main contributions are

\[
F \approx - \frac{2\pi^3}{45\hbar} \left( \frac{2GM}{\beta} \right)^4 - \frac{2}{9\pi} L^3 \int_m^\infty dE \frac{(E^2 - m^2)^{3/2}}{e^{\beta E} - 1} .
\]  

(5.15)

The second part is the usual contribution from the vacuum surrounding the system at large distances and is of little relevance here. The first part is an intrinsic contribution of the horizon
and is seen to diverge linearly as $h \to 0$.

The contribution of the horizon to the total energy is

$$U = \frac{\partial}{\partial \beta}(\beta F) = \frac{2\pi^3}{15h} \left( \frac{2GM}{\beta} \right)^4 Z,$$

and to the entropy

$$S = \beta(U - F) = \frac{8\pi^3}{45h} 2GM \left( \frac{2GM}{\beta} \right)^3 Z,$$

where a factor $Z$ has been added in both cases to denote the total number of particle types.

Now let’s adjust the parameters such that the total entropy becomes the right expression for a Schwarzschild black hole

$$S = 4\pi GM^2,$$

and use for $\beta$ the inverse Hawking temperature

$$\beta = \frac{2\pi}{\kappa}. \tag{5.19}$$

This allows one to determine the value for $h$

$$h = \frac{Z}{720\pi M}. \tag{5.20}$$

Note also that now the total energy becomes

$$U = \frac{3}{8} M, \tag{5.21}$$

which is independent of $Z$ and forms a sizeable fraction of the total mass $M$ of the black hole. It also follows that it does not make much sense to let $h$ decrease much below the value (5.20) because then more than the black hole mass would be concentrated at the outer side of the horizon.

Equation (5.20) also seems to suggest that $h$ depends on $M$, but this is merely a coordinate artifact. The invariant distance is

$$\int_{r=2GM}^{r=2GM+h} ds = \int \frac{dr}{\sqrt{1 - 2GM/r}} = 2\sqrt{2GMh} = \sqrt{\frac{Z}{90\pi}}. \tag{5.22}$$

Thus, the brick wall may be seen as a property of the horizon, independent of the size of the black hole.

The conclusion here is that the infinity of modes near the horizon should be cut off. Quantum fields seem to contain too many degrees of freedom to faithfully describe a black hole. Moreover, it appears that the value for the cut-off parameter is determined by nature, and a property of the horizon only. The model above could be considered as a reasonable description of a black hole.
as long as the particles near the horizon are kept at the Hawking temperature and all chemical potentials are kept close to zero. The interesting point is that there exists a classical analog of this brick wall, namely the stretched horizon that was introduced in the context of the membrane paradigm presented in chapter 3. So in both the classical and the quantum description of black holes there appears to be a physical role for this thin boundary layer at the horizon.

By restricting the wave functions to the outer side of the horizon, the model is unitary by definition. But clearly, it also has it’s shortcomings. Only the picture for an outside observer has been treated consistently here, the above description is definitely not valid for infalling observers. So the invariance under general coordinate transformations is broken. This results in a clear conservation of baryon number for example, something which is definitely not the case for the true physical situation of black hole evaporation as was explained in section 4.6.2 of the previous chapter. In the sections below a principle will be presented that adresses the question of how to keep not only unitarity but also invariance under general coordinate transformations while dropping all global conservation laws.

5.2 Problems with information in the Hawking radiation

In the previous section we only worried about the outside observer. Here however, we will again take into account the equivalence principle. So let us again consider the picture where we foliate the spacetime of black hole formation and evaporation with a complete family of Cauchy surfaces. This was already done on figure 4.7 in the previous chapter, where it was used to argue that at first sight there is a conflict between black hole evaporation and information conservation. But here we will look at the same figure from a different point of view. We will simply impose unitary evolution in the process of black hole formation and evaporation and see where this leads us. The line of thought below was presented in [91, 131]. For convenience, figure 4.7 is repeated here as figure 5.1.

Again, we will assume that state vectors on one Cauchy surface evolve to another Cauchy surface in the future by a linear and local evolution equation. With this equation, an initial state \(|\psi(\Sigma)\rangle\) defined on some Cauchy surface \(\Sigma\) which does not intersect the black hole can be evolved without encountering any singularity until the surface \(\Sigma_P\) is reached. \(\Sigma_P\) is the surface which contains the point \(P\) where horizon and singularity meet, as can be seen on figure 5.1. \(P\) divides \(\Sigma_P\) in \(\Sigma_{bh}\) and \(\Sigma_{out}\), which respectively lie inside and outside the black hole. The Hilbert space of states on \(\Sigma_P\) can be written as a tensor product space of functionals of the fields on \(\Sigma_{bh}\) and \(\Sigma_{out}\), i.e. \(H_P = H_{bh} \otimes H_{out}\).

Now consider on figure 5.1 the Cauchy surface \(\Sigma'\) long after the black hole has evaporated. If we assume unitarity, then the state \(|\psi(\Sigma')\rangle\) on this surface has to be pure, of course assuming that \(|\psi(\Sigma)\rangle\) was pure. In other words, there exists a unitary scattering matrix \(S\) such that \(|\psi(\Sigma')\rangle = S|\psi(\Sigma)\rangle\). By assumption, \(|\psi(\Sigma')\rangle\) has evolved from some state \(|\chi(\Sigma_{out})\rangle\) defined on \(\Sigma_{out}\) by a linear and local evolution equation. So \(|\chi(\Sigma_{out})\rangle\) also has to be pure. This, in turn, implies that \(|\psi(\Sigma_P)\rangle\) must be a product state

\[
|\Psi(\Sigma_P)\rangle = |\Phi(\Sigma_{bh})\rangle \otimes |\chi(\Sigma_{out})\rangle,
\]

(5.23)
where $|\Phi(\Sigma_{bh})\rangle \in \mathcal{H}_{bh}$ and $|\chi(\Sigma_{out})\rangle \in \mathcal{H}_{out}$. This product state is obtained from linear, local evolution from the initial state $|\Psi(\Sigma)\rangle$. But as argued above, $|\chi(\Sigma_{out})\rangle$ alone depends linearly on $|\Psi(\Sigma)\rangle$. So we arrive at the conclusion that the state $|\Phi(\Sigma_{bh})\rangle$ inside the black hole must be independent of the initial state!

Another way to look at the situation is the following. Construct a Cauchy surface that crosses most of the outgoing Hawking radiation and also crosses the collapsing body well inside the horizon. Of course, this surface is constructed such that it stays far from the singularity in regions of low curvature, so that we are confident that we know the causal structure reliably. Let $\{|i\rangle\}$ denote a basis for the initial quantum state of the collapsing body, and take the extreme view that each of these states evolves to a state on the Cauchy surface constructed above, such that the radiation and the collapsing body are completely uncorrelated. So the final state is the tensor product of a pure state inside the horizon and a pure state outside

$$|i\rangle \rightarrow |i\rangle_{\text{inside}} \otimes |i\rangle_{\text{outside}}.$$

But one may also consider a superposition of these basis states, which evolves as

$$\sum_i c_i|i\rangle \rightarrow \sum_i c_i(|i\rangle_{\text{inside}} \otimes |i\rangle_{\text{outside}}).$$

In general, the state inside and outside will be correlated, unless all of the states $|i\rangle_{\text{inside}}$ are actually the same state. So the radiation will always be in a pure state only if the body is in a unique state. More generally, if the radiation state is nearly pure, then the body’s state must
be nearly unique.

The above arguments imply that if the information really propagates out encoded in the Hawking radiation, then there must be a mechanism that strips away all information about the collapsing body as the body falls through the horizon, thus long before it reaches the singularity. This bleaching of information clearly is in contrast with the equivalence principle since to a freely falling observer the horizon is not a special place. If this bleaching of information at the horizon does not occur, then macroscopic violation of causality seems to be required to transport the information from the collapsing body to the outgoing radiation.

It’s instructive to compare the viewpoints of this section and the previous. In the previous section, the introduction of a ‘brick wall’ lead to a model that was manifestly unitary. In this section, imposing unitarity results in the conclusion that there must happen something special around the horizon. Namely, the information seems to ‘bounce back’ of the horizon without ever entering the black hole. These two very different approaches seem to be remarkable consistent in the sense that they both predict a special thin boundary layer at the horizon which plays a physical role. On top of that, this thin boundary layer has a classical analogon in the membrane paradigm. However, the two viewpoints focus only on the outside observer and they are both in conflict with the equivalence principle. So there appears to be a missing ingredient.

5.3 Average information in the Hawking radiation

Unitary evolution implies that if the matter collapsed to form a black hole was in a pure state, the black hole and its surrounding Hawking radiation are two subsystems of a combined system which also is in a pure state. Tracing over the black hole subsystem gives a density matrix for the radiation subsystem that generically is mixed. In this section we would like to find out what the typical information in the radiation subsystem is at various stages of the black hole evaporation. In order to give the exact answer to this question the precise mechanism behind the unitary evolution needs to be known, something which is not the case at present times. Therefore, it will be examined what the generic behaviour will be by taking the black hole and the Hawking radiation in a random pure state. The analysis is done according to [132].

To control the dimensions of the Hilbert spaces involved, we imagine forming the black hole from a pure state of radiation or matter in a box. We take the dimension of the total Hilbert space, i.e. black hole plus radiation, to be $nm$. $m$ is the dimension of the radiation subsystem and is related to its thermodynamic entropy $s_R$ as $m \sim e^{s_R}$. $n$ is the dimension of the black hole subsystem, with $n \sim e^{s_B}$. $s_B$ is the usual black hole entropy, so $s_B = A/4G$. The density matrices of the two subsystems are obtained by tracing out the other subsystem

$$\rho_R = \text{tr}_B \rho_{BR}$$

$$\rho_B = \text{tr}_R \rho_{BR},$$
where $R$ stands for the radiation subsystem, $B$ for the black hole system and $BR$ for the total pure system. Both systems have an entanglement entropy given by

$$S_R = -\text{tr}(\rho_R \ln \rho_R)$$

(5.28)

$$S_B = -\text{tr}(\rho_B \ln \rho_B).$$

(5.29)

Because the total system is pure, its entanglement entropy $S_{BR}$ is zero. So it follows from the subadditivity of entanglement entropy

$$|S_B - S_R| \leq S_{BR} \leq S_B + S_R$$

(5.30)

that $S_B = S_R$.

The information of a system is defined here as the deficit of the entanglement entropy from its maximum possible value. This definition follows from the interpretation of entropy as the 'lack of information'. So the black hole and radiation subsystem carry an information given by

$$I_R = \ln m - S_R$$

(5.31)

$$\approx s_R - S_R$$

(5.32)

$$I_B = \ln n - S_B$$

(5.33)

$$\approx s_B - S_B.$$  

(5.34)

To obtain the generic behavior of the quantities above, they are averaged over all random pure states of the total system. The average is defined with respect to the unitarily invariant Haar measure on the space of unit vectors in the $mn$-dimensional Hilbert space of the total system. This Haar measure is proportional to the standard geometric hypersurface volume on the unit sphere $S^{2mn-1}$ which those unit vectors give when the $mn$ complex-dimensional Hilbert space is viewed as the $2mn$ real-dimensional Euclidean space. For $m \leq n$, the average information in the radiation subsystem appears to be [133]

$$\langle I_R \rangle = \ln m + \frac{m - 1}{2n} - \sum_{k=n+1}^{mn} \frac{1}{k}.$$  

(5.35)

For $m \gg 1$, this can be shown to be [133]

$$\langle I_R \rangle \approx \frac{m}{2n} \sim e^{s_R - s_B}.$$  

(5.36)

By using (5.31) and (5.33), together with $S_R = S_B$, it follows that

$$I_B = \ln n - \ln m + I_R,$$  

(5.37)

which after averaging and using (5.36) becomes

$$\langle I_B \rangle = \ln n - \ln m + \frac{m}{2n}.$$  

(5.38)
So the results above imply that almost all the information giving the precise pure state of the entire system, \( \ln m + \ln n \) units, is in the correlations between the subsystems. Equation (5.36) shows that for a typical pure state of the entire system, very little of the information, roughly \( m/2n \) unit, is in the correlations within the smaller subsystem itself. Roughly \( \ln n - \ln m + m/2n \) units is in the correlations within the larger subsystem itself and the remaining roughly \( 2\ln m - m/n \) units of information are in the correlations between the larger and smaller subsystems.

If \( n \leq m \), one gets analogously

\[
\langle I_B \rangle = \ln n + \frac{n-1}{2m} - \sum_{k=m+1}^{mn} \frac{1}{k}.
\]  

(5.39)

Now (5.37) can be rewritten as

\[
I_R = \ln m - \ln n + I_B.
\]  

(5.40)

So for \( n \leq m \) and using (5.39), this gives

\[
\langle I_R \rangle = \ln m + \frac{n-1}{2m} - \sum_{k=m+1}^{mn} \frac{1}{k} \approx \ln m - \ln n + \frac{n}{2m}.
\]  

(5.41)

The average information in the radiation subsystem \( \langle I_R \rangle \), together with the average entanglement entropy \( \langle S_R \rangle = \ln m - \langle S_R \rangle \), is plotted in figure 5.2 against the thermodynamic entropy \( s_R = \ln m \) of the radiation. This is done for \( mn = 291600 \), whose 105 integer divisors are taken to be the values for \( m \).

The above analysis allows us to conclude that when the radiation emitted from a black hole has a smaller Hilbert space dimension than that of the remaining black hole, the radiation would typically have very little information in it and would be very nearly maximally mixed. Alternatively, consider the case in which the black hole has emitted most of its energy so that the radiation has the larger dimension. If one then examines only part of the radiation at a time so that each part has a smaller dimension than the rest of the system, one would expect to see in the separate parts only a very tiny amount of the information. The total information is instead mostly encoded in the correlations between all the parts. From figure 5.2 is also clear that information typically starts to 'leak out' of a black hole after it has evaporated about one half of its initial entropy. The time it takes for a black hole, starting from its initial state, to reach the point where it starts to release its information is called the 'information retention time' or 'Page time'. This point of time is clearly visible on figure 5.2. A black hole that has already past its Page time is called an old black hole.

### 5.4 The postulates

Based on the observations of the previous sections it is clear that there is something missing in the quantum framework of black holes. This missing ingredient goes under the name of black
Black hole complementarity. In its simplest form it just states [9]

**Black hole Complementarity** *No observer ever witnesses a violation of the laws of physics.*

Basically, the idea is that for an outside observer, the black hole is a hot membrane which can absorb, thermalize and eventually re-emit all information in the form of Hawking radiation. The number of degrees of freedom on this membrane is the exponential of the entropy of the black hole. The surface density of these degrees of freedom is constant on the horizon, namely about 1 degree of freedom per Planck area, so an incoming energy flux or outgoing Hawking radiation will cause degrees of freedom to pop into or out existence in order to keep the density constant. This boundary layer is called the stretched horizon and the idea is of course imported from the brick wall calculation of section 5.1 and the membrane paradigm of chapter 3, from where it has taken its name. To an outside observer, the microphysical degrees of freedom on the horizon appear in the quantum Hamiltonian used to describe the observable world. These degrees of freedom must be of sufficient complexity such that they behave ergodically and lead to a coarse-grained, dissipative description of the membrane.

To give a more exact definition of the stretched horizon, one can proceed as follows. At a point on the global event horizon, contruct the radial null geodesic which does not lie in the horizon. That ray intersects the stretched horizon at a point where the area of the transverse two-sphere has increased by an amount of order one Planck unit relative to its value at the corresponding point on the event horizon. The generators of the horizon can be thought of as a two-dimensional fluid. The points of this fluid can be mapped to the stretched horizon, thereby
defining a fluid flow on that surface. As seen in chapter 3, at the classical level the stretched horizon behaves as a continuous, viscous fluid. A natural candidate for the microphysics of the stretched horizon is to replace the continuous classical fluid with a fluid of discrete 'atoms'.

When a shell of matter collapses to form a black hole, it will be blue-shifted relative to stationary observers. So when it arrives at the stretched horizon, it has Planckian wavelengths. Thereupon it interacts with the 'atoms' of the stretched horizon leading to an approximately thermal state. The subsequent evaporation yields approximately thermal radiation but with non-thermal long time correlations. These non-thermal effects not only depend on the incoming pure state but also on the precise nature of the Planck-scale 'atoms' and their interaction with the blue-shifted matter. The evaporation products then climb out of the gravitational well and are red-shifted to low energy. The result is that the very-low energy Hawking radiation from a massive black hole has non-thermal correlations which contain detailed information about Planck-scale physics. Thus, the blueshift can be seen as a 'magnifying glass' to expose the physics at the Planck scale. This phenomenon is reminiscent of the imprinting of Planckian fluctuations onto the microwave background radiation by inflation.

Now consider an observer at the stretched horizon who counts the number of particles emitted per unit proper time. Since the stretched horizon is always at the Planck temperature, the number of particles emitted per unit area per unit proper time is order one in Planck units. If all these particles made it out to infinity, then a distant observer would estimate a number of particles emitted per unit time which is obtained by multiplying by the black hole area and the time dilatation factor (in Planck units)

\[
\frac{dN}{dt} \sim M^2 \frac{d\tau}{dt} \sim M \quad (5.43)
\]

On the other hand, the number per unit time of particles that actually emerge to infinity is obtained by multiplying the black hole luminosity \( L \sim M^{-2} \) by the inverse energy of a typical thermal particle at the Hawking temperature. This gives

\[
\frac{dN}{dt} \sim \frac{1}{M} \quad (5.44)
\]

So it seems that most of the particles emitted from the stretched horizon do not get to infinity. In fact, as we saw in section 2.4, only those particles emitted with essentially zero angular momentum reach distant observers, the rest scatters back into the hole. This gives rise to a thermal atmosphere above the stretched horizon which only slowly evaporates and whose repeated interaction with the stretched horizon ensures thermal equilibrium.

From the reasoning above, it is clear that the analysis of section 5.3 is particularly appropriate to the complex and ergodic behavior of the stretched horizon. This conclusion is only enforced by the existence of the thermal atmosphere. Therefore, we arrive at the following picture of the evaporation process. At the beginning, the total entanglement entropy of the combined system of stretched horizon and radiation is zero, but the radiation is correlated to the degrees of freedom of the stretched horizon. More time elapses, and the stretched horizon emits more quanta. The previous correlations between the stretched horizon and the radiation field are now replaced by correlations between the early part of the radiation and the newly emitted quanta.
In other words, the features of the exact radiation state which allow the entanglement entropy of the radiation system to return to zero are long time correlations spread over the entire time occupied by the outgoing flux of energy. The local properties of the radiation are expected to be thermal. For example, the average energy density, short time radiation field correlations, and similar quantities that play an important role in the semi-classical dynamics should be thermal. The long time correlations which restore the entanglement entropy to zero are not important to average coarse grained behavior.

To conclude, in the stretched horizon picture, a black hole evaporates in complete analogy to the burning up of a normal object. However, because the stretched horizon is a very complex and chaotic system, computing an $S$-matrix would be as daunting as computing the scattering of laser light from a piece of coal. The validity of quantum field theory in this case is not assured by exhibiting an $S$-matrix, but by identifying the underlying atomic structure and constructing a Schrödinger equation for the many particles composing the coal and the photon field to which it is coupled. Although the equations cannot be solved, we nevertheless think we understand the route from quantum theory to apparently thermal radiation via statistical mechanics. In the case of the stretched horizon, the underlying microphysics is not yet understood.

For an infalling observer, black hole complementarity states that the equivalence principle is respected. So as long as the black hole is much larger than the infalling system, the horizon is just flat spacetime without any special properties. No high temperatures or other anomalies are detected.

These outside and infalling viewpoints, together with the idea that they are not at all in conflict with each other is the basic idea of black hole complementarity. So the 'bleaching’ of information that was encountered in section 5.2 does not happen as far as the infalling observer is concerned. However, in the description of the outside observer it will have taken place since he will detect that same information in the Hawking radiation. The key idea is that the two observers will never be able to compare their constatations. Only a 'superobserver’ outside our universe would be able to see the information twice. So the picture coming from conventional quantum field theory in an evaporating black hole background that a single state vector describes both the interior and exterior of the black hole must be wrong if black hole complementarity is correct.

Black hole complementarity is usually formulated via a set of 4 postulates [131]:

**Postulate 1** The process of formation and evaporation of a black hole, as viewed by a distant observer, can be described entirely within the context of standard quantum theory. In particular, there exists a unitary $S$-matrix which describes the evolution from infalling matter to outgoing Hawking-like radiation.

**Postulate 2** Outside the stretched horizon of a massive black hole, physics can be described to good approximation by a set of semiclassical field equations.

**Postulate 3** To a distant observer, a black hole appears to be a quantum system with discrete energy levels. The dimension of the subspace of states describing a black hole of mass $M$ is the exponential of the black hole entropy.
Postulate 4 A freely falling observer experiences nothing out of the ordinary when crossing the horizon.

The first postulate just expresses the unitary evolution of black hole formation and evaporation. The second expresses the validity of the semiclassical approach outside a massive black hole. The third postulate states that the origin of the thermodynamic behavior of a black hole is the coarse graining of a large, complex, ergodic but conventionally quantum mechanical system. The fourth postulate is a formulation of the equivalence principle. The first three postulates involve an outside observer and the fourth applies to infalling observers.

At first sight, the idea of black hole complementarity seems a wild leap of faith. It definitely challenges the conventional way of thinking about black holes. To the skeptic, black hole complementarity might seem a way to deny the problems instead of seeking a solution to them. Nevertheless, the idea has hold stand for a long time now and there are many thought experiments that indicate it is true. In the next section we will take a look at some of these thought experiments.

5.5 Thought experiments

In the early part of the past century, the contradictions between the wave and the particle theories of light seemed irreconcilable. But careful thought could not reveal any logical contradiction. Experiments of one kind or the other revealed either particle or wave behavior, but neither both. The present situation in black hole physics is similar. An experiment of one kind will detect a quantum membrane, while an experiment of another kind will not. However, no possibility exists for any observer to know the results of both. The results of the two kinds of experiments are complementary. Here, we will analyse this situation by a set of gedanken experiments [134] which will provide us with examples of 'black hole complementarity at work'. The main conclusion of the gedanken experiments below will appear to be that any violation of black hole complementarity requires Planck-scale physics.

5.5.1 Verification of the stretched horizon

A first experiment that directly comes to mind is for an outside observer to simply check the existence of the stretched horizon by going to the horizon and seeing if he really finds this hot membrane containing all the information. Since the stretched horizon is defined as the time-like surface where the area of the transverse two-sphere is larger than at the null event horizon by order one in Planck units, the proper acceleration of a point on the stretched horizon at fixed angular position is approximately one Planck unit. So any observer who penetrates all the way to the stretched horizon will have to undergo Planck scale acceleration to return. As a result this experiment cannot be analyzed in terms of known physics and therefore it cannot at present be used to rule out the existence of the stretched horizon.

Next, consider an experiment in which a freely falling observer, who passes through the event
horizon, attempts to continuously send messages to the outside reporting the lack of substance of the membrane. First suppose that these messages are carried by radiation of bounded frequency in the freely falling frame. Because the observer has only a finite proper time before crossing the Rindler horizon only a finite number of bits of information can be sent. The last few bits get enormously stretched by the red shift factor and are drowned by the thermal noise. Therefore, there is in a sense a last useful bit. If the carrier frequency is less than the Planck frequency the last useful bit will be emitted before the stretched horizon is reached. In order to get a message from behind the stretched horizon, the observer must use super-Planckian frequencies. Again, the experiment cannot be analyzed using conventional physics.

So in both these experiments, efforts made to investigate the physical nature of the stretched horizon are frustrated by our lack of knowledge of Planck scale physics.

5.5.2 Baryon number violation

As argued in the previous chapter, the evaporation of black holes leads to the violation of conservation of baryon number. Here, we will look at this phenomenon in the context of black hole complementarity.

The conservation of baryon number is the basis for the stability of ordinary matter. Nevertheless, there are reasons to believe that baryon number, unlike electric charge, can at best be an approximate conservation law. This idea is supported by the observed matter anti-matter asymmetry in our universe [135]. The difference between baryon number and electric charge is that baryon number is not the source of a long range gauge field. Thus it can disappear without some flux having to suddenly change at infinity. In fact, most modern theories beyond the standard model predict baryon number violation by ordinary quantum field theoretic processes [136].

So let us here study a toy model for these processes. Suppose there is a heavy scalar particle $X$ which can mediate a transition between an proton and a positron, as well as between two positrons. Since the $X$-boson is described by a real field, it cannot carry any quantum numbers, and the transition evidently violates baryon conservation. The proton could then decay into a positron and an electron-positron pair. Let’s also assume that the coupling has the usual Yukawa form

$$g[\bar{\psi}_p \psi_e + X + \bar{\psi}_e \psi_p X],$$

where $g$ is a dimensionless coupling. If the mass of the $X$-boson $M_X$ is sufficiently large, baryon conservation will be a very good symmetry at the atomic energy scale, ensuring the stability of matter.

Now one can ask the question where the baryon violation takes place in the process of black hole formation and evaporation. A possible answer would be that it occurs when the freely falling proton encounters very large curvature as the singularity is approached. From the proton’s viewpoint, there is nothing that would cause it to decay before that. On the other hand, in the eyes of an outside observer, the proton encounters Planckian temperatures when it approaches the stretched horizon. Temperatures higher than $M_X$ can certainly excite the proton to decay.
So the external observer will conclude that baryon violation takes place at the horizon. Again, the freely falling and the outside observer viewpoint clearly are in conflict with each other.

However, the real proton propagating through spacetime is not the simple structureless bare proton. The Yukawa terms (5.45) cause it to make virtual transitions from the bare proton to a state with an $X$-boson and a positron. The complicated history of the proton is described by Feynman diagrams such as shown in figure 5.3. These diagrams make it clear that the real proton is a superposition of states with different baryon number. In the particular processes depicted in figure 5.3, the intermediate state has vanishing baryon number.

![Figure 5.3: Proton virtual fluctuations.](image)

There is nothing surprising about virtual baryon non-conservation. As long as $M_X$ is sufficiently large, the rate for real proton decay will be negligible, and the proton will be effectively stable. However, the probability for finding the proton in a configuration with vanishing baryon number is not small. This probability is closely related to the wave function renormalization of the proton and is of the order [9]

$$P \sim \frac{g^2}{4\pi} \log \frac{\mu}{M_X},$$

(5.46)

where $\mu$ is the cutoff in the field theory. For example, for $g \sim 1$, $\mu$ of the order of the Planck mass, and $M_X$ of the order $10^{16}$GeV, the probability that the proton has the ‘wrong’ baryon number is order unity. The transitions between baryon number states take place on a time scale of order $\delta t \sim M_X^{-1}$. So ordinary observations of the proton do not see these very rapid fluctuations. The quantity that is normally called baryon number is really the time averaged baryon number normalized to unity for the proton.

So by the arguments above, it not unlikely that when a proton passen the horizon, its instantaneous baryon number is zero. But a fluctuation that is much too rapid to be seen by a low energy observer falling with the proton appears to be a real proton decay lasting to eternity to an outside observer. This is of course a result from the time dilatation effect discussed in section 1.4.2. As the proton or any other system approaches the horizon, internal oscillations or fluctuations appear to slow down indefinitely so that a short lived virtual fluctuation becomes stretched out into a real process. This situation is depicted on figure 5.4. This explanation of baryon number violation to an outside observer is completely consistent with its perception of the stretched horizon as a hot membrane at Planckian temperatures.

An interesting question is now whether an observer falling with the proton can observe the baryon number just before crossing the horizon, and then send a message to the outside world
that the proton has not decayed. In order to make an observation while the proton is in a region of temperature $\leq M_X$, the observer must do so very quickly. In the proton’s frame, the time spent at the stretched horizon is $M_X^{-1}$. Thus, the uncertainty principle states the observer has to probe it with a quantum with an energy of order $M_X$. But such an interaction between the proton and the probe quantum is at high enough energy that it can cause a baryon number violating interaction. Thus, the observer cannot measure and report the absence of baryon number violation at the horizon without causing it himself.

5.5.3 Entangled spins

In section 5.2 we argued that at first sight, unitary black hole evaporation implies either a cloning of information or a mysterious bleaching of information. The latter was in conflict with the equivalence principle or with causality. And as we will show, the cloning of quantum states is in conflict with two foundations of quantum mechanics, namely the superposition principle and linearity. Suppose there exists some operator $D$ which has the following action

$$D|\psi\rangle = |\psi\rangle \otimes |\psi\rangle.$$  \hspace{1cm} (5.47)

Now assume that $|\psi\rangle$ is a superposition of two other states. For concreteness, take it to be the following state

$$|\psi\rangle = \frac{1}{\sqrt{2}}(|\uparrow\rangle - |\downarrow\rangle).$$  \hspace{1cm} (5.48)

Then linearity of quantum mechanics implies that acting with $D$ on $|\psi\rangle$ gives

$$D|\psi\rangle = \frac{1}{\sqrt{2}}(|\uparrow\rangle \otimes |\uparrow\rangle - |\downarrow\rangle \otimes |\downarrow\rangle).$$  \hspace{1cm} (5.49)
But this is clearly not equal to (5.47). So there appears to be no self-consistent definition of the operator $D$. Therefore, cloning of quantum states is not allowed.

However, the reasoning of section 5.5 states that the information to the infalling observer and the information to the outside observer are in fact two complementary versions of the same reality. Neither of these two observers will see a cloning of information.

The argument goes as follows. Consider a pair of particles that is prepared in a spin singlet. One member $a$ of the pair is sent into a black hole along with an apparatus $A$ which can measure the spin and send out signals. The other member $b$ remains outside. We assume that the energy associated with the apparatus is small compared to the black hole mass $M$ and that it is initially at rest outside the black hole.

Now the idea is the following. The outside observer waits a while after $a$ has been thrown into the black hole until the information about the spin of $a$ has been radiated away by the Hawking radiation. At that point, he can do a measurement on the radiation which is equivalent to a determination of any component of the original spin $a$. Meanwhile, the infalling spin $a$ has been measured by the apparatus $A$ which accompanied it. From the point of view of an external observer the 'spin in the Hawking radiation' $h$ must be maximally entangled with the member $b$ of the original pair which remained outside the black hole. If the spin $b$ is measured along any axis, then the Hawking spin $h$ must be found anti-aligned if it too is measured along the same axis. On the other hand, the original spin which fell through the horizon was also correlated to the other member of the pair $b$. It would seem that the two separate spins ($a$ and $h$) are maximally entangled with a third ($b$) so as to be anti-aligned with it. So we would need to have following evolution

$$\frac{1}{\sqrt{2}} (|\uparrow\rangle_a |\downarrow\rangle_b - |\downarrow\rangle_a |\uparrow\rangle_b) \rightarrow \frac{1}{\sqrt{2}} (|\uparrow\rangle_a |\downarrow\rangle_b - |\downarrow\rangle_a |\uparrow\rangle_b) \otimes \frac{1}{\sqrt{2}} (|\uparrow\rangle_b |\downarrow\rangle_h - |\downarrow\rangle_b |\uparrow\rangle_h),$$

which is not allowed by the arguments above.

There are a two important remarks to this reasoning. The first is of practical concern. For an outside observer to be able to find the information of the spin in the Hawking radiation, he would have to know the initial pure state of the matter that collapsed to form the black hole and the scattering matrix describing the unitary evolution of black hole evaporation. On top of that, as mentioned in the previous sections, the information in the Hawking radiation is very diffusely spread and comes out at a tremendously slow rate. So it should be clear that impossible in practice to find the information about the spin in the Hawking radiation. However, this gedanken experiment only adresses the question if it could be done in principle.

The second remark is of a more philosophical nature. It is well known by the principles of quantum mechanics that a measurement destroys the wave function. So to measure correlations, one needs to set up an ensemble of identical prepared systems. In the experiment above, this means one has to select a large number of identically prepared spin states $\{(a_1, b_1), (a_2, b_2), \ldots\}$. But then, an observer who first measures $b_1$ and subsequently jumps into the black hole to measure $a_1$ will not be able to get back out again and repeat the experiment. On the other hand, if one would take a large number of different observers and different black holes, they will never
be able to communicate the result of their measurement inside the horizon. So in this way, it cannot be checked if the anti-alignment of $b$ and $a$ is just a coincidence or a true correlation. The only way to check the correlation is if the outside observer first measures all the $b_i$ and then jumps in to check the $a_i$. So, how more certain the outside observer wants to be of the correlation between $a$ and $b$, the more measurements he has to make before jumping into the black hole and therefore the longer he has to wait before he can jump in. As we will see, this only favours the point we will make below.

To explain why $(5.50)$ is an invalid description of the situation, we address the question of how long the outside observer has to wait before jumping in so he is able to find the information about $a$ in the Hawking radiation. First, we consider the case where $a$ gets thrown into a young black hole, i.e. a black hole that has not yet reached the Page time. (The situation where the black hole is old is more complicated and will be discussed in the next sections.) In section 5.3 we saw that information starts to leak out when the black hole has evaporated half its initial entropy. And in section 4.6.3, we found that the time for a black hole to evaporate is of the order $\sim M^3$. Therefore, the time for a black hole to reach the Page time will also be of the order $\sim M^3$.

To do the further analysis, it is convenient to work in the Kruskal-Szekeres coordinates introduced in section 1.5. They are repeated here for convenience

$$U = -e^{\kappa(r^*-t)}$$
$$V = e^{\kappa(r^*+t)}.$$  (5.51)  (5.52)

where $r^*$ is the usual tortoise coordinate. It is evident that the value of $U$ where the outside observer runs into the singularity becomes very small if the observer delays for a long time before entering the black hole. This in turn constrains the time which the apparatus $A$ has available to emit its message. Let us choose the origin of the tortoise time coordinate such that the apparatus passes through the stretched horizon at $V = 1$. The observer will go through the stretched horizon after a period of order $M^3$ has passed in tortoise time, i.e. at $\log V \sim M^2$ since $\kappa \sim M^{-1}$. Recall from section 1.5 that the singularity is given by $UV = 1$. This implies that the message from $A$ must be sent before the apparatus reaches $U \sim \exp(-M^2)$. Near $V = 1$ this corresponds to a very short proper time $\tau \sim M^2 \exp(-M^2)$. The uncertainty principle then dictates that the message must be encoded into radiation with super-Planckian frequency $\omega \sim M^{-2} \exp(M^2)$. The backreaction on the geometry due to such a high energy pulse would be quite violent. It is apparent that the apparatus $A$ cannot physically communicate the result of its measurement to the observer in this experiment without running into unjustified extrapolation far beyond the Planck scale. The situation is depicted on figure 5.5.

Of course, the analysis above is not the full story. The thought experiment gives a flavor of how black hole complementarity works, but we have only considered the specific situation in which the black hole is young. In the next sections we will investigate the no-cloning experiment in more detail.
5.6 Old black holes as quantum mirrors

We saw in section 5.3 that a black hole starts to release its information after the Page time. Now we would like to refine our knowledge about information escape from black holes by asking how fast a certain amount of information of particular interest that gets thrown into a black hole comes back out in the Hawking radiation. Not only would we like to know this for young black holes, but also for old black holes. To get an idea of the information retention time, it is assumed that a black hole thermalizes information arbitrarily quick so that it is allowed to model the internal black hole dynamics by an instantaneous random unitary transformation. So we are taking the view of an outside observer who sees the black hole as a hot, radiating membrane. The analysis below was done in [137].

The quantum information that will be thrown into the black hole is stored in a $k$-qubit quantum memory. If a quantum memory stores $k$ qubits, this means that the stored quantum states live in a Hilbert space of dimension $2^k$. But actually, it also means something more: that the Hilbert space has a physically natural decomposition as a tensor product of $k$ two-level systems. For example, one might envision the memory as a system of $k$ spin-$\frac{1}{2}$ particles. However, this tensor product decomposition will not be central to the discussion below, so it will for the most part be adequate to regard the message system $M$ as a Hilbert space of dimension $|M| = 2^k$ without any special structure.

It is useful to imagine a reference system $N$ with dimension $|N| = |M|$ that is maximally
entangled with the message system $M$. That is, the initial joint state of the message and reference system may be written as

$$|\Psi\rangle^{MN} = \frac{1}{\sqrt{|M|}} \sum_{a=1}^{|M|} |a\rangle^M \otimes |a\rangle^N. \quad (5.53)$$

$N$ is said to provide a purification of the state of $M$. The density matrix for $N$ or $M$ separately is maximally mixed. If $M$ gets thrown into a black hole and after some time an outside observer finds a subsystem in the Hawking radiation that is maximally entangled with $N$, then one may say that the outside observer has recovered the quantum information that had been stored in $M$. This would imply in particular that if the initial state of $M$ had been the pure state $|\psi\rangle$, i.e. not entangled with any reference system, then the outside observer would be able to recover $|\psi\rangle$ in this chosen subsystem. So actually, the reference system is a tool to determine whether or not the information is recovered.

As already mentioned, we will consider the situation where $M$ gets tossed into an \textit{old} black hole, i.e. $|E| \geq |B|$, where $|E|$ and $|B|$ denote the dimension of the radiation and black hole subsystem respectively. Just after a black hole’s formation, it holds that $|E| \ll |B|$, and one can argue [138] that the radiation is nearly maximally entangled with a subsystem of the black hole. However, after the Page time $\ln |B|$ has decayed to less than half its initial value, so soon it holds that $|E| \gg |B|$. Then, we may expect that the black hole is nearly maximally entangled with a subsystem of the radiation. It should be noted that the analysis presented here tries to figure out how fast an outside observer can recover the information \textit{in principle}. This is because we will assume here that the outside observer has unlimited access to the information in the Hawking radiation so that by the reasoning above, the black hole is maximally entangled with a system that the outside observer controls. Of course, controlling the Hawking radiation is impossible \textit{in practice}. It comes out an immense slow rate, it is spread over a gigantic part of space and the correlations it contains are very subtle. So it is clear that only a super-civilization would be able to control it perfectly. Nevertheless, we only want to find out how nature works without worrying about the practical problems, so we will assume that the outside observer has unlimited control over the Hawking radiation.

The internal dynamics of the black hole are governed by deterministic unitary transformations that thoroughly mix the infalling information into the black hole’s preexisting $(n-k)$-qubit state. Then the black hole’s qubits are released, one by one, in the Hawking radiation. Now we would like to find out how many qubits it takes for a black hole to emit such that all the thrown-in information is returned to the outside observer.

Right after the information system $M$ has been tossed into the black hole, the $n$-qubit black hole system $B$ is maximally entangled with the system $NE$, where $E$ denotes the previously emitted ‘early’ Hawking radiation. Note that $B$ now contains $M$. The black hole continues to emit Hawking radiation. The number of qubits that have been emitted after $M$ has been thrown in is called $s$. The subsystem of $B$ that has been emitted by these $s$ qubits is called $R$. The black hole system containing $n-s$ qubits which remains after the emission of the $s$ qubits is called $B'$. We assume that the emitted subsystem $R$ of $B$ is chosen uniformly at random. That is, we imagine that $B$ is divided into two parts, one with $s$ qubits and the other with $n-s$
qubits. Then a unitary transformation $V$ chosen uniformly with respect to the Haar measure on $U(2^n)$ is applied to $B$. After that, the $s$-qubit system is identified as $R$.

As the Hawking radiation leaks out, the correlations between the evaporating black hole $B'$ and the reference system gradually weaken. Once $R$ is large enough, the surviving correlation of $N$ with $B'$ becomes negligible. At that point, since the overall state of $B'RNE$ is pure, the state of $N$ is very nearly purified by the radiation system $RE$ that Bob controls. The original information in the system $M$ has fallen into the hands of the outside observer. The complete situation is depicted in figure 5.6.

![Figure 5.6: The release of information thrown into an old black hole.](image)

Let $\rho^{BNE}$ denote the pure density matrix of the system $BNE$ at the point which the information has been thrown into the black hole. The reduced density matrix of the reference system and the black hole, i.e. the $BN$ system, is given by

$$\rho^{BN} = \text{tr}_E(\rho^{BNE}).$$

Then, the mixing by the black hole takes place, which is modeled by the unitary transformation $V$

$$\rho^{NB}(V) = (I^N \otimes V^B) \rho^{NB} \left( I^N \otimes V^B \right)^\dagger$$

After emission of the subsystem $R$, the reduced density operator on the remaining $NB'$ system is

$$\rho^{NB'}(V) = \text{tr}_R \left[ \rho^{NB}(V) \right].$$
Chapter 5. Black hole complementarity

The distance of $\rho_{NB}'$ from a product state, averaged over $V$ and hence over the choice of the subsystem $R$, can be bounded as

$$\int dV \|\rho_{NB}'(V) - \rho^{N}(V) \otimes \rho_{B}'_{\text{max}}\|^2 \leq \frac{|NB|}{|R|} \text{tr} \left( (\rho_{NB})^2 \right),$$

(5.57)

where $|NB|$ denotes the dimension of the Hilbert space of the $NB$ system. In the left hand side

$$\rho^{N}(V) = \text{tr}_B \left[ \rho_{NB}'(V) \right]$$

(5.58)

is the reduced density operator of $N$, and

$$\rho_{B}' = \frac{1}{|B'|} I_{B'}$$

(5.59)

is the maximally mixed density matrix on $B'$. The norm in (5.57) is defined by $\|A\| = \text{tr}\sqrt{A^\dagger A}$ and is an appropriate measure because two states that are close in this norm cannot be well distinguished by any measurement [140].

Because we are considering an old black hole, $B$ is maximally entangled with $NE$. So $\rho^{NB}$ is maximally mixed on a system of dimension $|E| = |N|/|B|$. (Recall that $B$ already contains the information system $M$ and that $|M| = |N|$.) So it holds that

$$\text{tr} \left[ (\rho_{NB})^2 \right] = \frac{|N|}{|B|},$$

(5.60)

Hence, (5.57) becomes

$$\int dV \|\rho_{NB}'(V) - \rho^{N}(V) \otimes \rho_{B}'_{\text{max}}\|^2 \leq \frac{|N|^2}{|R|^2} = \frac{22k}{22s} = \frac{1}{2^{2(s-k)}}.$$

(5.61)

So we see that if the number of emitted bits $s$ becomes bigger then the $k$ bits that were thrown in, the state of the $NB'$ system is nearly maximally mixed. The $k$ qubits that were thrown in have been ‘forgotten’ by the black hole and have been acquired by the outside observer.

Inconveniently, the information orginally encoded in the system $M$ has become encoded in a subsystem $M'$ of $RE$ that is very diffusely distributed among the emitted radiation quanta. But in principle, the outside observer could do a quantum computation that maps $M'$ to a compact system $\hat{M}$ localized in his laboratory. For any fixed value of the unitary transformation $V$, the outside observer’s decoding map can be chosen such that, after decoding, the density operator $\rho_{MN}$ is close to the maximally entangled state $|\Psi\rangle_{MN}$

$$F(V) \equiv \langle \Psi | \rho_{MN} | \Psi \rangle \geq 1 - \|\rho_{NB'}(V) - \rho^{N}(V) \otimes \rho_{B}'_{\text{max}}\|.$$  

(5.62)

Now (5.61) implies that, after averaging over $V$, the fidelity $F(V)$ deviates from one by no more than $2^{-(s-k)}$. So apart from a small error, the outside observer holds the purification of the reference system $N$ which, as explained above, means that he has recovered the information that originally was in the system $M$. 

The outside observer was able to extract $k$ qubits of high fidelity quantum information because of the pre-existing quantum entanglement that he shared with the black hole. Suppose on the other hand that the information system $M$ was thrown into a young black hole, such that $|E|/|B| \ll 1$. In that event, the previously emitted Hawking radiation $E$ will be nearly maximally entangled with a subsystem of $B$. The radiation will continue to be essentially informationless, revealing none of the information contained in $M$, until $|B'| = |NRE|$. Soon after, the black hole will be nearly maximally entangled with its surroundings and (5.61) (or more specifically, (5.60)) will begin to apply. At that point, the information contained in $M$ spills out. This is consistent with what was previously found in section 5.3. But the analysis here extends the one in section 5.3, since here we focused on when a fixed amount of quantum information of particular interest can be recovered while in section 5.3 we only considered the time-dependence of the quantum entanglement of the black hole with its surroundings.

So under the assumption that the outside observer has unlimited control over the Hawking radiation, the simple model of quantum black holes treated in this section leads to two main conclusions. First, if $k$ qubits are thrown into a black hole after the Page time, the information bounces right back. The outside observer has to wait not much longer than $k$ qubits to be evaporated back to obtain the original information with high fidelity. In other words, an old black hole behaves like a quantum mirror. On the other hand, if the $k$ qubits are thrown into a young black hole then the outside observer has to wait until the Page time is reached. At that point, the information pops out almost immediately.

This latter statement seems rather strange. Because who is it to say which $k$ qubits are the ones that were thrown in? In fact, no matter which $k$ qubits of quantum information swallowed by the black hole are of particular interest, these $k$ qubits are revealed almost right away when the Page time is reached. There is nothing special about the subsystem $M$ of $B$ that is maximally entangled with $N$. For any other $k$-qubit subsystem the conclusion would been the same, namely that $N$ becomes very nearly maximally entangled with a $k$-qubit subsystem of $RE$. Therefore, when a black hole that initially contained $n$ qubits has evaporated past the Page time, so that $(n + s)/2$ qubits have been emitted, the outside observer gets to decide which $k$ qubits of quantum information he will retrieve from the Hawking radiation. When he makes up his mind he performs the decoding operation on $RE$ that maps those $k$ qubits to the quantum memory in his laboratory. But the catch it that, although the outside observer can recover almost any $k$-qubit subsystem at this stage, he cannot recover more than $k$ qubits.

At the moment, the conclusions above seems to invalidate the principle of black hole complementarity. If we again consider the no-cloning thought experiment concerning entangled spins of the previous section, it is obvious that now there could occur quantum cloning by throwing one of the two entangled spins into an old black hole since it would just bounce right back. However, in this section we simply assumed that the mixing of information, modelled by the unitary transformation $V$, was instantaneous. A physical black hole will do the mixing or thermalization process in a finite amount of time. In the next section we investigate this thermalization process and see if it can save black hole complementarity.
5.7 Fast scrambling

When information gets thrown into a black hole, an outside observer will see it end up in the stretched horizon. There, it gets thermalized by the complex and ergodic behavior of that membrane. After the thermalization process, the information will be released in the Hawking radiation, ready to be detected by the outside observer. As was discussed in the previous section, this information release is very efficient when the black hole has become old. So in order to see if an outside observer could detect quantum cloning by throwing an entangled spin in an old black hole, we investigate how fast a black hole thermalizes information and what this tells us about its dynamics and the principle of black hole complementarity.

5.7.1 Scrambling in general quantum systems

Before directing our attention towards black holes, we first consider the general problem of how fast a quantum system can thermalize or scramble information [141]. To define the scrambling time, consider a complex chaotic system of many degrees of freedom, that has originally been prepared in some pure state. After a long time the system thermalizes although its quantum state remains pure. To see what is meant by this statement, consider the density matrix of a subsystem of \( m \ll N \) degrees of freedom, where \( N \) denotes the total number of degrees of freedom. It is well known that the small subsystem’s density matrix will tend toward thermal equilibrium with an average energy given by appropriately partitioning the original average energy of the big system. In other words, the entanglement entropy of the subsystem will approach the maximal value. In fact, the subsystem does not have to be small. The analysis of section 5.3 makes it very plausible that the subsystem will be extremely close to thermal for any \( m \) less than \( N/2 \). When this condition is achieved, i.e. when any subsystem smaller than half the whole system has maximum entanglement entropy, the system is called 'scrambled'. Intuitively this means that any information contained in the original state is mixed up so thoroughly that it can only be recovered by studying at least half the number of degrees of freedom.

Now let us start with a scrambled system and add a single degree of freedom in a pure state. Alternatively, we could perturb a small collection of degrees of freedom. The system will no longer be completely scrambled since one can recover information by looking at a single degree of freedom. But if one waits a little while, the bit of added information will eventually diffuse over all the degrees of freedom and the system will return to a scrambled state. The time needed to re-scramble when a bit is added is defined to be the scrambling time. We will denote the scrambling time by \( t_s \). (Actually, the scrambling time defined in this way is not completely precise since one needs to specify a precision in how close the subsystem’s entropies are to the maximal as was done in (5.57). Here, however, this complication will be ignored.)

The quantum systems that will be looked at here are supposed to have interactions that are between bounded clusters of degrees of freedom. Pairwise interactions would be an example. So if the system is described by a conventional Hamiltonian \( H \), then \( H \) consists of terms, each of which involves clusters of a fixed, finite amount of degrees of freedom \( l \). The total number of degrees of freedom scales with a parameter \( N \) and they may either be commuting or anti-commuting. Now for such a system, what is the smallest that the scrambling time can be?
Suppose that the degrees of freedom are arranged in a $d$ dimensional periodic array so that each degree of freedom interacts with only a few near neighbors. The linear dimension of the system is proportional to $N^{1/d}$. In this case the time for a signal to propagate from a single cluster to the most distant cluster obviously grows with $N$ at a rate that satisfies

$$t_* \leq c N^{1/d},$$

(5.63)

where $c$ is a coefficient that does not depend on $N$.

In many examples the effective rate of interaction is temperature dependent. Thus the co-efficient $c$ depends on $\beta$. A convenient parameterization is

$$\tau \equiv \frac{t_*}{\beta} \leq C(\beta) N^{1/d},$$

(5.64)

where $C$ is dimensionless.

In most known examples, thermalization is a process of diffusion in which the initial perturbation spreads in space to a distance of order $\sqrt{t}$. In that case the bound becomes

$$\tau \equiv \frac{t_*}{\beta} \leq C(\beta) N^{2/d}.$$  

(5.65)

Now let us eliminate the restrictions implied by the finite dimensionality. In other words, we allow arbitrary interactions between any degrees of freedom as long as the individual interaction terms involve no more than $l$ of them. Roughly speaking, we are going to the limit of infinite dimension. The 'Fast scrambling conjecture' is then that (5.65) is replaced by

$$\tau \leq C(\beta) \log N.$$  

(5.66)

Systems that saturate the bound (5.65) or (5.66) are called 'fast scramblers'.

An indication for the validity of the fast scrambling conjecture comes from quantum circuits. The simplest quantum circuit involving $N$ qubits is constructed as follows. Time is divided into intervals and in each interval a pair of qubits are selected at random and allowed to 'scatter' by means of a randomly chosen $U(4)$ operator. The number of timesteps is called the depth of the circuit. The circuit acts on any input state of the $N$ qubits and unitarily transforms it to an output state. It is known that this system scrambles in a number of steps that increases with $N$ like $N \log N$.

But faster scrambling can be achieved by a 'parallel processing' in which multiple disjoint pairs are allowed to interact simultaneously. The time between steps will be called $\beta$ since it will roughly correspond to the inverse temperature in Hamiltonian systems. In the example we will consider here, every qubit interacts one in each timestep. Every step begins by randomly pairing the qubits into $N/2$ pairs. Any qubit may pair with any other qubit, but none interact with more than one other. Next, we pick $N/2$ random $U(4)$ matrices and allow the qubit pairs to scatter. As before, the total number of $U(4)$ operations required to scramble the system is
Chapter 5. Black hole complementarity

\(N \log N\), but now the parallel processing assembles them into only \(\log N\) timesteps, taking a total time \(t_* = \beta \log N\). So in the notation used before

\[
\tau = \frac{t_*}{\beta} = C \log N,
\]

with \(C\) being independent of \(\beta\) in this case.

As mentioned above, the precise definition of scrambling is technical. A simple definition in the qubit model is that the final state has been randomized with respect to the Haar measure over the entire \(2^N\) dimensional Hilbert space. But such randomization is known to be inefficient, it requires a non-polynomial number of timesteps. However, here we rely on a weaker definition of scrambling that requires only quadratic functions of the density matrix elements to completely randomize, i.e. approach their Haar-scrambled values. With that definition, scrambling takes place on a time scale of order \(\log N\) and not smaller. The result also does not depend on the assumption of two-body interactions. As long as the number of qubits in the elementary operations is finite, the minimum scrambling time grows like \(\log N\).

The logarithmic growth of \(t_*\) can be understood as follows. Suppose the state of the first qubit is fixed in some manner. Then after one timestep that qubit has influenced two qubits, namely itself and the one that it interacted with. After \(n\) timesteps the first qubit has influenced \(2^n\) qubits. Obviously the system is not completely scrambled until that first qubit has influenced all the others. Thus the scrambling time cannot be smaller than order \(\log N\). That the quantum circuit above saturates this bound shows how efficient a scrambler it is. The validness of the fast scrambling conjecture has also been supported by the proof of a logarithmic lower bound on the scrambling time for systems with finite norm terms in their Hamiltonian [142]. The bound holds in spite of any nonlocal structure in the Hamiltonian, which might permit every degree of freedom to interact directly with every other one.

An interesting question concerning the relation between the discrete models and the continuous Hamiltonian evolution, is what time scale in the latter corresponds to a single step in the discrete theory. The answer obviously depends on the state of the system. Increasing the energy or temperature will speed things up. Therefore, a good guess is that the discrete timesteps should be identified with time intervals of order

\[
\delta t \sim \frac{1}{\epsilon},
\]

where \(\epsilon\) is the energy per degree of freedom. In many cases it is proportional to the temperature. This time scale is the time interval during which every degree of freedom interacts about once. For that reason it is identified with the discrete timesteps in the parallel processing circuit.

There is another definition of scrambling that is suggested by the analysis of section 5.3. Consider any subsystem of \(k\) qubits with \(k < N/2\). In section 5.3 it was shown that the entanglement entropy on the subsystem is close to maximal in a Haar-scrambled state. In fact, the entropy differs from maximal by less than a single bit even if the subsystem is just a little smaller than
N/2. We found in section 5.3 that

\[ S_k = \ln(2^k) - \frac{2^k}{2^{2N-k}} \approx k - O(e^{2k-N}). \]  

Any state that satisfies (5.69) will be called Page-scrambled. So Haar-scrambled implies Page-scrambled, but the converse is not true, i.e. Page-scrambled does not imply Haar-scrambled. In particular, the scrambler described above is sufficient to Page-scramble despite the fact that it only takes \( N \log N \) operations.

### 5.7.2 Scrambling in black holes

We now turn our attention back to black holes and we would like to know how long it takes for a bit of information to diffuse over the entire horizon. The simplest situation is a localized perturbation created on the stretched horizon, thereby disturbing the thermal equilibrium. The perturbation then spreads out until it uniformly covers the horizon. Although there is no mathematical proof, it seems reasonable to identify that time with the scrambling time.

One could drop a mass into the black hole and watch the energy and temperature spread out on the stretched horizon. But, in section 3.3, we calculated how fast the charge density equilibrated after we dropped a point charge in the black hole

\[ t_* = 2GM \log \left( \frac{2GM}{\rho_0} \right) \approx \frac{1}{\kappa} \log \left( \frac{2GM}{\rho_0} \right) \sim \beta \log \left( \frac{2GM}{\rho_0} \right), \]  

where \( \rho_0 \) is the thickness of the stretched horizon. This can of course also be used as the scrambling time. If we assume that \( \rho_0 \) is of the order of the Planck length, then we can write

\[ t_* \sim \beta \log S, \]  

since the black hole entropy \( S = c^3 \pi R_s^2 / Gh \) is the square of \( R_s / l_p \). So if we think of the entropy of the black hole as the number of its degrees of freedom then \( \tau = C \log S \) shows that black holes are fast scramblers.

In a sense the fast scrambling property of black holes is the quantum mechanical analogon of the classical no hair conjecture. In classical theory, the mass contracts beyond its Schwarzschild radius after which it will settle down to a Kerr-Newman black hole. Once this stationary regime is reached, the only information left about the collapsed matter is its mass, angular momentum and charge. So during the time the black hole is evolving towards its equilibrium state, information gets lost. In black hole complementarity, an outside observer will see the matter contracting into the stretched horizon. Because of the fast scrambling property of that membrane all the quantum information of the collapsed matter will be spread and hidden across the entire
horizon area. The chaotic dynamics make it very hard to recover the information. So where black holes destroy information in the classical theory, they effectively hide it in quantum theory.

Based on the observations of the previous section, it is surprising that a real physical system can scramble that fast. One might argue that as the number of degrees of freedom increases they have to spread out in space, either along a line, a plane, or in a space-filling way. One can imagine connecting distant degrees of freedom by wires and simulating non-locality, or a higher dimensional system, but eventually the wires will get so dense that there will not be room for more. The fastest scramblers in three spatial dimensions would have a scrambling time of order $N^{2/3}$. This seems likely to be the case for anything made of ordinary matter.

But that intuition is wrong when gravity is involved: gravity brings something entirely new into the game, something that looks so non-local that black holes effectively are infinite dimensional. They are the fastest scramblers in nature by a wide margin.

This observation gives us a condition that must be satisfied by the dynamics of the microphysical degrees of freedom on the stretched horizon. It therefore provides us with a hint of what exactly is going on in this thin boundary layer. Another observation made in [141] is that matrix quantum mechanics (M theory) satisfies the bound (5.66). This means that string theory could possibly account for the fast scrambling behavior of the stretched horizon. The authors also strengthen this possibility with arguments from D0-brane black holes and Ads/CFT.

### 5.7.3 The entangled spin experiment revisited

Let us now see if black hole complementarity survives when we use the scrambling time $t_s \sim R_s \log(R_s/l_p)$ as the information retention time. Or in Planck units

$$\frac{t_s}{l_p} \sim \frac{R_s}{l_p} \log \left( \frac{R_s}{l_p} \right) \rightarrow t_s \sim R_s \log R_s. \quad (5.72)$$

We consider again the entangled spin experiment of section 5.6.3.

The outside observer crosses the horizon at $V_o$. He then reaches the singularity at $U \leq V_o^{-1}$. The freely falling apparatus $A$ has a proper time $\tau$ between crossing the horizon at $V = V_A$ and reaching $U = V_o^{-1}$ that is given by [137]

$$\tau = CR_s \frac{V_A}{V_o}, \quad (5.73)$$

where $C$ is a numerical constant that depends on the apparatus’s initial data. $C = e^{-1}$ if the apparatus falls from rest starting at infinity. In terms of the Schwarzschild time, the outside observer’s fall into the black hole is delayed relative to the one of the apparatus by $\Delta t$, where $V_o/V_A = \exp(\Delta t/2R_s)$. Therefore

$$\tau = CR_s e^{-\Delta t/2R_s}. \quad (5.74)$$
Thus the apparatus’s proper time is of order the Planck time or shorter if

\[ 1 \leq CR_s e^{-\Delta t/2R_s}, \]

in Planck units. This gives

\[ \Delta t \geq R_s \log R_s. \]

which is equal to the scrambling time. So it follows that complementarity is only just compatible with black holes as fast scramblers.

We can conclude that the fact that black holes are fast scramblers is not just an interesting curiosity. The principle of black hole complementarity requires that no observer be able to detect cloning of quantum information. This places a bound on how fast an outside observer can retrieve information that was thrown into a black hole. At first, the situation in section 5.6.3 was satisfied by a huge ’overkill’. But complementarity would have been more compelling if it had just barely escaped inconsistency. A good example is the Heisenberg microscope experiment which not only showed that the uncertainty principle could not be violated, but that it could be saturated.

So the experiment of throwing information in an old black hole gives, by the reasons of the previous and this section, a very gratifying situation: the retrieval time roughly saturates the complementarity bound derived from un-observability of quantum cloning. This conclusion greatly favors the principle of black hole complementarity. It indicates that we are not looking at just a trivial fact but really at a fundamental principle of nature.

### 5.8 Complementarity in the semiclassical framework

In a series of papers [143–147] it was argued that the idea of complementarity also is present in the semiclassical framework of black hole evaporation. All that is needed to expose it is an incorporation of backreaction in the derivation of the Hawking radiation. The resulting formalism is related to the stretched horizon concept by the ’magnifying glass mechanism’ mentioned in section 5.4.

Normally, if one draws a Cauchy surface in a spacetime diagram like that in figure 5.7, one expects that all operators on this surface which are space-like separated commute with each other. This assumption in fact was essential to the original derivation of the Hawking radiation in section 2.3.1. It also leads to the non-unitary evolution of the previous chapter since it is one of the main foundations to argue that the asymptotic Hilbert space of out-modes is incomplete.

Here, it will be argued that this reasoning becomes incorrect when backreaction effects are taken into account. The result will imply a drastic revision of the standard semiclassical picture of the evaporation process.

In section 2.3.1 we assumed that the incoming particles described by \( \psi_{in}(v) \) with \( v > v_0 \) and the outgoing particles described by \( \psi_{out}(u) \) form independent sectors of the Hilbert space, and that the corresponding field operators commute with each other. The underlying classical intuition is that the fields \( \psi_{out}(u) \) will propagate into the region behind the black hole horizon and thus
Figure 5.7: A Cauchy surface with in and out modes.

become unobservable from the outside. However, this intuition ignores the important fact that the infalling particles in fact do interact with the outgoing radiation because they slightly change the black hole geometry. In the spherically symmetric case of an infalling $s$-wave particle, this change in the geometry is represented by a small shift in the black hole mass $M$ and the time $v_0$ at which the black hole horizon was formed. Note that we only consider $s$-wave particles because of the arguments in section 2.4.

Assume that a spherical shell of matter with energy $\delta M$ falls into a black hole at some later time $v_1 > v_0$. The Schwarzschild radius will then increase slightly with an amount $2G\delta M$, and the time of the formation of the horizon $v_0$ will also change very slightly by

\[
\delta v_0 = -4c\delta Me^{-\left(\frac{v_1 - v_0}{4GM}\right)}.
\] (5.77)

At first it seems reasonable to ignore this effect as long as the change $\delta M$ is much smaller than $M$. However, this will appear not to be the case. The exponential $v$-dependence that occurs in (5.77) is typical of black holes and has to do with the diverging redshift. This time it helped in our favour because it exponentially suppressed the effect on $u_0$ of the ingoing matter. But in other physical quantities it is easy to get exponentially growing factors that enhance physical effects that seemed to be unimportant at first. This will be the ‘magnifying glass effect’, exposing some Planck-scale physics to a distant observer. For example, the variation in $u_0$, although very small, has an enormous effect on the wavefunction $\psi_{\text{out}}(u)$ of an outgoing particle.

For large $u$, the reparametrization $u(v)$ takes the asymptotic form [147]

\[
u(v) = v - 4GM \ln \left(\frac{v_0 - v}{4GM}\right).
\] (5.78)
With this one can write the relation between the \textit{in} and \textit{out} fields as

\[ \psi_{\text{in}}(v) = \psi_{\text{out}}(u(v)) = \psi_{\text{out}} \left( v - 4GM \ln \left( \frac{v_0 - u}{4GM} \right) \right). \] (5.79)

Now using (5.77) one can verify that as a result of the infalling shell, the outgoing particle-wave is delayed by an amount that grows rapidly as a function of \( v \)

\[ \psi_{\text{out}}(u) \rightarrow \psi_{\text{out}} \left( v - 4GM \ln \left( \frac{v_0 - u}{4GM} \right) \right) \]

\[ = \psi_{\text{out}} \left( u - 4GM \ln \left( 1 - 4e \frac{\delta M}{4GM} e^{-\left( u - v_0 + 4GM \ln(\frac{4GM}{v_0 - v}) \right)/4GM} \right) \right). \] (5.80)

Notice that even for a very small perturbation \( \delta M \) the argument of the field \( \psi_{\text{out}} \) goes to infinity after a finite time \( u_{\text{lim}} - u_1 \sim -4GM \ln(\delta M/M) \). The physical interpretation of this fact is that a matter-particle that is on its way to reach the asymptotic observer at some time \( u > u_{\text{lim}} \) will, as a result of the additional infalling shell, get trapped inside the black hole horizon.

The arguments above imply that the asymptotic wave function of an \textit{individual} particle is very sensitive to the gravitational backreaction. To see what this means for the \textit{collective} state of the outgoing radiation is clearly a much more subtle matter. For example, the transformation (5.80) can be a symmetry of the Hawking state. Approximately, this indeed appears to be the case [145]. This implies that the thermality of the Hawking radiation will approximately survive the inclusion of backreaction. However, the fact that the gravitational backreaction is important for individual particles is sufficient to substantially change the usual semiclassical picture.

To take the effect of (5.80) into account, let us divide up the infalling matter in a classical piece plus a small quantum part that is described in terms of the quantum field \( \psi_{\text{in}}(v) \). The classical piece obviously represents the matter that collapsed to form the black hole. As a counter-intuitive consequence, the parameter \( v_0 \) is now not just a classical number but should be treated as a quantum operator. More explicitly, \( v_0 \) can be written as

\[ v_0 = v_0^c - 4e \int_{v_0^d}^{\infty} dv e^{(v_0^d - v)/4GM} T_{\text{in}}(v) = v_0^c + \delta v_0, \] (5.81)

where \( T_{\text{in}}(v) \) denotes the energy-momentum tensor of \( \psi_{\text{in}}(v) \) with support on \( v > v_0^c \). The classical part \( v_0^c \) is determined by the collapsed matter.

The goal is now to calculate the algebra of the outgoing field \( \psi_{\text{out}}(u) \) for late times with the incoming field \( \psi_{\text{in}}(v) \) for \( v > v_0^c \). First one finds from (5.81) that

\[ [\delta v_0, \psi_{\text{in}}(v)] = -i4e e^{(v_0^d - v)/4GM} \partial_v \psi_{\text{in}}(v), \] (5.82)

where it was used that the energy-momentum tensor generates coordinate transformations.
Now the relation
\[ \psi_{\text{out}}(u) = \psi_{\text{in}}(v(u) + \delta v_0), \] (5.83)
where
\[ v(u) = v_0^{cl} - 4GM e^{(v_0^{cl} - u)/4GM} \] (5.84)
is the inverted form of (5.78), can be written as
\[ \psi_{\text{out}}(u) = \exp(-e^{(u - v_0^{cl})/4GM} \delta v_0 \partial_u) \psi_{\text{in}}(v(u)) \]
(5.85)
\[ = \psi_{\text{in}}(v(u)) - e^{(u - v_0^{cl})/4GM} \delta v_0 \partial_u \psi_{\text{in}}(v_0^{cl} - 4GM e^{(v_0^{cl} - u)/4GM}) \partial_u \psi_{\text{in}}(v) + \ldots \]
\[ = \psi_{\text{in}}(v(u)) - \delta v_0 \partial_u \psi_{\text{in}}(v(u)) + \ldots \]
\[ = \psi_{\text{in}}(v(u) + \delta v_0). \]

Actually, since \( \delta v_0 \) is an operator-valued quantity this could in principle introduce a problem with normal ordering at higher orders in the expansion of the exponential. In first instance, however, this point will be ignored and the linearized interaction between the in and out modes will simply be exponentiated. This procedure amounts to the ladder approximation to linearized gravity, which, in the kinematical regime of interest, is known to provide the correct leading order result [148].

With the results above one can compute the exchange algebra between the in and out-fields. One finds by using (5.85) and (5.82)
\[ \psi_{\text{out}}(u) \psi_{\text{in}}(v) = \exp(-e^{(u - v_0^{cl})/4GM} \delta v_0 \partial_u) \psi_{\text{in}}(v(u)) \exp(-e^{(u - v_0^{cl})/4GM} \delta v_0 \partial_u) \psi_{\text{out}}(u) \]
(5.86)
\[ = \exp(i4e^{(v-u)/4GM} \partial_u \partial_v) \psi_{\text{in}}(v) \psi_{\text{out}}(u), \] (5.87)
which is valid for \( v > v_0^{cl} \). This exchange algebra is the quantum implementation of the gravitational backreaction (5.80) and can be seen to be highly non-local.

It should be noted that to derive this result no use was made of any assumption other than those already made in the usual derivation of the Hawking radiation. The only difference compared to section 2.3.1 is that now the seemingly negligible quantum contribution from \( \psi_{\text{in}}(v) \) to \( v_0^{cl} \) is taken into account.

The found commutators grow exponentially in time. This implies that the standard semiclassical picture of the black hole evaporation process needs to be revised drastically. In particular, it tells us that, due to the quantum uncertainty principle, we should be very careful in making simultaneous statements about the infalling and outgoing fields. Mathematically, the Hilbert space of the scalar fields on a Cauchy surface as depicted on figure 5.7 does not decompose into a simple tensor product of a Hilbert space inside the black hole and one outside. Instead, in view of the exponentially non-local nature of the commutator between the in and out fields, it is clear that the out Hilbert space is not even approximately independent of the Hilbert space of the infalling matter. This result supports the physical picture that there is a certain complementarity between the physical realities as seen by an asymptotic observer and by an infalling observer.
So although the principle of black hole complementarity was introduced as being founded on the existence of a thin Planck-scale membrane, it does have some roots in the semiclassical framework. The derivation above obviously does not prove the validness of black hole complementarity, or neither does it provide us with a detailed mechanism of how it should work. Nevertheless, it indicates that black hole complementarity is an essential feature of quantum black holes.

At this point the main features of black hole complementarity are presented. This was done in 5 precise postulates, which make black hole complementarity a concrete statement rather than just some vague idea. After that, the consequences were investigated via thought experiments. The idea of black hole complementarity was strengthened via results from quantum information theory and the semiclassical framework. Although its validity can only be confirmed with certainty once we have a satisfactory quantum theory of gravity, it is a very promising principle since it ties together most of the loose ends about quantum black holes. However, in the next chapter we will discuss a loophole in the black hole complementarity picture that could possibly invalidate the complete principle.
Chapter 6

The Firewall

"The world we have created is a product of our thinking; it cannot be changed without changing our thinking."
- A. Einstein (1908)

Throughout the previous chapters, a long way has been travelled to reconcile quantum theory with black holes. Up to chapter 4, everything went well with the quantum mechanical confirmation of thermodynamical aspects of horizons. However, it then became clear that unitarity was endangered by the process of black hole formation and evaporation. This appeared to have catastrophic consequences for effective quantum field theory. The alternatives didn’t provide any less alarming solution. Ultimately, this convinced people to keep unitarity in quantum gravity.

However, implying unitarity to the microscopic degrees of freedom of a black hole seemed to result in the cloning of arbitrary quantum states. Resolving this issue lead to the principle of black hole complementarity which provided us with a phenomenological description of how unitary quantum black holes must behave.

In this chapter however, a possible loophole in the complementarity picture will be investigated. As explained below, black hole complementarity is threatened by a firewall. This firewall is the reason for the word ‘persistent’ in the title of this thesis. In a more general perspective, the information paradox can be seen as the difficulty to reconcile black holes with unitarity. In this point of view, today the information paradox is alive more than ever.

6.1 The AMPS argument

From the analysis in sections 5.3 and 5.6 of the previous chapter we know that if black hole evaporation is unitary, the black hole is maximally entangled with the Hawking radiation once it has evaporated half of its initial entropy. From that point on, information starts to leak out of the black hole under the form of correlations between the newly emitted Hawking quanta and the earlier emitted radiation. So the Hawking quanta emitted by old black holes are entangled with the previously emitted radiation.
On the other hand, the equivalence principle requires entanglement between modes on different sides on the horizon. This can be understood by first looking at the Unruh effect. There, the Minkowski vacuum resulted into entangled modes in the left and right Rindler wedges for accelerating observers. Here we will use the reverse argument: in order to be in the Minkowski vacuum state, one needs entangled modes in Rindler spacetime. Now the equivalence principle dictates that the the freefalling coordinate frame in the black hole spacetimes is Minkowski. Based on the observations on the Unruh effect, one expects that in order to be in the freely falling Minkowski vacuum state, there should be entangled modes on both sides of the horizon. This is confirmed by the explicit derivation of the state (4.81). That this state truely is the Minkowski vacuum follows from the derivation of the Hawking radiation in section 2.3.1. There, the modes at asymptotic late times were related to modes in the asymptotic past where the matter that will collapse to form the black hole was still very, very diffusely spread so that spacetime was flat and thus Minkowski. So to conclude, if this entanglement between the modes on different sides of the horizon were not present, the field would not be in the freely falling vacuum state and an infalling observer would detect particles. This is completely analogous to the observation that without the entangled modes between left and right Rindler wedges, one would not have the Minkowsksi vacuum.

The equivalence principle and unitarity are believed to be two foundations of quantum black holes. The AMPS argument however, named after its discoverers Almheiri, Marolf, Polchinski and Sully, states that they are inconsistent and cannot be combined within the framework of black hole complementarity [149]. If this indeed would be the case, then complementarity would be completely ruled out. The AMPS argument is based on two observations in the semiclassical picture which will be presented in the subsections below.

6.1.1 The entropy argument

Consider the black hole evaporation process and assume that it has reached the Page time. This means that the black hole is maximally entangled with the Hawking radiation $R$ that has already been emitted up to that point. We will call $R$ the early radiation. Call the next Hawking quanta that gets emitted $O$ and its interior partnermode $I$. Strong subadditivity of the entanglement entropy in the ROI system gives

\[ S_{RO} + S_{OI} \geq S_O + S_{ROI}. \]  \hfill (6.1)

Unitary evolution implies that after the Page time, the entanglement entropy of the black hole has to decrease. Because the total system is in a pure state, the entanglement entropy of the Hawking radiation is equal to that of the black hole at all times. This implies that the entanglement entropy of the radiation before the emission of the $O$ quantum has to be bigger than afterwards.

\[ S_{RO} < S_R \]  \hfill (6.2)

Now for an infalling observer to experience the vacuum, maximal entanglement between the outgoing quantum $O$ and its interior partnermode $I$ is required. So $I$ purifies the state of $O$

\[ S_{IO} = 0. \]  \hfill (6.3)
Because the IO system is in a pure state it follows that

\[ S_{RIO} = S_R. \] (6.4)

Now using (6.2) and (6.4), equation (6.1) becomes

\[ S_R \geq S_O + S_R, \] (6.5)

which clearly is a contradiction because \( O \) by itself is definitely not in a pure state so \( S_O \neq 0 \).

To summarize, for an infalling observer to experience the horizon as harmless the outgoing mode has to be maximally entangled with an interior partner mode. On the other hand, the entanglement entropy of the black hole has to decrease after the Page time in order to have unitary evolution. This can only be done if the outgoing mode is entangled with the early emitted radiation. The analysis above shows that these two types of entanglement the outgoing mode needs to have are not compatible.

### 6.1.2 The projection argument

For a second argument we again select a certain point of time in the evaporation process later than the Page time. The radiation that has been emitted before that point is called the early radiation, the radiation emitted after that point is the late radiation. Because black hole evaporation is unitary by postulate 1, the final state of the Hawking radiation after the black hole has disappeared completely is pure, again assuming that the collapsed matter was in a pure state. So we can write it as

\[ |\Psi\rangle = \sum_i |\psi_i\rangle_E \otimes |i\rangle_L, \] (6.6)

where \( |i\rangle_L \) is a complete, orthonormal basis for the late radiation. It is crucial to realize that the division between early and late radiation was done after the Page time so that the dimension of the late Hilbert space is much smaller than the early Hilbert space. Therefore, the number of basis states of the late radiation \( |i\rangle_L \) is very small compared to the number of basis states of the early radiation. This implies that the states \( |\psi_i\rangle \) can definitely not be a complete and orthonormal basis for the late radiation.

We will now show that we can construct operators, acting on the early radiation, whose action on \( |\Psi\rangle \) is equal to that of a projection operator onto any given subspace of the late radiation. Because the stretched horizon is a chaotic system, the state of the Hawking radiation is assumed to be effectively random within its Hilbert space. We also assume, just as in section 5.6 that the observer knows the initial state of the matter that collapsed to form the black hole and also the black hole S-matrix.

Consider the projection operator onto the state \( |i\rangle_L \) in some orthonormal basis for the late radiation

\[ P^i = |i\rangle \langle i|_L. \] (6.7)
This projection operator represents a measurement of the state $|i\rangle_L$ of the late radiation. Also introduce the operator

$$\hat{P}^i = L|\psi_i\rangle_E \langle \psi_i|_E ,$$

which represents a measurement of the state $|\psi_i\rangle_E$ of the early radiation. Here, $E$ and $L$ represent the dimensions of the early and late radiation Hilbert spaces. It will now be shown that this measurement of the early radiation will allow one to anticipate the measurement of the late radiation. That is

$$\hat{P}^i |\Psi\rangle \approx P^i |\Psi\rangle = |\psi_i\rangle_E \otimes |i\rangle_L .$$

If the $|\psi_i\rangle_E$ were an orthonormal basis, this would be an equality. However, from the analysis below it will appear to be an approximate equality when $L \gg E$.

The relative error between $P^i |\Psi\rangle$ and $\hat{P}^i |\Psi\rangle$ is

$$\epsilon = \frac{|(P^i - \hat{P}^i)| |\Psi\rangle|^2}{|P^i |\Psi\rangle|^2}$$

$$= \frac{1}{\langle \psi_i|\psi_i\rangle_E} \left( \langle \psi_i|\psi_i\rangle_E - 2L\langle \psi_i|\psi_i\rangle_E^2 + L^2 \sum_j \langle \psi_i|\psi_j\rangle_E \langle \psi_j|\psi_i\rangle_E \right)$$

$$= (1 - L\langle \psi_i|\psi_i\rangle_E)^2 + L^2 \sum_{j \neq i} \langle \psi_i|\psi_j\rangle_E \langle \psi_j|\psi_i\rangle_E^2$$

Now expand the states of the early radiation in an orthonormal basis

$$|\psi_i\rangle_E = \sum_{a=1}^E c_{ia} |a\rangle_E .$$

Then the average over the Hawking state $|\Psi\rangle$ with the uniform measure, as in the microcanonical ensemble, gives

$$\overline{c_{ia} c^*_{jb}} = \frac{1}{LE} \delta_{ij} \delta_{ab}$$

$$\overline{c_{ia} c^*_{jb} c_{kc} c^*_{ld}} = \frac{1}{L^2 E^2} (\delta_{ij} \delta_{kl} \delta_{ab} \delta_{cd} + \delta_{il} \delta_{jk} \delta_{ad} \delta_{bc}) .$$

So it follows that

$$\overline{\langle \psi_i| \psi_j\rangle_E} = \frac{1}{L} \delta_{ij}$$

$$\overline{\langle \psi_i| \psi_j\rangle_E \langle \psi_k| \psi_l\rangle_E} = \frac{1}{L^2} \delta_{ij} \delta_{kl} + \frac{1}{L^2 E} \delta_{il} \delta_{jk} .$$

Then, for $E \gg L \gg 1$, one finds for the averaged relative error

$$\bar{\epsilon} = L^2 \sum_{j \neq i} \left( \frac{1}{L^2} \delta_{ij} \delta_{ij} + \frac{1}{L^2 E} \delta_{il} \delta_{jj} \right)$$

$$= L^2 \frac{L}{L^2 E}$$

$$= \frac{L}{E} ,$$

(6.16)
which decreases exponentially after the Page time. While the calculations above refer to projection onto a one-dimensional space, (6.16) also holds for more general projections given by sums of the $\hat{P}_i$.

Now consider an outgoing Hawking mode at infinity in the later part of the radiation. We take this mode to be a localized wave packet with width or order $R_s$, corresponding to a superposition of frequencies $O\left(R_s^{-1}\right)$. Postulate 2, which states the validity of effective quantum field theory outside the stretched horizon, then implies that one can assign a unique observer-independent creation operator $b^\dagger$ to this mode. Now we can take the basis $|i\rangle_L$ of the analysis above to be the eigenstates of the number operator $N_b = b^\dagger b$. This means that an observer making measurements on the early radiation can know the number of Hawking quanta that will be present in a given mode of the late radiation.

Next consider an infalling observer and his associated set of modes with creation operators $a^\dagger$. The vacuum state for this observer, which will guarantee him a safe passage through the horizon, is defined by $a|0\rangle = 0$. But now recall from the derivation of the Hawking radiation in section 2.3.1 that the two sets of operators $(a, a^\dagger)$ and $(b, b^\dagger)$ are related by a Bogoliubov transformation. It is therefore impossible for the state $|\Psi\rangle$ to be both a $N_b$ eigenstate and an $a$-vacuum.

So we come across a contradiction. The almighty outside observer knows the initial state of the collapsed matter and he can simply act on it with the known black hole $S$-matrix. This allows him to know the state (6.6) the radiation will have after the black hole has evaporated completely. When the black hole is old he can measure the early radiation which leads him to (6.35). Combining his measurement results with the knowledge of the total radiation state he therefore knows with very high precision how many Hawking quanta in the mode associated with $b^\dagger$ are yet to come. This is equivalent to stating that the late radiation is in an eigenstate of $N_b$. But this implies that $a|\Psi\rangle \neq 0$. So an infalling observer will not experience the vacuum but encounters high energy quanta. That these quanta have a destructively high energy can be seen by tracing back a typical Hawking quantum to just outside the horizon where it will be exponentially blue-shifted.

Note that the infalling observer need not have actually made the measurement on the early radiation. To guarantee the presence of high energy quanta it is enough that it is possible, just as shining light on a two-slit experiment destroys the fringed even if we do not observer the scattered light. The line of reasoning used in the analysis above is very similar to the one when scattering two electrons. Assume the initial state (momentum) of the two particles is known. One then works on this state with the $S$-matrix which can be calculated from the underlying theory, QED in this case. Because energy-momentum conservation is contained in the $S$-matrix, the final calculated state will be a superposition of all possible outcomes, i.e. all momentum combinations for the two electrons that add up to the total initial momentum. So if one measures the momentum of one of the electrons, the other one is known automatically.

There are two explanations for the name 'firewall'. The first refers to the high energy quanta an infalling observer will encounter at the horizon and cause him to 'burn up'. The second interpretation states there is a singularity at the horizon which 'breaks' the entanglement between the
outgoing modes and their interior partner modes. An infalling observer is simply ‘terminated’ at this singularity. In section 6.4 we will come back to this second interpretation and examine the link between the firewall and the true black hole singularity.

6.2 The thermal zone and mining

As seen in section 2.4, there is a centrifugal barrier at a distance of order $R_s$ from the horizon which reflects almost all but the $s$-waves. The occupation numbers of higher modes are exponentially suppressed by the tunneling barrier. So the Hawking radiation consists almost completely out of $s$-quanta.

The region behind the centrifugal barrier is also the region that can be approximated by Rindler space. The proper temperature varies from near Planckian to the Hawking temperature. As long as we keep away from the Planckian end, postulate 2 states that this region should be describable by ordinary quantum field theory. The entropy stored in this portion of space is part of the total black hole entropy. And although it is a small fraction of the total, it contains enough heat to be dangerous to anyone hovering above the horizon. The entropy is distributed over all angular momenta from $l = 0$ to $l = R_s m_p$, where $m_p$ is the Planck mass. The higher the angular momentum, the closer the modes are to the horizon. The correct picture is that the high $l$ quanta are emitted and absorbed by the stretched horizon and thereby thermalized.

So to an outside observer, a black hole can be thought of as an object consisting of two subsystems which constantly interact, namely the stretched horizon $H$ and the thermal zone $B$. The thermal zone is the shell of proper width of order $R_s$ just outside the membrane. Operationally, the difference between $H$ and $B$ is that $B$ can be probed by an outside observer without experiencing accelerations greater than the Planck scale, while $H$ cannot.

But now the argument of section 6.1.2 uses the purity of the total, final state of the Hawking radiation. Since the actual outgoing quanta in the radiation are primarily low angular momentum quanta, this argument applies to these modes and not directly to the vast reservoir of high angular momenta degrees of freedom that comprise most of the entropy of the black hole. On the other hand, the low angular momentum degrees of freedom are very dilute. The black hole emits only one $s$-wave quantum every Schwarzschild time, and that quantum is spread over the entire horizon area. Even if the $s$-wave degrees of freedom are completely entangled with the early radiation, which implies that an infalling observer would encounter them, this observer would probably not be seriously affected by them. To make the argument that there is a dangerous firewall, the degrees of freedom in the thermal zone must also be entangled with the early radiation. It is difficult to see how the analysis of section 6.1.2 can access these modes.

However, it is long known that it is possible to ‘mine’ energy from the modes trapped behind the centrifugal barrier [150]. This can be done by the same basic procedure we already encountered in section 2.5.1. One lowers some object quasistatically below the barrier, let the object absorb the trapped modes and then raise the object back above the barrier. In section 2.5.1 this object was a box that could be opened to collect ambient radiation and then closed
to keep the radiation from escaping. If one does not trust the box argument because the high energy radiation could make holes in it, one may also visualize the object as a particle detector or even a cosmic string [151].

In the context of such a mining operation, the arguments of section 6.1.2 can be applied to higher angular momentum degrees of freedom as well. One need only consider the internal state of the mining equipment to be part of the late-time Hawking radiation. In particular, the validness of effective field theory can be used to evolve the mode $b$ to be mined backward in time and to conclude for an old black hole that, even before the mining process took place, the mode must be fully entangled with the early-time radiation. The equivalence principle is then violated for these modes as well, suggesting that the infalling observer encounters a Planck density of Planck scale radiation and burns up.

The mining construction might seem artificial, and to some extent it is. But it seems there is no fundamental constraint why it should not be valid. In any case, if the argument can be made rigorously for the $s$-waves, it would seem strange that it would not apply also to the high angular momenta. There is no reason to assume that in the chaotic system the stretched horizon is, only the emitted $s$-wave quanta would be entangled with the early radiation.

### 6.3 Why complementarity is not enough

At first sight, it seems that one can resolve the firewall paradox by fully exploiting the freedom offered by complementarity which implies that outside and infalling observers can have different theories for predicting their observations. Each theory must be consistent with quantum mechanics and with semi-classical gravity in its regime of validity. But those theories need only agree on observations that the two kinds of observers can communicate without violating causality or leaving the regime of semi-classical gravity. For example, what the theory for an infalling observer predicts at or behind the stretched horizon cannot be communicated to an outside observer. Another way of saying this is that at the stretched horizon he has no longer a choice whether he wants to end up as an outside or inside observer. The theory describing his observations then need not be consistent with an outside observer’s theory. Especially, the combination of both theories into a global picture may yield a contradiction. The prime example of this was the no-cloning or entangled spin experiment of the previous chapter.

A similar type of resolution could be envisaged for the firewall paradox [152]. Consider two observers outside the black hole who both have access to the early emitted Hawking radiation. The first observer stays outside at all times and will find by unitarity that the Hawking radiation at late times is purified by a subsystem of the early radiation. He does not have access to the black hole interior. Therefore he cannot detect a contradiction by verifying that the late radiation is also purified by a different system behind the horizon.

The second observer on the other hand, jumps into the black hole and thus cannot measure the late Hawking radiation. Therefore, he cannot verify the entanglement between late and early radiation. Because of the constatation that he can freely fall through the horizon he will
implicitly detect that the late radiation is entangled with the modes behind the horizon. However, at the time he experiences this vacuum at the horizon it’s too late to communicate this to the first observer who stayed outside or to return himself as an outside observer.

But in order for this resolution to be valid, it must pass a consistency check. It must be impossible for an observer hovering in the thermal zone to measure the modes there before he reaches the stretched horizon. Because at that point the observer can still decide to fire his rockets and go back to spatial infinity, so his observations should match the ones of the outside observer. This implies that he will find the modes in the thermal zone entangled with the early radiation. But if he would then stop hovering and start to fall freely through the horizon, he finds a contradiction. So if an observer who will eventually fall into the black hole can measure the modes in the thermal zone before crossing the horizon, then complementarity is not enough to evade the firewall argument.

One can argue that such measurements are difficult. Remaining in the thermal zone for a long time requires a large acceleration outwards, which might pollute the setup due to emissions from her detector. However, in the limit of a large black hole \( R_s \to \infty \) the thermal zone becomes arbitrarily large and this complication appears to break down. Also, the validity of the firewall argument does not rely on the ability to measure any particular near-horizon mode with arbitrarily high accuracy, some finite fidelity is sufficient.

It is possible that a fundamental obstruction to the measurement of near-horizon modes prior to horizon crossing arises from some constraint that has been overlooked. But at this point, it is reasonable to conclude that the consistency check fails. Thus, complementarity appears to be insufficient. However, we will take a second look at this conclusion in 1.8.

### 6.4 Migrating singularity

In the previous sections it was argued that after a black hole has become old, the horizon is replaced by a firewall at which infalling observers burn up, in apparent violation of one of the postulates of black hole complementarity. Here, an alternative interpretation of the firewall phenomenon will be given in which the properties of the horizon are conventional, but the dynamics of the singularity are strongly modified [153, 154].

The existence of the firewall implies that there must be a singular, or at least highly excited, region at the horizon which prevents the entanglement of modes on both sides. One may even go further following [155, 156] and say that the lack of entanglement of the two sides of the horizon means the spacetime behind the horizon does not exist at all.

Another way to see this is the following. Initially, the thermal zone just outside the horizon is maximally entangled with the region behind the horizon. As the black hole emits Hawking radiation and becomes old, these modes behind the horizon are transferred to the radiation. In this process, the density matrix of the thermal zone is unchanged but the entanglement is transferred from behind the horizon to the radiation. One may say that there is a conservation
of entanglement. In this picture, instead of blowing up, the infalling observer finds fewer degrees of freedom after the thermal zone is passed. The argument of [156] would then say that there is no space behind the horizon for the infalling observer to exist in.

If one looks at the part of the black hole Penrose diagram in figure 6.1(a), then one sees that it is not consistent with the idea of the non-existence of spacetime behind the horizon. An observer could cross the conventional horizon and migrate to the region behind the firewall. A diagram which is more consistent with the hypothesis that the firewall is the end of spacetime is shown on figure 6.1(b). Instead of thinking of the firewall as part of the horizon, figure 6.1(b) suggests that we think of it as an extension of the singularity. The horizon only consists of the black part of the light sheet. A pleasing consequence of this interpretation is that now there is no conflict between postulates 2 and 4.

\[
\left( \frac{A_H}{4G} - \frac{A_F}{4G} \right) + S_R = S_0, \tag{6.17}
\]

where \(A_H\) is the spatial cross section area of the horizon and \(A_S\) is the spatial cross section area of the firewall. The first term in (6.17) is then the covariant entropy bound [125] on the light sheet crossing these two points. \(S_R\) represents the thermal entropy in the Hawking radiation passing the light sheet and \(S_0\) is the initial entropy of the black hole. The actual details are undoubtedly more complicated.

Another interesting observation is that when an observer jumps into an old black hole, the true horizon will shift outwards. The original horizon is merely an apparent horizon. This is also depicted on figure 6.2. As already explained in section 1.4.1 this is a consequence of the fact that the horizon is a global phenomenon, determined by all future events. Following the arguments of section 6.1, it is clear the firewall was located at the apparent horizon, and not at the true horizon. Adding the information of the infalling observer to the black hole makes it no
longer maximally entangled with the early radiation. It will take about the scrambling time and
the emission of a few Hawking quanta before the black hole is again maximally entangled with
the radiation. Therefore, it will also take a while before a new firewall forms at the true horizon.

The global nature of the horizon makes it clear that the firewall phenomenon does not pre-
vent information from entering the black hole in the infalling frame. This implies that the
firewall does not automatically solve the cloning problem since by the same line of reasoning
a second observer that was hovering outside at first could also enter the black hole. If, in the
exterior frame, the information is in the Hawking radiation, then complementarity has to be
invoked even after the Page-time.

With the shift from the singularity’s usual classical place to a location much closer to the
horizon, the infalling observer can still safely cross the actual horizon but the infalling time un-
til the singularity is extremely short. The further away an observer starts to freely fall towards
the horizon, the larger its momentum will be at arrival. This will cause a greater perturbation of
the Schwarzschild radius. Since the horizon shifts along a null geodesic and the observer moves
on a time-like geodesic, this will increase the survival time of the observer after the crossing of
the horizon. The same effect could be reached by jumping in alongside a very large mass. Note
however that this last possibility does not increase the total longevity of the infalling observer
but only the survival time between passing the horizon and arriving at the firewall.

In this way, the survival time becomes very sensitive to the mass of the infalling system. In
classical black holes, the opposite is true since the survival time is the classical geodesic distance
from the point where the system crosses the horizon to the singularity. Typically, this is of order
$M$, the mass of the black hole. However, since the horizon does respond to the infalling energy
there is always some small dependence of that geodesic distance on the infalling mass. In the
case where the singularity includes the firewall, the mass of the black hole becomes irrelevant.

![Figure 6.2: The shift of the horizon due to an infalling observer in Kruskal-Szekeres coordinates.](image-url)
6.5 Formation time of the firewall

A question that is not directly answered by the arguments of section 6.1 is at which point the firewall forms. The answer seems to be trivial, i.e. when the black hole is maximally entangled with the radiation. However, after a second look, the matter appears to be much more subtle. Basically, there are two possibilities. The first one is that the firewall arises after the scrambling time as argued in [149]. The second puts the time of birth at the Page time [153, 154]. To make the distinction between the two arguments, a set of subtle definitions is needed.

6.5.1 Generic and scrambled

First it will be explained what is meant by a generic state. Generic refers to what is true for the vast majority of states of a system. What is generic is what is true for a density matrix which maximizes the entropy, subject to whatever constraints may be relevant. For example, if there is no constraint whatsoever, the density matrix which maximizes the entropy is

$$\rho = \sum_{i=1}^{N} \frac{1}{N} |i\rangle \langle i|.$$  \hspace{1cm} (6.18)

Each basis state $|i\rangle$ has an equal probability, no matter what basis is chosen. From this it is clear why a property satisfied by this density matrix must be true for the majority of the states $|i\rangle$. It is also clear that any pure state is non-generic for some certain quantities. The state (6.18) is actually never achieved since it corresponds to infinite temperature. On the other hand, if there is a constraint on the total energy, the density matrix with maximal entropy will be thermal

$$\rho = \sum_{i=1}^{N} e^{-\beta E_i} |E_i\rangle \langle E_i|.$$ \hspace{1cm} (6.19)

effectively truncating the space of states when the individual degrees of freedom have energy greater than the temperature. Within the truncated space, a thermal density matrix is close to a completely incoherent state.

Now consider a large macroscopic system in a pure state. Denote the energy levels $E_n$ and let the corresponding eigenvectors be $|n\rangle$. A general pure state has the form

$$|\psi\rangle = \sum_{n} F_n |n\rangle.$$ \hspace{1cm} (6.20)

Now consider a small part of the system and trace out over the rest. The small subsystem will be described by a thermal density matrix with a temperature which is chosen to reproduce the average energy in the small subsystem. In other words, the state of the small subsystem is maximally incoherent subject to the constraint. The entropy of the small subsystem is maximal until the size of the system exceeds half of the total system. This follows from the analysis of section 5.3. That same analysis also showed that when the subsystem exceeds the half way point, the entropy in the case of an overall pure state starts to decrease. As seen in section 5.7 this phenomenon of maximal entropy for all small subsystems is called scrambling. In a
scrambled state almost everything that we normally measure has the generic thermal value. That is because the things we measure usually can be constructed from the observables of small subsystems.

On the other hand there are global observables which generally do not exhibit generic behavior. These are not the usual things we measure and they depend on the details of the pure state (6.20). Whether they are generic or not cannot be determined on the basis of whether the system is scrambled, for the simple reason that the definition of scrambling only involves small subsystems. Typically they involve at least half the degrees of freedom in an extremely intricate way. These global observables do not become generic in a scrambling time.

So it is important to notice that scrambled is not equivalent to generic. For many properties of a system they are completely different. The reason for conflating the two is that for most of the usual observables that are experimentally accessible, generic and scrambled are in fact the same.

In analyzing the time scales for firewall formation one may or may not have to take into account the evaporation process. If we want to know whether a firewall has formed by the Page-time then the evaporation process is of crucial importance. Because by definition the Page time has to do with evaporation. It is the point at which the remaining subsystem that represents the black hole is described by a thermal density matrix. At that point the black hole will have generic behavior as explained above. In particular, if a black hole has a firewall after the Page-time, then firewalls are generic features of black holes which means they exist for the vast majority of black hole states.

On the other hand, if we want to know whether a firewall has formed by the scrambling time, evaporation is not relevant. For an evaporating black hole of mass $M$, at the scrambling time the number of emitted quanta is only $\log M$, a negligible fraction of the total entropy. It is not evaporation but rather the unitary evolution of the entire system which causes scrambling.

The question is then, in what sense has the pure state of a black hole become generic by the time it is scrambled? And does that degree of genericity imply the existence of a firewall?

### 6.5.2 Fine grained and coarse grained

In most cases when we deal with a large system of many degrees of freedom we are interested in coarse grained quantities. To illustrate the difference between coarse and fine grained, consider a large system such as a box with perfectly reflecting walls. The box is filled with radiation and also some electrons to scatter the radiation and bring it to equilibrium. There are two cases to compare.

In the first case the photons and electrons are put into the box in a pure state with a given expectation value of the total energy. The quantum state at time zero is

$$|\Psi(0)\rangle = \sum_n F_n |n\rangle,$$

(6.21)
where the index $n$ represents the $n^{th}$ energy eigenstate in the box. For convenience we will define the states $|n\rangle$ such that the $F_n$ are real. At a later time $t$ the state evolves to

$$|\Psi(t)\rangle = \sum_n F_n e^{iE_n t} |n\rangle.$$  

(6.22)

The probability for the energy level $|n\rangle$ is

$$P_n = F_n^2.$$  

(6.23)

The other situation assumes that the degrees of freedom in the interior of the box are entangled with a heat bath on the outside. One could imagine that the entanglement took place at a time when there was a hole in the box, which was subsequently sealed. One may assume that the density matrix has the form

$$\rho = \sum_n P_n |n\rangle\langle n|$$  

(6.24)

at all times.

Now by fine-grained is meant that an observable is very sensitive to the relative phases between neighboring energy states in (6.22). Coarse grained means the opposite: a coarse grained operator is insensitive or has an exponentially small sensitivity in the size of the system to those phases. This implies that coarse grained operators practically have the same expectation values in the pure state (6.22) and in the mixed state (6.24).

For large closed systems, we saw that quantities built out of a small fraction of the degrees of freedom will take on their thermal values after a suitable scrambling time. For example, take a sub-volume of a box filled with radiation, consisting of a small fraction of the total volume. To exponential precision all expectation values involving fields within the sub-volume tend to the same value in the pure and the mixed states. In fact the analysis of section 5.3 suggests that this remains so as long as the sub-volume is smaller than half the size of the box. Whenever this is true the state is said to be scrambled. So the definition of coarse grained operators automatically includes observables who depend on a small number of degrees of freedom. This is because in the procedure of tracing out the irrelevant part of the system the information contained in the phases gets lost automatically. Therefore, this observable cannot depend on these phases.

On the other hand there are some quantities built out of more than half the system which are sensitive to the relative phases. Those are by definition fine-grained. Obviously, any quantity which can probe the purity of $|\Psi\rangle$ is fine-grained. Such quantities are extremely complicated functions of at least half the degrees of freedom in the box.

### 6.5.3 Special states and generic states

Let’s assume that the initial state has some special property. An example would be a reflecting box filled with highly coherent laser radiation. It is obvious that such a state is far from generic, and that the phases $\phi_n$ are special. It is also far from being scrambled.
Now consider the evolution of the phases in (6.22). In the initial state the phases were zero. If
the energy levels are characteristic of a chaotic system they will eventually be randomly sprinkled
over the unit circle. In other words, the typical state will be characterized by a classical
gas of indistinguishable particles on the unit circle, with random unpredictable positions. The
timescale for this to happen can be estimated by asking how long it takes for two neighboring
phases $\phi_n = E_n t$ and $\phi_{n+1} = E_{n+1} t$ to separate by an order 1 angle.

If we again suppose that a black hole with entropy $S$ has $e^S$ microstates, then the separation
between the energy levels is of order
$$\delta E \sim e^{-S}. \quad (6.25)$$

After a time $t$ the phase difference between neighboring energy levels will be
$$\delta \phi = t \delta E \sim te^{-S}. \quad (6.26)$$
The time scale for the phases to randomize will be the classical recurrence time
$$t_{\text{rec}} \sim e^S. \quad (6.27)$$

By contrast, as argued in section 5.7, the scrambling time $t_s$ for a black hole of mass $M$ is only
$$t_s \sim M \log M \sim \sqrt{S} \log S. \quad (6.28)$$
At the scrambling time neighboring phases have only separated by an exponentially small amount
$$\delta \phi \sim t_s \delta E = \sqrt{S} \log S e^{-S}. \quad (6.29)$$
Thus at the scrambling time the phases are extremely coherent. Evidently, the scrambling time
has nothing to do with the time for the state of a complex system to become generic. This
again illustrates the difference between scrambled and generic.

Fine-grained operators were introduced in the previous subsection as being sensitive to the
relative phases. An example of a fine-grained operator is
$$F = \sum_n (|n\rangle\langle n + 1| + |n + 1\rangle\langle n|). \quad (6.30)$$
The expectation value of $F$ varies with time like
$$\langle \Psi(t)|F|\Psi(t)\rangle = \sum_n F_n^* F_{n+1} e^{i(E_n - E_{n+1})t} + c.c., \quad (6.31)$$
where $E_n - E_{n+1}$ is of order $e^{-S}$. For $t \ll e^S$ the phase factors can be ignored since they
are extremely close to one. If one also assumes that $F$ is a smooth function of $n$ then one finds that
$$\langle \Psi(t)|F|\Psi(t)\rangle \approx 1. \quad (6.32)$$
However, as $t$ increases past the recurrence time the relative phases become random and

$$\langle \Psi(t)|\mathcal{F}|\Psi(t)\rangle \approx 0.$$

This is the same value that the expectation value of $\mathcal{F}$ would have in the incoherent density matrix (6.24). Note that nothing special happens at the scrambling time. At $t_*$ the neighboring phase differences are exponentially small and the expectation value of $\mathcal{F}$ is close to its value at $t = 0$.

Of course there are many degrees of fine-grained. The operator in (6.30) is maximally fine-grained because it depends on the phases of nearest-neighbor energy levels. If instead, the operator coupled second nearest neighbor energy levels the time scale for it to relax to zero would be more rapid. There are of course many other highly fine-grained operators but (6.30) is typical of them. In general they will achieve the value that they have in the incoherent density matrix only when the phases become random. By contrast, coarse grained observables tend to their incoherent counterparts much more rapidly, namely by the scrambling time.

It is intuitively very clear that a black hole which has just formed by a collapsing shell of matter is in a special state and will therefore not possess a firewall. This idea is strengthened by the analysis of section 1.4.1 which showed that the horizon forms before the shell reaches its Schwarzschild radius. There exist observers whose world line enters this part of the horizon while still out of causal contact with the shell. Locality then insures that nothing happens when the observer crosses the horizon. Another argument in favor of the specialness of young black holes is that the number of ways to make a black hole by collapse is probably much smaller than the exponential of the black hole entropy. Because as mentioned before, the entropy of ordinary matter which could collapse to form a black hole is much smaller than the corresponding black hole entropy. This fact supports the idea that young black holes are special states in the total black hole Hilbert space.

Now consider the description of the black hole in the frame of an outside observer. Let’s suppose that there exists a firewall-operator in the Hilbert space of black holes that detects the existence of a firewall. Call the firewall-operator $\mathcal{F}'$ and define it such that the existence of a firewall is indicated by $\langle \mathcal{F}' \rangle = 0$. The arguments of section 6.1 then imply that $\langle \mathcal{F}' \rangle = 0$ at the Page time when the black hole is maximally entangled with the radiation and is described by a thermal density matrix. This means that one should have $\langle \mathcal{F}' \rangle = 0$ in the vast majority of energy eigenstates $|n\rangle$. So in almost every eigenstate of the Hamiltonian, a firewall exists.

If the firewall-operator is similar to the fine-grained operator (6.30) than its expectation value in almost all states, i.e. states with random phases, are very close to zero. But in special states with smooth phase relations between neighboring states, the expectation value of $\mathcal{F} \sim 1$. Moreover, $\langle \mathcal{F} \rangle$ is time-dependent in the same way as the firewall-operator is. $\langle \mathcal{F}' \rangle = 1$ for young black holes and $\langle \mathcal{F} \rangle = 0$ for old black holes.

Now the question of how long it takes to form a firewall depends on just how fine-grained the operator $\mathcal{F}'$ is. If $\mathcal{F}'$ is maximally fine-grained then it takes a very long time to form a firewall. The evaporation process will have to bring the black hole to the Page time so that it
Chapter 6. The Firewall

is described by a thermal density matrix.

We can now summerize the conclusions of this section as follows. The existence of a firewall in the maximal entangled state at the Page time implies that the typical black hole state has a firewall. However, a black hole 'starts' in a special state without firewall. The scrambling time is the time for half the system to become typical, and not for the entire system to become typical. There are many subtle global observables that do not become typical until much longer times. If the existence of a firewall is one of these subtle questions then the timescale for the formation can be long.

6.6 Non-local dynamics

The arguments of section 6.1 imply that the following postulates are not mutually consistent:

- An infalling observer experiences nothing out of the ordinary at the horizon.
- The formation and evaporation of a black hole is a unitary process.
- Effective quantum field theory is valid outside the stretched horizon.

In the previous sections we’ve treated in detail the situation where we abandon the first postulate and place a firewall at the horizon. The consequences of and alternatives to non-unitary evolution were discussed in chapter 4. In this section we will examine the possibility of giving up effective field theory near the horizon in a very specific way.

As argued in section 6.3, complementarity is not sufficient to evade a firewall because the theory of an infalling observer alone is not consistent as it stands. In the thermal zone he should measure modes entangled with the early radiation since he can still return to infinity. But if he wants to pass the horizon safely this entanglement cannot be present. Thus, in order for both to be possible effective quantum field theory must break down well outside the stretched horizon, at least for an infalling observer.

Of course one would like to keep such novel physics in effective field theory to a minimum. But if we are to relax postulate 2 then the modified dynamics must be much larger in magnitude than expected. It is generally believed and argued in section 5.4 that the return of information in the Hawking radiation requires modification of the Hawking calculation only for observables involving a number of quanta of order $S$, or for a small number of quanta over extremely long time-scales. However, if one would like to preserve the equivalence principle and unitarity, then the arguments of section 6.1.2 show that an $N_a$ eigenstate has to evolve into a $N_b$ eigenstate, an effect visible in the two-point function over time-scales not much larger than the light-crossing time. In the remainder of this section we will discuss the revision of effective field theory by giving up locality and see how this addresses the firewall-paradox.

The idea is the following. For an old black hole, modify effective field theory by adding nonlocal
interactions in the black hole exterior which extend to a distance of order $R_s$ from the horizon. These nonlocal interactions must allow information to 'jump over' the thermal zone. In this way, the information transfer becomes nonviolent since it takes place only when the frequency of the Hawking quanta falls to of order the black hole temperature.

To sketch the mechanism in more detail we use a simple bit model. Consider an old black hole containing $N$ bits in a basis state $|j\rangle$. Because of the entanglement with the early Hawking radiation, the full state of the system is given by a sum over $j$. Now assume that the black hole emits a Hawking quanta, which we will idealize as a single bit. The equivalence principle requires this bit to be entangled with a bit behind the horizon. We must therefore use a state of $N+1$ bits to describe the black hole after emission, so that the evolution is

$$|j\rangle \rightarrow \sum_k |j, k\rangle_{bh} |k\rangle_m \equiv \sum_k |j, k; k\rangle,$$  \hspace{1cm} (6.34)

where $|j, k\rangle_{bh}$ represents the black hole state after emission and $|k\rangle_m$ is the outgoing Hawking bit.

After the Hawking bit is emitted, the black hole no longer is in equilibrium. Only after a scrambling time the black hole will again be in a typical state. During this scrambling or thermalization process the following evolution takes place

$$\sum_k |j, k; k\rangle \rightarrow \sum_{l, m} |l; m\rangle \langle l; m|j\rangle = \sum_{l=1}^{2^{N-1}} \sum_{m=1}^{2} |l\rangle_{bh} |m\rangle_m \langle l; m|j\rangle.$$  \hspace{1cm} (6.35)

The $N$ bits of $j$ are mapped onto the $N-1$ bits $l$, which now constitute the black hole, and the outgoing bit $m$. The effect is that one bit of entanglement with the early radiation is transferred to the outgoing bit $k$. This entanglement is produced by the coefficients $\langle l; m|j\rangle$. After the thermalization, $|l\rangle$ runs through the same proces as $|j\rangle$ started in (6.34).

Equation (6.35) describes unitary evolution from an $N$ bit space labeled by $j$ to $(N-1) + 1$ bit spaces labeled by $l$ and $k$. The state on the left is embedded in a space of $N+2$ bits, but the evolution has been specified only when two are in a definite state. Note that the evolution cannot be seen as a simple thermalization of the black hole because it evolves from a Hilbert space of $N+1$ bits to one of $N-1$ bits. Rather, it acts unitarily on the whole (black hole plus outside Hawking radiation) system.

The mechanism above can be summarized as follows. First, there is an emission process. This pulls the entangled pair denoted by $\sum_k |k; k\rangle$ 'out of the vacuum'. This entanglement is required to avoid high energy quanta at the horizon. One member of the pair starts to travel to infinity as Hawking radiation and the other ends up in the black hole. Then, the crucial new concept is to give up the idea of scrambling as a local operation at the stretched horizon. Instead, the scrambling transformation involves the entire state, including the emitted bit that is far from the horizon. It induces the required entanglement between the outside bit and the bits that make up the black hole. At this point the outgoing Hawking bit is far enough from the horizon so that there no longer is a threat to create a firewall.
However, as argued in [149] the modifications above proposed in [157–159] are insufficient. Suppose that we mine a close to the horizon and that the mining equipment can manipulate the quantum data in the storage bit. This manipulation is represented by an arbitrary unitary transformation $U$ on the storage bit. Instead of (6.34), we now get

$$ |j\rangle \rightarrow \sum_k |j, k; U_k\rangle . \tag{6.36} $$

For each $|j\rangle$, allowing $U$ to range over all unitary operations generates a basis for a Hilbert space of dimension 4 ($U(2)$ has 3 generators, plus the identity). In this sense, the right hand side of (6.36) spans a full $N+2$ bit Hilbert space. There can thus be no $U$-independent analogue of equation (6.34) involving only a remaining $N-1$ bit black hole and 1 additional storage bit. Explicit dependence of the Hamiltonian on $U$ would violate the usual rules of quantum mechanics.

An alternative to fix the two-bit mismatch in (6.36) might be to couple to the infinite number of states associated with the occupation numbers in outgoing radiative modes, though one would expect such a coupling to modify even the mean rate at which energy and information escape from the black hole.

A second alternative would be that some yet unknown physics, or some effect that has been neglected, simply prevents energy from being mined closer to the horizon than some distance $L_{\text{new}}$. This might be a new fixed scale or some geometric mean of $l_p$ and $R_s$. There would then be no obvious reason to believe that infalling observers experience radiation above the energy scale $L_{\text{new}}^{-1}$. The firewall at the horizon would be replaced by a much more innocent version at a distance $L_{\text{new}}$, since it is in this region the entanglement between the outgoing modes and the interior partner modes would start to get lost by the evolution (6.35). And at that distance, the exponential blueshift is much less strong.

### 6.7 The Harlow-Hayden conjecture

In section 5.6 we studied how fast an old black hole releases its information to an observer who has unlimited control over the Hawking radiation. Here however, we will address the practical question of precisely how long it takes to extract information from the Hawking radiation [160]. This timescale will then be compared to the black hole lifetime. The final goal is to put an operational constraint on the testability of the firewall-paradox.

The black hole entropy is proportional to $M^2$ in Planck units. As calculated in section 4.6.3, a black hole evaporates in a time proportional to $M^3$. An observer thus has to extract information from $n \sim M^2$ bits of Hawking radiation in a time

$$ t \sim n^{3/2} \tag{6.37} $$

to be able to jump in before the black hole evaporates. By going through the three subsections below we will compare this time to the time that follows from basic quantum information theory calculations.
6.7.1 The decoding process

In the black hole evaporation process, the system on the outside of the horizon is described by a pure state $|\Psi\rangle$ at all times. This state lives in the Hilbert space $\mathcal{H}_{\text{outside}}$. At any given time in the unitary evolution, one can factorize $\mathcal{H}_{\text{outside}}$ into subfactors with simple semiclassical interpretations

$$\mathcal{H}_{\text{outside}} = \mathcal{H}_H \otimes \mathcal{H}_B \otimes \mathcal{H}_R,$$

where $H$ again represents the stretched horizon, $B$ the thermal zone and $R$ the Hawking radiation. The time evolution of $|\Psi\rangle$ does not respect this factorization and cannot be computed using effective field theory due to the presence of $H$. But for the purposes here it is enough to consider the state at a given time.

If $|H|$ and $|B|$ are the dimensionality of $\mathcal{H}_H$ and $\mathcal{H}_B$ respectively, then the corresponding entropies $\log|H|$ and $\log|B|$ are both proportional to the area of the black hole horizon in Planck units at the time at which we study $|\Psi\rangle$. Thus, their size decreases with time. We will also consider $R$ as the part of the radiation that is nontrivially entangled with $\mathcal{H}_H$ and $\mathcal{H}_B$. This part of the radiation is enough to write the state on the outside as a pure state. Thus the size of $\mathcal{H}_R$ grows with time.

We will again consider the situation where the black hole has become old so that it is nearly maximally entangled with the radiation. This means that the combined system $BH$ has a density operator which is close to being proportional to the identity operator

$$\rho_{BH} \approx \frac{1}{|B||H|} I_B \otimes I_H.$$  

More carefully one would expect a thermal distribution in the Schwarzschild energy at the usual Hawking temperature. But since these low-energy modes can have very high proper energy near the horizon, the thermal density matrix for $\mathcal{H}_H \otimes \mathcal{H}_B$ is quite close to (6.39).

We can describe the state $|\Psi\rangle$ more accurately by considering the purifications $R_H$ and $R_B$. In other words, we make the $|\Psi\rangle$-dependent decomposition of $\mathcal{H}_R$

$$\mathcal{H}_R = (\mathcal{H}_{R_H} \otimes \mathcal{H}_{R_B}) \oplus \mathcal{H}_{\text{other}},$$

with $|R_H| = |H|$ and $|R_B| = |B|$, such that we can write the state of the full system, to a good approximation, in its Schmidt decomposition as

$$|\Psi\rangle = \left( \frac{1}{\sqrt{|H|}} \sum_h |h\rangle_H |h\rangle_{R_H} \right) \otimes \left( \frac{1}{\sqrt{|B|}} \sum_b |b\rangle_B |b\rangle_{R_B} \right).$$

Here, $h$ and $b$ label orthonormal bases for $\mathcal{H}_H$ and $\mathcal{H}_B$ respectively, and we have chosen convenient complementary bases for $\mathcal{H}_{R_H}$ and $\mathcal{H}_{R_B}$. $R_H$ is the purification of the stretched horizon and, just as in the previous section, $R_B$ is the purification of the thermal zone.

If we want to describe an infalling observer, the Schmidt basis is inconvenient to describe the state of the old black hole. From here on a basis for the radiation field will be used which
Chapter 6. The Firewall

242

is simple for an infalling observer to work with, and whose elements will be written as

\[ |bhr \rangle_R \equiv |b_1, \ldots, b_k, h_1, \ldots, h_m, r_1, \ldots, r_{n-k-m} \rangle_R. \]  

(6.42)

There are \( n \equiv \log_2 |R| \) total qubits, each of which we assume the observer can manipulate easily. \( b_1 \ldots b_k \) are the first \( k \) of these qubits, where \( k \) is the number of bits in \( \mathcal{H}_B \), and \( m \) is the number of bits in \( \mathcal{H}_H \). One can think of \( k + m \) as the number of qubits remaining in the black hole. The \( r_i \) qubits make up the remainder of the modes which have non-trivial occupation from the Hawking radiation.

Roughly one might expect that

\[ n \approx S - k - m, \]  

(6.43)

where \( S \) is the initial horizon area of the black hole in Planck units. This follows from postulate 3 of black hole complementarity which states that the number of states of a quantum black hole is given by \( e^S \sim 2^S \).

We will refer to (6.42) as the computational basis. In the computational basis one can write the state (6.41) as

\[ |\Psi \rangle = \frac{1}{\sqrt{|B||H|}} \sum_{b,h} |b \rangle_B |h \rangle_H U_R |bh\rangle_0, \]  

(6.44)

where \( U_R \) is some complicated unitary transformation on \( \mathcal{H}_R \). What unitary transformation it is will depend on the details of quantum gravity, as well as the initial state of the black hole. For simplicity, we have defined it to act on the state where all of the \( r_i \) qubits are zero.

Now the challenge for an observer is to act on the state of the Hawking radiation (6.44) with \( U_R^\dagger \). When this is done, it will be easy for him to confirm the entanglement between \( \mathcal{H}_B \) and \( \mathcal{H}_RB \). Engineering a particular unitary transformation to act on some set of qubits is precisely the challenge of quantum computation, and we will therefore often refer to the observer’s task as a computation.

So far we have been interpreting \( \mathcal{H}_B \) as the thermal atmosphere of the black hole, but to actually test the entanglement between the radiation and the thermal zone, it would be overkill for an observer to try to decode all of the atmosphere. Indeed the separation between \( \mathcal{H}_B \) and \( \mathcal{H}_H \) is rather ambiguous, and we are free to push some of the thermal zone modes we are not interested in into \( \mathcal{H}_H \). So from here on we will mostly take \( k \) to be \( O(n^0) \). In other words, we will consider the case where the observer is only trying to check the entanglement for a few of the bits in the atmosphere. This simplifies his computation, because in any event he only needs to implement \( U_R \) up to an arbitrary element of \( U(2^{n-k}) \), acting on the last \( n - k \) qubits of the radiation. In other words, the set of things he is really after is elements of \( U(2^n)/U(2^{n-k}) \).

Because the unitary group is continuous the observer will not be able to do the computation exactly. We thus need a good definition of how close he needs to get to reliably test the entanglement. A way to quantify the closeness of operators is the trace norm, which for an operator \( A \) is defined as

\[ \|A\|_1 \equiv \text{tr} \left( \sqrt{A^\dagger A} \right). \]  

(6.45)
If $A$ is hermitian this is just the sum of the absolute values of its eigenvalues. The motivation for this definition is as follows. Say $\rho_1$ and $\rho_2$ are two density matrices, and $\Pi_a$ is a projection operator for some measurement to give result $a$. Then

$$|P_1(a) - P_2(a)| = |\text{tr}[(\rho_1 - \rho_2)\Pi_a]| \leq \|\rho_1 - \rho_2\|_1.$$  \hspace{1cm} (6.46)

So if the trace norm of the difference of two density matrices is less than $\epsilon$ then the probabilities they predict for any experimental result will differ by at most $\epsilon$. The trace norm of their difference is clearly preserved by unitary evolution.

If both density matrices describe a pure state then the trace norm of their difference has a simple interpretation. For any two pure states $|\Psi_1\rangle$ and $|\Psi_2\rangle$ we can write that

$$|\Psi_2\rangle = e^{i\alpha} \left( \sqrt{1 - \frac{\delta^2}{4}} |\Psi_1\rangle + \frac{\delta}{2} |\chi\rangle \right), \hspace{1cm} (6.47)$$

where $|\chi\rangle$ is orthogonal to $|\Psi_1\rangle$, $\alpha$ is real and $\delta$ is real and positive. One then has

$$\| \langle \Psi_2 | \Psi_2 \rangle - \langle \Psi_1 | \Psi_1 \rangle \|_1 = \delta. \hspace{1cm} (6.48)$$

In order for the observer to do his computation, he needs to adjoin the radiation to a computer, whose initial state lives in a new Hilbert space $\mathcal{H}_C$, and then wait for the natural unitary evolution $U_C$ on $\mathcal{H}_R \otimes \mathcal{H}_C$ to undo $U_R$ and put the bits which are entangled with $B$ into an easily accessible form, let's say the first $k$ qubits of the memory of the computer. This process is sketched on figure 6.3. The connecting lines at the top and bottom indicate entanglement, and time goes up. During the computation, the subsystem $H$ just goes along for the ride, and after the computation its purification is split between $R$ and $C$ in some complicated way.

\hspace{1cm} \hspace{1cm} \hspace{1cm}

![Figure 6.3: Decoding the information in the Hawking radiation.](image)

$U_C$ is determined by the laws of physics and cannot be changed, so the only way that we can have any hope of getting the computer to do what we want is by carefully choosing its initial state. Without loss of generality we can take this initial state to be pure. But how many initial pure states are there to choose from? Of course there are infinitely many, but in any Hilbert
space $\mathcal{H}$ of dimension $d$ one can find a finite set $S_\epsilon \subset \mathcal{H}$ with the property that any pure state in $\mathcal{H}$ is within trace norm distance $\epsilon$ of at least one element of $S_\epsilon$. Such a set is called an $\epsilon$-net.

From (6.48) one observes that half of the trace norm difference is weakly bounded by the Hilbert space norm

$$\|\langle \Psi_2 | - |\Psi_1 \rangle \|^2 = 2 \left( 1 - \cos \alpha \sqrt{1 - \frac{\delta^2}{4}} \right) \geq \left( \frac{\delta}{2} \right)^2 = \left( \frac{1}{2} \|\Psi_2 \rangle \langle \Psi_2 | - |\Psi_1 \rangle \langle \Psi_1 | \| \right)^2,$$

(6.49)

where $\| \cdot \|_2$ denotes the Hilbert space norm. Thus, an $\epsilon/2$-net for the Hilbert space norm is also an $\epsilon$-net for the trace norm. The minimal size of an $\epsilon/2$-net for the Hilbert space norm is the number of balls of radius $\epsilon/2$ centered on points on the unit sphere in $\mathbb{R}^{2d}$ that are needed to cover it, which at large $d$ is proportional to some small power of $d$ times $(\epsilon/2)^{1/2}$. Intuitively we may just think of unitary evolution as an inner-product preserving permutation of the $(\epsilon/2)^{1/2}$ states.

We can now apply this discussion to the computer. The number of possible states that can come out of $U_C$ and which are distinguishable above the desired precision is

$$\left( \frac{\epsilon}{2} \right)^{1-2|C||R|} \approx \left( \frac{\epsilon}{2} \right)^{-2|C||R|}.$$

(6.50)

Out of these states a fraction

$$\left( \frac{\epsilon}{2} \right)^{-2|C||R|} \left( \frac{\epsilon}{2} \right)^{2|C||R|(1-2^{-k})} = \left( \frac{\epsilon}{2} \right)^{2|C||R|(1-2^{-k})}$$

(6.51)

have the property that the first $k$ qubits in the memory are entangled with $B$. The numerator on the left side is determined by the remaining freedom in the $\log_2(|C||R|) - k$ bits after the entangled state has been chosen for the first $k$ bits. For a generic $U_C$ we can interpret this ratio as the probability that any particular initial state will be sent to one of the desirable final states. The number of initial states for the computer is $(\epsilon/2)^{-2|C|}$. So, heuristically, the probability that we will be able to find an initial state for the computer that we can match to the radiation state so that after a single timestep, i.e. one action of $U_C$, we get one of the desired states is

$$P = \left( \frac{\epsilon}{2} \right)^{2|C||R|(1-2^{-k})-1}.$$

(6.52)

For any nontrivial $k$ and with $|R| = 2^n$, it is clear that this probability is extraordinarily small. Making the computer bigger, and thus increasing $|C|$, only makes the situation worse since it becomes even more unlikely the computation can be done! In order to beat this, one would need to allow the computer to run for of order

$$t \sim e^{2 \log \left( \frac{\epsilon}{2} \right)|R||C|}$$

(6.53)

timesteps, which is unimaginably long for any reasonable system size.

Equation (6.53) has the following physical interpretation. With no further assumptions about $U_R$ and $U_C$, the only way to do the computation with any certainty is to sit around and wait for
a quantum recurrence of the computer/radiation system. The quantum recurrence time, over which the system comes close to any given quantum state, is double-exponential in the number of bits of the whole system.

The main conclusion is that no amount of preparation of the initial state of the computer will allow the observer to do the decoding in any reasonable amount of time without imposing special assumptions about the dynamics both of the black hole and the computer. In the following assumptions we will take a look at such assumptions and argue that although they allow the observer to beat the double exponential down to a single exponential, that will most likely be all he gets.

As a final remark it is interesting to note that the result of this subsection is actually special to quantum mechanics. In the classical analogon of the setup used here the situation can easily be solved by making the computer bigger [160].

### 6.7.2 General unitary transformation with quantum gates

If doing any quantum computation takes a time double exponential in the number of bits we would never be able to do any quantum computation at all. It is of course wrong and in this section we will again study the quantum circuit model. Just as in section 5.7, we work on the 'quantum memory' of $n$ bits with some finite set of two-qubit unitary transformations, called quantum gates, on any two of the qubits. The computer builds up larger unitary transformations by applying the various gates successively. Interestingly enough, the number of different types of gates needed to generate arbitrary unitary transformations is quite small. A set of gates that has this property is called universal.

We can now ask how many gate-operations are needed to make a complicated unitary transformation like $U_R$ in (6.44). This is a good measure for the amount of time/space needed to actually do the computation, since we can imagine that the gates can be implemented one after another in a time that scales at most as a small power of $n$. For a set of $f$ fundamental gates, the number of circuits we can make which use $T$ total gates is clearly

$$
\binom{n}{f}^T.
$$

(6.54)

We have

$$
\ln \left( \frac{n!}{f!(n-f)!} \right) = \ln(n!) - \ln(f!) - \ln((n-f)!) \\
\approx n \ln n - n - \ln(f!) - (n-f) \ln(n) + (n-f) + O(f^2/n^2) \quad (6.55)
$$

$$
\approx f \ln n, \quad (6.56)
$$

where it was used that $n \gg f > 1$. So the number of circuits (6.54) can to a good approximation be written as

$$
\binom{n}{f}^T \approx n^{fT}. \quad (6.57)
$$
To proceed further we need some idea of size and distance for the unitary group. The unitary group on $n$ qubits is a compact manifold of dimension $2^{2n}$, and one parametrizes its elements as

$$U = \exp(i \sum_{a=1}^{2^{2n}} c_a t^a),$$

(6.58)

where $t^a$ are generators of the Lie algebra of $U(2^n)$, and one can very roughly think of the $c_a$'s as parametrizing a unit cube in $\mathbb{R}^{2^{2n}}$. So also roughly, one can think of linear distance in this cube as a measure of distance between the unitaries. For example say we wish to compute the difference between acting on some pure state $|\Psi\rangle$ with two different unitary matrices $U_1$ and $U_2$, and then projecting on onto some other state $|\chi\rangle$

$$\langle \chi|(U_1 - U_2)|\Psi\rangle = \langle \chi|(I - U_2 U_1^\dagger)U_1|\Psi\rangle \approx -i \langle \chi| \sum_a \delta c_a t^a U_1 |\Psi\rangle,$$

(6.59)

where $\delta c_a = c_a^2 - c_a^1$ and the approximation is to first order in $\delta c_a$. If the sum of the squares of the $\delta c_a$ is less than $\epsilon^2$, the right hand side will be at most some low order polynomial in $2^n$ times $\epsilon$. However, this polynomial is irrelevant.

Around each of our $n^{fT}$ circuits we can imagine a ball of radius $\epsilon$ in $\mathbb{R}^{2^{2n}}$. The total volume of all these balls will be of order of the full volume of the unitary group when

$$n^{fT} \epsilon^{2^{2n}} \approx 1,$$

(6.60)

where the right hand side represents the volume of the unit cube. Thus we see that in order to be able to make generic elements of $U(2^n)$ we need at least

$$T \sim 2^{2n} f^{-1} \log \left( \frac{1}{\epsilon} \right)$$

(6.61)

gates. Because $\epsilon$ appears inside a logarithm the crude nature of the definition of distance used here does not matter.

The important conclusion is that the number of gates is now only a single exponential in the number of bits. So the quantum circuit model is able to do arbitrary quantum computations much faster than the calculation of the previous subsection suggested. Given that we have so quickly beaten down a double exponential to a single exponential, one might be optimistic that further reduction in computing time is possible. Unfortunately, this does not seem to be the case. Simple modifications of the model such as changing the set of fundamental gates or considering higher spin objects instead of qubits make only small modifications to the analysis and don’t change the main $2^{2n}$ scaling. One could imagine trying to engineer gates that act on some finite fraction of the $n$ qubits all at once, perhaps by connecting them all together with wires or some such, but it is easy to see that any such construction requires a number of wires exponential in $n$. This implies the travel time between the various parts of the computer will be exponential in $n$. It is very reasonable to assume, as is widely done, that the quantum circuit model accurately describes what are physically realistic expectations for the power of a quantum computer. Thus if $U_R$ has no special structure, the observer cannot implement it (or its inverse) in a time shorter than $2^{2n}$. 
6.7.3 Decoding is slower than black hole dynamics

Now we turn to the question of whether or not the black hole dynamics constrain $U_R$ in any way that could help the observer to implement his computation faster. Because we know a black hole produces the state (6.44) relatively quick, i.e. after the Page time which scales like $\frac{n^3}{2}$, this seems to suggest that the observer might be able to implement $U_R^\dagger$ very quickly by some sort of time-reversal. This turns out not to be the case. To explain this we introduce a slightly more detailed model of the dynamics that produce the state (6.44). The conclusion of this subsection provides the crucial insight behind the Harlow-Hayden conjecture.

To describe the evaporation process it is clearly necessary to have a Hilbert space in which we can have black holes of different sizes. We can write this as

$$\mathcal{H} = \bigoplus_{n=0}^{n_f} \left( \mathcal{H}_{BH,n} \otimes \mathcal{H}_{R,n} \right).$$

The subscripts $n$ and $n_f - n$ indicate the number of qubits in the indicated Hilbert spaces. The dimensionality of $\mathcal{H}$ is $n_f 2^n$. One can imagine starting in the subspace with $n = 0$ and then in each timestep acting with a unitary transformation that increases $n$ by one. The evolution on the radiation will be taken to be trivial. The black hole becomes old after $n_f/2$ steps. This model could be seen as 'adiabatic' since it conserves the number of bits which have the physical interpretation as thermal entropy. So this model assumes (6.43) to be exact. This is not a bad approximation since the evaporation process takes a time of order $M^3$ and the thermal entropy of a one dimensional gas is given by $n \approx LT \sim M^3 M^{-1} = M^2 = S$. So the number of Hawking quanta produced is of order of the entropy of the black hole.

An actual black hole formed in collapse will have some width in energy, which here means a width in $n$. But by ignoring this we can make a further simplification. Starting in one of the $2^{n_f}$ states with $n = 0$, the evolution never produces superpositions of different $n$. So we can actually recast the whole dynamics as unitary evolution on a smaller Hilbert space of dimension $2^{n_f}$, but in which the interpretation of the subfactors change with time. This is illustrated with a circuit diagram in figure 6.4 which represents the black hole dynamics for a 7-bit black hole. With each step the subfactor we interpret as the radiation gets larger.

![Figure 6.4: A 7-bit black hole with an increasing subsystem which represents the Hawking radiation.](image)

With this simplification we can now combine all of the timesteps together into one big unitary matrix $U_{\text{dyn}}$ acting on the $2^n_f$ dimensional Hilbert space. The matrix $U_R$ appearing in the state (6.44) is a result from acting with $U_{\text{dyn}}$ on the initial state. Therefore, $U_R$ will depend rather sensitively on the initial state while $U_{\text{dyn}}$ clearly does not. Because the observer only needs to be able to do the computation for some particular initial state, we will for simplicity choose it to just have all the bits set to zero. For $n > \frac{n_f}{2}$ we thus expect the following to be true

$$U_{\text{dyn}} |00000\rangle_{\text{init}} = \frac{1}{\sqrt{|B||H|}} \sum_{b,h} |b\rangle_B |h\rangle_H U_R |bh0\rangle_R.$$  

(6.63)

This equation tells us something about $U_R$, whose complexity we are interested in understanding. To proceed further we need to make some sort of assumption about $U_{\text{dyn}}$. This is a question about the dynamics of quantum gravity so we can’t say anything too precise, for those black holes which are well understood in matrix theory [161] or AdS/CFT [119, 162] the dynamics are always some matrix quantum mechanics or matrix field theory. As mentioned in section 5.7.2, the observation that black holes are fast scramblers strongly supports the idea that this is true for all black holes. Theories of this type can usually be simulated using polynomial-sized quantum circuits [163]. Therefore, it seems quite reasonable that $U_{\text{dyn}}$ can be generated by a polynomial number of gate-operations. Such circuits are usually called ‘small’. So more precisely, we want to know the following: does the existence of a small circuit for $U_{\text{dyn}}$ imply the existence of a small circuit for $U_R$? If the answer is yes, then our model would imply that Alice can decode $R_B$ out of the Hawking radiation fairly easily.

We can now define a new operator 

$$\tilde{U}_R = U_{\text{dyn}} U^\dagger_{\text{mix}},$$  

(6.65)

which has the property that 

$$\tilde{U}_R \frac{1}{\sqrt{|B||H|}} \sum_{b,h} |b\rangle_B |h\rangle_H |bh0\rangle_R = \frac{1}{\sqrt{|B||H|}} \sum_{b,h} |b\rangle_B |h\rangle_H U_R |bh0\rangle_R.$$  

(6.66)

Since $U_{\text{mix}}$ is a standard operation in quantum computation which can be implemented very easily and $U_{\text{dyn}}$ is described by a small circuit, it is clear that $\tilde{U}_R$ can be implemented with a small circuit. Apparently, this seems to be exactly what the observer needs. He can just apply the inverse circuit to the state (6.44) and the decoding is accomplished.

But now it is crucial to realize that this is not possible. Although the operator $\tilde{U}_R$ appears to act only on the radiation, the circuit this construction provides involves gates that act on all of the qubits, thus also on the bits in the thermal zone and at the stretched horizon! But while doing the decoding, the observer has no access to the qubits in $B$ and $H$. 

Chapter 6. The Firewall 248
Of course, if the circuit $\tilde{U}_R$ really acted as the identity operator on $B$ and $H$ for any initial state this would not matter since the observer could just throw in some ancillary bits in an arbitrary state to replace those in $B$ and $H$ and still use the $\tilde{U}_R^\dagger$ to undo $U_R$. The problem with this is that (6.67) holds only when $\tilde{U}_R$ acts on the particular state $(|B||H|) - 1/2 \sum_{b,h} |b\rangle_B |h\rangle_H |bh\rangle_R$. This can be traced back to the fact that the definition of $U_R$ in the first place depended on the initial state of the black hole that $U_{dyn}$ acts on.

The lesson of this section is that because the observer does not have access to all of the qubits in the system, he is unable to simply time-reverse the black hole dynamics and extract $R_B$ in a time that is polynomial in the number of bits (= the entropy). Without such a simple construction, he will in general be left with no option to brute-force his construction of $U_R^\dagger$ using of order $2^{2n}$ gates.

It is still possible that some yet-unknown special feature of black hole dynamics will conspire to provide a simple circuit for $U_R$, but this would be rather surprising.

It is interesting to note that if the Harlow-Hayden conjecture is correct, it supersedes many of the black hole thought experiments of the previous chapter. In particular, the argument that the scrambling of information by a black hole in a time no faster than $R_s \log(R_s/l_p)$ would no longer be needed. This indicates that the standard conceptions about black hole complementarity might need some rethinking, which will be done in the next section.

### 6.8 Next generation complementarity

The AMPS argument together with the Harlow-Hayden conjecture lead to a major rethinking about the nature of complementarity [164]. The AMPS argument made clear that the role of entanglement was much more subtle than the one it had in the original formulation of chapter 5. Therefore, we will review the evolution of a quantum black hole by the principles of complementarity with a special emphasis on entanglement. After that we will see how the Harlow-Hayden conjecture can be combined with the ideas of sections 6.3 and 6.6.2 to give rise to a modified complementarity principle that might make the need for a firewall superfluous.

#### 6.8.1 The stretched horizon as a hologram

To an infalling observer, the modes in the region behind the horizon $A$ are entangled with the modes in the thermal zone $B$. In fact, we even know that every exterior mode has an interior partner mode with which it is maximally entangled. So there is a definite pairing of modes in $B$ and $A$. Let’s denote a particular mode in $B$ by $B_i$ and the corresponding partner in $A$ by $A_i$.

This discussion of $(A, B)$ entanglement in the infalling frame must be translatable to the language of the exterior degrees of freedom. Because by assumption, the interior degrees of freedom are constructed from the exterior degrees of freedom. The exterior description is thermal, and it can be thought of as a scrambled system. Appendix E gives a review of the difference between
scrambled entanglement and ground state entanglement. So complementarity states that the ordered entanglement of a ground state is dual to the scrambled entanglement of a random (thermal) state. This duality between the infalling and exterior frame already was the central issue of complementarity as formulated in chapter 5, but here the emphasis is on the duality of the two kinds of entanglement.

Since most of the exterior degrees of freedom are in the stretched horizon $H$, one can assume that the $A_i$ are constructs made of the $H$ qubits. Identifying $A_i$ in $H$ is a matter of finding a unique subsystem of $H$ which is maximally entangled with $B_i$, the partner of $A_i$. In general, there is no guarantee that such a subsystem of $H$ exists. However, for the case of a relatively young black hole we can be sure that it does. By relatively young is meant that the black hole has already scrambled, but the evaporation is negligible. In that case the $HB$ system is in a pure but scrambled state.

Given that the state of $HB$ is pure, there is an important consequence of this fact. Namely, that to an equally high degree of approximation, every small subsystem is maximally entangled with the rest of the system. Since $B$ is a small subsystem of the $HB$ system, it follows that $B$ is maximally entangled with $H$. Furthermore, each qubit of $B$ is almost exactly maximally entangled with a unique subsystem of $H$. Because scrambled systems hide their entanglement very well, it is unlikely to easily recognize the subsystem of $H$ that is maximally entangled with $B_i$. But we can be sure that it exists. Call this subsystem $H_{B_i}$.

So we’ve come to the conclusion that $B_i$ is maximally entangled with $A_i$ and with $H_{B_i}$. But maximal entanglement is monogamous. Therefore, it follows that $A_i$ and $H_{B_i}$ must be the same thing. The formal equation

$$A = H_B$$

expresses this identification. This identification can only be made if not a single observer can detect $A_i$ and $H_{B_i}$. Because in that case both can be considered as one fundamental mode that manifests itself as $A_i$ to one observer and as $H_{B_i}$ to another. Another way to say this is that the stretched horizon $H$ is a hologram at the horizon that represents the interior $A$. From the discussion above it is clear that the relation between the interior and exterior degrees of freedom is extremely fine-grained.

An issue that is bound to come up is the non-linearity of the $H \leftrightarrow A$ mapping. By non-linearity is meant that the relation between $A_i$ and operators in $H$ depends on the initial state of the black hole. That is because the particular form of $H_{B_i}$ is state-dependent since it is formed by scrambling the initial state. Although this does not imply an observable non-linear violation of quantum mechanics in either the exterior or infalling frames, it does seem to violate the linear spirit of quantum mechanics. We will return to this issue of state-dependence in section 6.9.2.
6.8.2 The transfer of (distillable) entanglement

We would like to have a quantitative concept of how entangled $B$ and $H$ are during the evaporation process [164, 165]. To define the amount of entanglement between $B$ and $H$ we will introduce the concept from quantum information theory called distillable entanglement [166], represented by the symbol $D$. We will not attempt to be too precise in its definition; in essence $D$ is the number of Bell pairs shared by two subsystems.

Since the subsystems we consider are always nearly maximally entangled we will take $D$ to count the number of ‘regulated’ Bell pairs. They obey the following two conditions. One is that the density matrix of the union of the two qubits is almost pure. In other words, the entanglement entropy of the union is less than $\epsilon$, with $\epsilon$ very small. And secondly, the density matrices of the individual qubits are almost maximally random. The entanglement entropy of each qubit is almost maximal and greater than $\ln 2 - \epsilon$.

If the $HB$ system is not in a pure state, we obtain the distillable entanglement as follows. Consider a unitary operator constructed as the tensor product of a $2^NH \times 2^NH$ matrix in the Hilbert space of the subsystem $H$, and the identity matrix in $B$. Apply it to the density matrix $\rho_{BH}$ of the $HB$ system

$$\hat{\rho}_{BH} = U^\dagger \rho_{BH} U.$$ \hspace{1cm} (6.68)

Next, pair the $B$-qubits with a subset of the $H$-qubits and count the number of regulated Bell pairs. Finally, maximize that number with respect to all $2^NH \times 2^NH$ unitary transformation. Basically, this is the unscrambling procedure described in the previous section. The resulting number of Bell pairs is the distillable entanglement $D$.

A useful bound on $D$ can be found by defining

$$\mu = \frac{1}{2}(S_B + S_H - S_{BH}),$$ \hspace{1cm} (6.69)

where $S_B$, $S_H$ and $S_{BH}$ represent the entanglement entropy of the thermal zone, the stretched horizon and the combined system respectively. Thus, $\mu$ is defined to be the half of the mutual information [79]. The $HB$ subsystem is purfied by the radiation subsytem $R$, so we can write

$$\mu = \frac{1}{2}(S_B + S_H - S_R).$$ \hspace{1cm} (6.70)

It is known that the distillable entanglement is bounded by $\mu$ [166]

$$D \leq \mu.$$ \hspace{1cm} (6.71)

There are two situations in which $D$ equals $\mu$ or is very close to it. The first case is $\mu = 0$. Since both $\mu$ and $D$ are never negative it follows that if $\mu$ vanishes, so does $D$. The second less trivial case is when $\mu$ is maximal or close to it [165].

Now we again represent the black hole as a system of $N$ qubits where $N$ is the black hole
entropy shortly after collapse. The qubits are assigned to the subsystems $H$, $B$ and $R$ according to

$$N = N_H + N_B + N_R.$$  \hspace{1cm} (6.72)

At any given time the black hole entropy is

$$S_{BH} = N_H + N_B.$$  \hspace{1cm} (6.73)

We also denote the fraction of the black hole entropy contained in the thermal zone by $f$

$$S_B = f S_{BH}.$$  \hspace{1cm} (6.74)

Based on the previous subsection, it is clear that a necessary condition for an uncorrupted black hole interior is that the distillable entanglement between $B$ and $H$ should be equal to the number of qubits in $A$. If the number is less than that there is not enough of an entanglement resource to define all the interior modes. Even worse, if $D = 0$ it is impossible to define any vacuum modes in $A$. So at that point the geometry seems to be terminated at the horizon by a firewall. This idea was already presented in section 6.4. So the AMPS argument can be formulated as a calculation which shows that the $HB$ distillable entanglement goes to zero before the black hole has evaporated.

In figure 6.5 a schematic representation of the evaporation process is given. The $N$ qubits are represented in a box which is scrambled at all times. As the evaporation process takes place, the part representing the Hawking radiation depicted on the right of the box gets bigger. In the box on the left of figure 6.5 representing the initial black hole state, the part representing $B$ has $fN$ qubits and the $H$-box has $(1-f)N$.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure6.5.png}
\caption{A schematic representation of the evaporation process showing the $H$, $B$ and $R$ subsystems.}
\end{figure}

Initially, in the infalling frame the number of Bell pairs is equal to $fN$, which is the number of qubits in the thermal zone $B$. As we will see, that amount of entanglement persists for a long time as the black hole evaporates. But at some point the entanglement begins to diminish, and by the Page time it vanishes. This is the heart of the AMPS argument. Because the original statement of section 6.1 was that after the Page time the emitted Hawking quanta are entangled with the early radiation such that the entanglement with modes behind the horizon disappeared.
By the arguments of the previous subsection, this entanglement with modes behind the horizon is the distillable entanglement between $H$ and $B$.

We now define the cusp time $t_c$ as the time at which $H$ becomes smaller than half the total system. Note that the cusp time is earlier than the Page time at which $N_R$ becomes half the system. Before $t_c$, $B$ and $R$ are small subsystems of a scrambled system. This means they are maximally random and their entanglement entropies are therefore given by

\begin{align}
S_B &= N_B \\
S_R &= N_R \\
S_{BR} &= N_B + N_R,
\end{align}

where we again omitted the factor $\ln 2$. Since the total state is pure we have

\begin{equation}
S_H = S_{BR} = N_B + N_R.
\end{equation}

This gives

\begin{equation}
\mu_{BH} = N_B.
\end{equation}

Since this is the maximal value for $\mu_{BH}$, we can also write

\begin{equation}
D_{BH} = N_B,
\end{equation}

or

\begin{equation}
\frac{D_{BH}}{S_{BH}} = \frac{N_B}{N_B + N_H} = f \ (t < t_c).
\end{equation}

After $t_c$ the fractional distillable entanglement decreases linearly with time. It vanishes at the Page time and stays equal to zero until the black hole evaporates. To see this, note that between the cusp time and the Page time all three subsystems have less than half the total number of qubits. Therefore

\begin{align}
S_B &= N_B \\
S_H &= N_H \\
S_R &= N_R,
\end{align}

such that

\begin{equation}
\mu_{BH} = \frac{1}{2}(N_B + N_H - N_R) = N_B - \frac{N_R - N_c}{2},
\end{equation}

where $N_c \equiv N_H - N_B$ is the number of bits in the radiation at $t_c$. In other words, the mutual information begins to decrease relative to $N_B$ once the cusp is passed. It is also easy to see that it vanishes at the Page time when $N_R = N_H + N_B$. From the fact that $\mu$ bounds $D$ we see that the distillable entanglement between $H$ and $B$ also decreases to zero at the Page time.

So we can conclude that $D$ remains large enough so that the interior degrees of freedom can be defined for a long time. Indeed, for small $f$, the number of degrees of freedom in $H$ is very large compared to $B$ so $t_c$ will almost be the Page time. This is good news for a long-lived interior geometry. But after the Page time there is no hope, the fine-grained quantities $A_i$ associated with $HB$ entanglement have disappeared altogether. As long as we insist that the interior be
built from near-horizon degrees of freedom the evaporation process will destroy the necessary entanglements and a firewall must replace the smooth horizon.

Note that we already discussed the formation time of the firewall in section 6.5. The conclusion of this section, that the firewall starts to form at the cusp time is only valid within the model of second generation complementarity where the modes inside the horizon are defined by their entanglement. Outside this model one again has to rely on the discussion of section 6.5.

6.8.3 Standard complementarity ($A = R_B \subset R$)

One implicit assumption in the previous subsection is that the degrees of freedom of the interior must be constructed from exterior degrees of freedom which are physically near the black hole. This is called the proximity assumption. So one can argue that the AMPS did not prove that the standard postulates of complementarity are inconsistent, but only that they are inconsistent with the proximity assumption.

A way to view this is the following. When the black hole is young, the information in $A$ is redundant with the information in $H$ through the identification (6.67). However, if the black hole starts to evaporate and releases its information, the information in $A$ must eventually become redundant with information in $R$.

Subsystem $R$ does provide a resource for entangled Bell pairs. Indeed, after the Page time the degrees of freedom $B_i$ continue to be entangled but with a subsystem of $R$ instead of $H$. The distillable entanglement between $B$ and the union $HR$ remains at all times large enough to define partner modes for $B_i$. In figure 6.6 the fractional distillable entanglement of $RB$ is plotted alongside that of $HB$. The total distillable entanglement is in fact conserved. Before $t_c$ the Bell pairs are shared between $H$ and $B$. After the Page time they are shared between $R$ and $B$. Between the cusp and Page time they are partly shared with $H$ and partly with $R$, but the number of Bell pairs shared by $B$ is constant.

After the Page time, the degree of freedom that is maximally entangled with $B_i$ lives in $R$ and can be called $R_{B_i}$. The hypothesis that at late times $A$ becomes redundant with the Hawking radiation would replace the $A \leftrightarrow H$ mapping by an $A \leftrightarrow R$ mapping,

$$A_i = R_{B_i}. \tag{6.86}$$

The relation between $B_i$ and $R_{B_i}$ is of course very fine-grained and depends on the precise initial state and dynamics of the black hole. Equation (6.86) expresses the identification of the two subsystems that purify the thermal zone $B$. This removes the need for a double maximal entanglement of the mode $B_i$ and therefore a firewall is no longer necessary.

Black hole complementarity in its most fundamental from states the non-invariant localization of information. The identification (6.86) implies a radically greater localization-ambiguity than $A_i = H_{B_i}$. Note however that such large scale delocalization of information is already present in any holographic theory [125]. So in a sense (6.86) is an extension of the standard complementarity idea of section 6.8.1. The only difference is that we have abandoned the idea
that the interior modes $A_i$ should be constructed from modes near the horizon. In other words, where we at first only allowed the $A_i$ to be built up from the stretched horizon modes, we now allow $A_i$ to be built from the early radiation mode $R_{Bi}$ which is at a macroscopic distance from the black hole.

In section 6.3 however, we actually already argued against the identification (6.86). An observer equipped with a very powerful quantum computer uses as input the early half of the Hawking radiation. The output is a specific qubit that the observer can hold and manipulate. Of course, this computer again knows again the initial black hole state and the black hole dynamics. Assume the observer can use this computer to distill $R_{Bi}$. Then the observer could start to freely fall and check whether $R_{Bi}$ is maximally entangled with $B_i$. If it is, then by the monogamy of entanglement, $B_i$ cannot also be entangled with $A_i$. In other words, because one observer can access both $R_{Bi}$ and $A_i$ they cannot be two manifestations of the same fundamental mode without leading to quantum cloning. In section 6.3 it was said that there might be some fundamental constraint which prevents an observer from accurately measuring $B_i$. But instead of a limit on the measurement of $B_i$ we can now use the Harlow-Hayden conjecture to exclude the measurement of $R_{Bi}$ in less than the evaporation time. This implies that $A_i$ and $R_{Bi}$ are not accessible to a single observer.

### 6.8.4 Strong complementarity

Besides standard complementarity, there is a second form of next generation complementarity that might evade a firewall. Consider the worldline of an infalling observer $\mathcal{F}$ to its end on the singularity. The causal past of this worldline defines the observer’s causal patch. This patch can be sliced by a family of space-like surfaces, one of which passes through the modes $A$, $B$ and asymptotes to the light-like boundary of the observer’s causal patch. This is shown on figure 6.7. Complementarity requires entanglement between $B$ and $A$ in this frame.
On the other hand, the causal patch of an outside observer $O$ contains $B$ as well as the outgoing radiation that was seen from $F$’s patch, but it does not contain $A$. On $O$’s space-like slice $B$ must be entangled with the outgoing radiation, i.e. with $R_B$.

We can now formulate a new version of complementarity called *strong complementarity*.

**Strong Complementarity** Each causal patch has its own quantum description.

In $F$’s quantum mechanics $B$ is entangled with $A$ and not with the outgoing radiation. In $O$’s description $B$ is entangled with $R_B$. At the level of coarse grained properties of the radiation, the descriptions must match in the overlap region of the causal patches, certainly where it is well understood that $F$ and $O$ see the same Hawking quanta. But the two descriptions must not match on the fine-grained level. $B$ is a coarse-grained object and therefore $F$ and $O$ should agree on it, but $R_B$ is extremely fine-grained. So the large-scale entanglements of a pure but scrambled state are not present in $F$’s patch, but they are in $O$’s patch. By invoking the Harlow-Hayden conjecture this does not lead to any observable contradiction.

However, a possible constraint to the Harlow-Hayden conjecture might be that $F$ may try to slow down the evaporation process while his computer is distilling $R_B$. One way to do that would be to surround the black hole by mirrors to keep it from radiating. But this might not help $F$ if the decoding time is exponential. An exponential time scale has multiple meanings.
for a complex closed system. For one thing, the time scale for resolving tiny energy difference between neighboring states is of order $\Delta t = 1/\Delta E$. For a system of entropy $S$ this is equal to $e^S \sim e^N$. Of greater relevance, over such time scales Poincaré recurrences will repeatedly occur, undoing and re-collapsing the black hole. It is unlikely that the identity of a mode $A_i$ has any meaning over such long times.

### 6.9 Problems with $A \subset R$

In the standard complementarity picture there was one global Hilbert space and the black hole interior was identified with the modes in the early radiation via common entanglement with $B$. In strong complementarity each causal patch had it’s own Hilbert space. Both pictures relied on the Harlow-Hayden conjecture to prevent quantum cloning.

In this section however we argue that the embedding of $A$ in $R$ is not as natural as it might seem and leads to some substantial difficulties [167].

#### 6.9.1 Measurements create a firewall

Suppose a hovering outside observer measures some particle of the early radiation $R$. He does not want to verify any entanglement so he does not have to do a complicated and time-consuming decoding operation. Denote the annihilation operator associated with the measured early mode by $e$. The modes as seen by an infalling observer again have annihilation operators $a$. We now want to show that the commutator of $e$ with $N_a$ is of order one.

For simplicity we work with the parity $(-1)^{N_e}$. Take a basis in which

$$(-1)^{N_e} = \sigma^z \otimes I.$$  \hspace{1cm} (6.87)

That is, we factor the Hilbert space into the measured parity and the rest.

To an outside hovering observer, the modes behind the horizon are the partnermodes of $B$. Consider such a partnermode and denote its annihilation operator by $\tilde{b}$. Now consider $(-1)^{N_{\tilde{b}}}$. If we take $A \subset R$, then we may expand

$$(-1)^{N_{\tilde{b}}} = I \otimes S^0 + \sigma^x \otimes S^x + \sigma^y \otimes S^y + \sigma^z \otimes S^z.$$ \hspace{1cm} (6.88)

The matrices $S^\mu$ are constrained only by $(-1)^{N_{\tilde{b}}}(-1)^{N_{\tilde{b}}} = 1$. As argued in the previous section, because ordered groundstate entanglement for the infalling observer is dual to scrambled entanglement for the outside observer, the relation between the complementary descriptions is expected to involve a scrambling of the Hilbert spaces. Therefore, the operators $S^\mu$ are generic and have typical eigenvalues of order one. It follows that the commutator of $(-1)^{N_e}$ and $(-1)^{N_{\tilde{b}}}$ is of order one, and so therefore is $[e, \tilde{b}]$.

This result is to be expected for the following reason. First, note that the commutator between
an early and a late mode \([e, b]\) is zero. They simply are annihilation operators corresponding to different modes of the same scalar field, so they act in the same Fock basis. Without the embedding, the interior partner mode \(\tilde{b}\) would also commute with \(e\) for the same reason. But recall that we saw that the bit in the early radiation which is entangled with some late mode in \(B\) is very scrambled. Therefore, to expose this bit we need to make the transformation to a completely different basis. If we now identify this bit as the mode we are considering behind the horizon, the associated annihilation operator \(\tilde{b}\) will definitely not work on the same Fock basis as \(b\) and \(e\). Therefore, \([e, \tilde{b}]\) will not be zero. In fact, the reasoning above even shows it is of order 1. This is because the outside is maximally scrambled.

This nonzero commutator has some major consequences. In particular, if we start with an eigenstate of \((-1)^N b\) and measure \((-1)^N e\), the eigenvalue of \((-1)^N b\) changes with probability of order one. For convenience, we show this for the process with the roles of \(\tilde{b}\) and \(e\) switched, which is equivalent but clearer in the basis \((6.87)\). Choose a state \(|\psi\rangle\) such that

\[(-1)^N e |\psi\rangle = + |\psi\rangle. \tag{6.89}\]

But if we now first measure the parity by means of the partner mode \(\tilde{b}\) this state changes to

\[|\psi\rangle \rightarrow (-1)^N \tilde{b} |\psi\rangle. \tag{6.90}\]

So after this measurement, the expectation value of \((-1)^N e\) becomes

\[
\langle \psi | (-1)^N b (-1)^N e (-1)^N \tilde{b} |\psi\rangle = \langle \psi | \sigma^z \otimes (S^0 S^0 + S^z S^z - S^x S^x - S^y S^y) |\psi\rangle + \text{cross terms}. \tag{6.91}
\]

We now average \((6.91)\) over all \(S^\mu\) consistent with \((-1)^N b (-1)^N \tilde{b} = 1\). The cross terms involve products of distinct \(S^\mu\) and so are on the average zero, since the constraint allows independent sign flips. The distinct \(S^\mu S^\mu\) are on average equal, so on average the expectation value is reduced from 1 to 0 by the measurement.

In terms of the modes of an infalling observer, \(\tilde{b}\) can be expanded as a sum of \(a\) and \(a^\dagger\) because of the Bogoliubov transformation that was the origin of the Hawking radiation (see section 2.3.1). This implies the commutator of \(e\) with one of these (generically both) is also of order one. Now suppose that there was no firewall, so that the infalling observer sees the vacuum \(a |\psi\rangle = 0\). However, after the hovering observer has measured his bit, the order one commutator \([e, a]\) means that the state has been heavily perturbed. This is true for every mode \(a\), so the hovering observer has created a firewall!

It may seem odd that measurement of a single bit can perturb many others, but this seems to be a manifestation of the butterfly effect: perturbation of a single bit, followed by a scrambling operation, perturbs all bits.

To summarize, what happens is the following. An observer is freely falling towards the horizon of an old black hole. We assume the identification \(A \subset R\) is true and that it evades the firewall. So the falling observer is happily detecting the vacuum \(a |\psi\rangle = 0\). But then, a second hovering, outside observer decides to measure one of the modes in the early radiation \(R\). This mode has an annihilation operator \(e\). Above, it was shown that because of the identification \(A \subset R\), the
lowering operator $\hat{b}$ of a mode behind the horizon associated with the outside observer does not commute with $e$. Since $\hat{b}$ is related to $a$ and $a^\dagger$ via a Bogoliubov transformation, also $a$ and $e$ do not commute. We have even argued their commutator is of order one. Because $\hat{b}$ corresponds to a very scrambled mode on the outside of the horizon, the measurement will affect all infalling modes $a$. So by the measurement of the outside observer, there is a change in eigenbasis and the infalling observer will no longer detect the vacuum. He will burn up at the horizon by a firewall.

There is a possible subtlety here. One measurement perturbs a second noncommuting measurement only if the latter is later in time. For local field theories, there is an unambiguous time ordering because operators at space-like separation commute. Here, we have to assign some foliation, and if $\hat{b}$ is effectively ‘earlier’ than $e$, the measurement of $e$ will not perturb it. However, the infalling observer will encounter $e$ before $\hat{b}$, so such a proposal could lead to a closed time-like loop because an observer would first measure $e$ and then find no firewall which implies the state would be back as it was before he made the measurement.

6.9.2 State dependence

The identification $A \subset R$ is based on the the fact that these have the same entanglement with $B$. However, the precise $R - B$ entanglement follows from the unitary evolution of the initial black hole state. So this entanglement depends on this initial state. This might cause problems, even when the Hilbert space is enlarged to contain the Hilbert space of initial black hole states. In this section we follow [168, 169], where explicit constructions of the Hilbert space of an infalling observer have been proposed using the idea that it can be identified by its entanglement. These papers are mainly in the context of stable AdS black holes, but as those authors note the construction extends to evaporating black holes.

Let the index $i$ label the space of initial black hole states $I$, $j$ the states of the early radiation $R$ and $N$ the states of the late radiation $B$ in a Fock basis (the late radiation is denoted by $B$ because it is equivalent to the thermal zone by the mining argument of section 6.2). Given the black hole $S$-matrix $S_{i,j,N}$, a particular initial state $i$ will decay as

$$|i\rangle_I \rightarrow S|i\rangle_I = \sum_{j,N} S_{i,j,N}|j,N\rangle_E,B .$$

(6.92)

Defining

$$|\hat{N}\rangle_{\hat{B}} = Z^{1/2} e^{-\beta E_N/2} \sum_j S_{i,j,N}|j\rangle_E ,$$

(6.93)

where $Z$ is a normalization constant and $\hat{B}$ are the interior partner modes of $B$, the late time state is

$$Z^{-1/2} \sum_N e^{-\beta E_N/2} |\hat{N},N\rangle_{\hat{B},B} .$$

(6.94)

If we identify $|\hat{N}\rangle$ as the Fock states of the interior Hawking modes, this is the infalling vacuum state as required by the equivalence principle. Thus, the identification (6.93) is the desired mapping from $A$ into $E$. 
Having identified the states $|\tilde{N}\rangle$, one can now define interior operators as linear combinations of the $|\tilde{N}\rangle\langle\tilde{N}'|$. For example, for the individual Hawking partner modes $\tilde{b}_k$ we have annihilation and creation operators

$$\tilde{b}_k|\tilde{N}\rangle_{\tilde{B}} = \tilde{N}_k^{1/2}|\tilde{N} - \tilde{k}\rangle_{\tilde{B}} \quad (6.95)$$

$$\tilde{b}_k^\dagger|\tilde{N}\rangle_{\tilde{B}} = (\tilde{N}_k + 1)^{1/2}|\tilde{N} + \tilde{k}\rangle_{\tilde{B}}. \quad (6.96)$$

'Early' and 'late' have been defined such that the dimension of $R$ is much larger than $B$ and $A = \tilde{B}$. As a result, states of the form (6.93) span a low-dimensional subspace of $R$ so (6.95) and (6.96) are an incomplete specification of $\tilde{b}, \tilde{b}^\dagger$ as operators on $R$. One option is to set all unconstrained matrix elements to zero. With this choice, we can fully define $\tilde{b}$ as

$$\tilde{b}_k(i) = \sum_{\tilde{N}} \tilde{N}_k^{1/2}|\tilde{N} - \tilde{k}\rangle \langle \tilde{N}| \quad (6.97)$$

$$= Z \sum_{\tilde{N}} \sum_j \sum_l e^{\beta(EN + E_{N-\tilde{k}})/2} \tilde{N}_k^{1/2} S_{i\tilde{N}N-\tilde{k}}^j \langle j| S_{i\tilde{N}N-\tilde{k}}^* |i\rangle. \quad (6.98)$$

Using (6.92), we find

$$\langle N - \tilde{k}| S| i\rangle = \sum_j S_{i\tilde{N}N-\tilde{k}}^j. \quad (6.99)$$

So by combining (6.98) and (6.99), $\tilde{b}$ can be written as

$$\tilde{b}_k(i) = Z \sum_{\tilde{N}} e^{\beta(EN + E_{N-\tilde{k}})/2} \tilde{N}_k^{1/2} \langle N - \tilde{k}| S| i\rangle \langle i| S^*| N\rangle. \quad (6.100)$$

Here $\langle N - \tilde{k}| S| i\rangle$ is a ket vector in $R$. So for a given initial state (6.100) manifestly maps $R \to R$.

Thus, there is a new problematic feature in that the embedding of the interior Hilbert space in the early radiation depends on the initial state $|i\rangle$: it is not just a mapping from $A \to R$, but from $I \otimes A \to R$, where $I$ is again the space of all initial states. Consequently, operators in the interior become maps from $R \to R$ that depend on the reference state $|i\rangle$. This state-dependence is outside the normal framework of quantum mechanics, and one must argue very carefully that it is consistent.

### 6.9.3 Arbitrariness and energy considerations

Another undesirable feature of the embedding is that the construction of $|\tilde{N}\rangle_{\tilde{B}}$ is dependent on the choice of separation time between the early and late Hilbert spaces. Any choice of more than half the Hawking modes may be used to define an early Hilbert space that has sufficient entanglement to embed $A \to E$ as above, but each leads to a different embedding because $|\tilde{N}\rangle_{\tilde{B}}$ depends by its definition (6.93) directly on $|j\rangle$, where $j$ labels the states of the early radiation.

The nonzero commutator $[e, \tilde{b}]$ has another unpalatable effect. The outside observer can capture $e$ without yet measuring it, and, if there is no firewall, see effects of the nonzero commutator.
when he falls past the horizon where he can directly compare the two. On the other hand, he might instead carry a physical identical bit which he has not captured form the early radiation but from the thermal zone. Call the annihilation operator associated with this second bit $\tilde{b}$. But then, $b$ and $\tilde{b}$ do in fact commute.

Note that an observer follows a time-like path and so encounters the $b, \tilde{b}$ bits at time-like separation. Therefore, causality does actually impose no direct requirement that they commute. But, all these bits are essentially outward-moving functions of the Kruskal-Szekeres coordinate $U$ and so the commutator does not depend on the $V$ value at which the observer measures them. The observer, moving at a geodesic of constant $U$ will encounter the bit $b$ at a larger value of $V$ than when he encounters the bit $\tilde{b}$. Therefore, an order one $[e, \tilde{b}]$ commutator and and a zero $[b, \tilde{b}]$ commutator are in fact inconsistent with local field theory. But of course, an observer can also send out probes to interact with space-like separated bits away from his worldline and then reassemble the results at a later time. In this way he can verify the difference between $[e, \tilde{b}] \neq 0$ and $[b, \tilde{b}] = 0$ at space-like separations.

So, letting the bits be physical particles like electrons, the observer finds that not all electrons are the same. But quantum mechanics does not allow the physics of a bit or particle to depend on either the bits history or on the degree to which it is entangled (of course it can depend on the specific entanglement, e.g. two spin 1/2’s can combine either to spin 0 or to spin 1).

A final objection to the embedding $A \subset R$ comes from energy considerations. The operators $\tilde{b}, \tilde{b}^\dagger$ defined in (6.95) and (6.96) change the energy of the early radiation, whereas the correct behind-horizon operators should change only the energy emitted at late times. Simple observables such as the gravitational field outside the horizon will be sensitive to this distinction.

In [170–172] it was suggested that the operators $\tilde{b}, \tilde{b}^\dagger$ may act non-trivially both on $R$ and on $I$. This can be done by not putting all unconstrained matrix elements to zero and therefore deviating from the form (6.97). In this way the authors tried to evade the state-dependence of the embedding. We might then parametrize the amount of action on $R$ by the expectation values of the commutators of some given $\tilde{b}, \tilde{b}^\dagger$ with the operators $e, e^\dagger$ associated with the early modes. At least one of these must be large if there is a significant commutator of $\tilde{b}, \tilde{b}^\dagger$ with any bit in $R$. But then, one still has the problem that not all physical particles are identical and one has to cope with the above mentioned energy considerations. On the other hand, if $\tilde{b}, \tilde{b}^\dagger$ have small commutators with all bits in $R$, then one may define slightly modified operators $c, c^\dagger$ that precisely commute with all qubits in $R$ and which define approximately the same notion of infalling vacuum as $\tilde{b}, \tilde{b}^\dagger$. But then the result is again an entanglement conflict with unitarity. This is to be expected since these small commutators imply a ‘small’ embedding. But this embedding was introduced to evade the need for double maximal entanglement of the same bit, and therefore evade the need for a firewall, in the first place.

The model of strong complementarity is not directly addressed by the difficulties above. However, it remains to provide a working example that evades these arguments. In particular, the most developed version [173] requires the restrictions of the quantum states to agree in the observers’ common causal past and thus appears to remain in direct conflict with the arguments of section 6.1 concerning the low energy limit.
Also note that we’ve argued in this section that $A \subset R$ is in conflict with effective field theory, which is assumed to be valid outside the stretched horizon. Therefore, these arguments do not apply to the embedding $A \subset H$ of section 6.8.1. This means the concept of the stretched horizon as a hologram of the black hole interior is not endangered, only the embedding $A \subset R$ used to evade the firewall is.

### 6.10 The Hilbert space for an infalling observer

The central question that runs through the alternatives to the firewall is the quantum description of the observations of the infalling observer: what is the nature of his Hilbert space? Here an overview of this question is given.

At some given time an asymptotic observer can observe the joint state of the black hole $H$, some outgoing Hawking modes $B$ emitted around that time, and the previously emitted radiation $R$. In an orthonormal basis for $H \otimes B \otimes R$ we have the state $\psi_{\text{IN}}$.

Now consider an observer who falls into the black hole at around this time. For his observations we need a density matrix for the inner and outer modes $(A = \tilde{B}) \otimes B$.

The natural way to try to relate these two descriptions is to imagine that $A$ is identified with some subspace of $H \otimes R$. Thus, we decompose $H \otimes R = A \otimes A^c$, and in a basis for $A \otimes B \otimes A^c$ the wavefunction is $\psi_{\text{IN}}$. The density matrix on the $A \otimes B$ subsystem is then given by the trace over the unobserved degrees of freedom

$$
\rho_{NN,NN'} = \sum_l \psi_{\text{IN}}^l \psi_{\text{IN}}^l \cdot
$$

(6.101)

If one wishes to avoid a firewall, this density matrix must correspond to the pure infalling vacuum. Thus $\psi_{\text{IN}}$ must factorize as $\phi_{\text{IN}}$, where

$$
\phi_{\text{IN}} = Z^{-1/2} e^{\beta E_N/2} \delta_{NN}.
$$

(6.102)

Now, for any fixed state of the black hole we can find an identification of $A \subset H \otimes E$ for which this is true. However, as we vary over the initial black hole states the necessary identification changes, being related by some generic unitary transformation. Thus, the construction (6.101) does not avoid a firewall, unless we extend the rules to allow the embedding of $B \otimes A^c$ to depend on the state of the black hole. This is a part of the state-dependence that was discussed in section 6.9.2. Another part is that, even when $\phi_{\text{IN}}$ is pure, it will not generally agree with the fixed $\psi_{\text{IN}}$-independent definition (6.102) of the infalling vacuum. The required state-dependence goes beyond the usual rules of quantum mechanics. The consistency of such a modification requires careful considerations.

In [170–172] a somewhat more elaborate construction than (6.101) is proposed. There, the
specific entanglement of $A$ with $B$ depends on the specific state in $E$, which is called the classical world. Thus, the interior mode operator is of the form

$$\tilde{b} = \sum_a P^{(a)} \tilde{b}^{(a)},$$

(6.103)

where the $P^{(a)}$ are projectors acting on the early radiation $R$, and the $\tilde{b}^{(a)}$ are different operators acting on the black hole Hilbert space. So (6.103) associates an interior quantum theory with each classical outside world selected via $P^{(a)}$.

For each classical world there exists a transformation $U^{(a)}$ acting on the initial stretched horizon containing all the information of the infalling matter. So initially, we can write $H = A^c \otimes A$ since the stretched horizon at that point is a complete hologram of the interior region. The transformation $U^{(a)}$ represents the black hole dynamics which transforms the wavefunction for a classical world $\psi_{hNNa}$ in $A^c \otimes A \otimes B \otimes R$ into the factorized form

$$\tilde{\psi}_{hNNa} = \sum_{h'M} U^{(a)}_{hN,h'M} \psi_{h'NNa} = \chi_{ha} \phi_{NN},$$

(6.104)

The authors of [171, 172] propose that the state seen by the infalling observer is $\tilde{\psi}_{hNNa}$, which gives a pure density matrix for $B \otimes A$. Again, this suffers from the state-dependence as above. For different initial black hole states one needs different $U$’s. In particular (6.104) is not invertible, and thus, in spite of appearances, is not actually a unitary transformation on the space of states of the black hole.

The classical-world model is in the framework of an overall Hilbert space, in which the internal Hilbert space is embedded in the radiation. Now let us consider strong complementarity. We again construct a density matrix $\rho_{NN,NN'}$, which at least for a black hole that forms from a collapse should be determined by $\psi_{1Nk}$ living in $H \otimes B \otimes R$. But now $A$ is not considered as embedded in $H \otimes R$. In this framework, one can find a density matrix on $BA$ with a number of good properties

$$\rho_{NN,NN'} = \phi_{NN} \phi_{NN'} + \tau_{NN'} \left( \sum_k \psi_{1Nk} \psi_{1N'k} - \tau_{NN'} \right),$$

(6.105)

where

$$\tau_{NN'} = Z^{-1} e^{-\beta E_N} \delta_{NN'},$$

(6.106)

is the thermal density matrix and $\phi$ is again given by (6.102). The main difference between $\phi_{NN}$ and $\tau_{NN'}$ is that $\phi_{NN}$ lives in a subspace $A \otimes B$ which extends on both sides of the horizon while $\tau_{NN'}$ lives in a subsystem on one side. Therefore, $\phi_{NN}$ is relevant to an infalling observer and $\tau_{NN'}$ to an outside observer. This distinction is possible because we advocate a form of strong complementarity where $A$ is not embedded in $R$.

The density matrix (6.105) has three very promising properties. First, (6.105) is bilinear in $\psi$, as required by the linearity of quantum mechanics. Secondly, tracing out $A$ by summing over $\tilde{N} = \tilde{N}'$, the reduced density matrix on $B$ for arbitrary $\psi$ is the same for the infalling observer as for asymptotic observers. And third, for $\psi$ typical in the microcanonical ensemble,
the difference in parentheses vanishes and the infalling density matrix is the pure vacuum: there is no firewall.

Unfortunately, (6.105) is not positive for general \( \psi \). In particular, if we consider \( \psi \) that has been projected along some subspace of \( B \), then in the subspace of \( AB \) that is orthogonal to both the projection and \( \phi \), only the negative definite \( \tau \tau \) term survives. It is very likely that one cannot improve on this, but it could be a possibly useful expression.

### 6.11 Static AdS black holes

Evaporating black holes provide the sharpest arguments that there is a problem with reconciling unitarity, effective field theory and the equivalence principle. In this section we will, perhaps somewhat surprisingly, argue that a firewall is typical based on a static non-evaporating AdS black hole. The basic tension that is explored is between the equivalence principle and supposing that the black hole is described by a fixed Hilbert space of finite size. This is done using counting arguments which may be considered more or less independent from the arguments of the previous sections.

In AdS/CFT there is a sharp dictionary relating the boundary limits of bulk fields to local operators in the CFT. To extend this further into the bulk requires some form of extrapolation, essentially integrating the bulk field equations. To in this way extend past the horizon of a black hole that is formed from collapse, it is necessary to integrate the field equations back in time prior to the formation of the black hole, and then outward to the boundary [174]. However, the backwards integration produces an exponential blueshift. After a time \( T^{-1} \ln R \), the backward integration depends on unknown trans-Planckian interaction between the Hawking quanta and the infalling body [165]. This implies that we cannot by this means explicitly construct the field operators behind the horizon.

Here we will give a simple argument which indicates the field operators behind the horizon do not exist even in principle. Consider the raising operator \( \tilde{b}^\dagger \) for an interior Hawking mode, which is assumed to have some image in the CFT. Because the partner modes behind the horizon have negative energy, \( \tilde{b}^\dagger \) lowers the global energy by some amount \( \omega \). Now consider all the CFT states which correspond to \( M < E < M + dM \). Here \( M \) is assumed to be the mass of a black hole after the Page time, so that the typical CFT states behave thermally. Take \( dM \) small but large enough so there are many states in the range. Labeling these states by

\[
| i \rangle : \quad M < E < M + dM . \tag{6.107}
\]

Because the black hole is considered to be a system with a discrete number of states given by \( e^S \), the number of corresponding CFT states \( | i \rangle \) is also finite. Now consider the states

\[
\tilde{b}^\dagger | i \rangle : \quad M - \omega < E < M - \omega + dM . \tag{6.108}
\]
In effective field theory, the raising operator $\tilde{b}^\dagger$ has a left inverse

$$\left( \frac{\tilde{b}}{\tilde{b}^\dagger b + 1} \right) \tilde{b}^\dagger = 1,$$  \hspace{1cm} (6.109)$$

where we assumed bosonic behavior. This implies the states in (6.108) must be independent. However, their number is smaller than that of the $|i\rangle$ by a factor $e^{-\beta \omega}$. So we arrive at a contradiction: there is a one-to-one mapping between the states of the two intervals, but because of the thermal behavior the number of states in the lower energy range should be smaller. If the field is fermionic, $\tilde{b}^\dagger$ will annihilate half of the states. However, this would still lead to a problem for modes with $e^{-\beta \omega} < \frac{1}{2}$. Therefore, the operator $\tilde{b}^\dagger$ cannot exist in the CFT.

This result has at least two possible interpretations. Because of the trans-Planckian problem in the original construction of behind-horizon operators, one may say that $\tilde{b}^\dagger$ annihilates states at the UV cutoff of the effective field theory. This means that the redundant states in (6.107) correspond to high energy modes beyond the cutoff in the bulk. Now, $e^{-\beta \omega}$ is $O(1/2)$, so $\tilde{b}^\dagger$ annihilates $O(1/2)$ of all states. So $(\tilde{b}^\dagger)^k$ annihilates a redundant fraction $1 - O(1/2^k)$ of all CFT states. With this interpretation, most CFT states correspond to highly excited bulk modes near the UV cutoff, and so firewalls are typical.

The other interpretation to explore is that the CFT contains an incomplete description of the black hole interior. Indeed, the notion that the CFT described only a subset of the states of the black hole, namely those that could have been formed from collapse, has been expressed before [175]. Since the state created by $\tilde{b}^\dagger$ is trans-Planckian in the past, there is no guarantee that this state can be formed from collapse, and the counting argument shows that in some cases it cannot. Of course, an infalling apparatus could emit a quantum in the mode $\tilde{b}^\dagger$. However, the mass of the apparatus adds to that of the black hole and so the full process is more than the creation of the mode. The $\tilde{b}^\dagger$ are also formed by the Hawking process, but always entangled with the $b^\dagger$ excitations outside so that there is no change in global energy.

On the other hand, an infalling observer who wishes to describe the physics behind the horizon would naturally use low energy effective field theory, including $\tilde{b}^\dagger$. Since evaporation is neglected, we may set aside the concerns of the previous section and take the point of view of strong complementarity. So each observer has its own quantum mechanics, allowing that the external observer can measure $|i\rangle$ but not $\tilde{b}^\dagger$, and the infalling observer can measure $\tilde{b}^\dagger$ but not $|i\rangle$.

Thus, the nonexistence of $\tilde{b}^\dagger$ does not by itself imply the nonexistence of the interior. Curiously, if $\tilde{b}, \tilde{b}^\dagger$ did exist in the CFT, we would immediately conclude that typical states would have a firewall.

### 6.12 Conclusion

By collecting all the arguments of the previous sections we can make a summary about the current status of the firewall-paradox.
It is beyond doubt that the AMPS argument exposes a failure in the original formulation of black hole complementarity of chapter 5. By replacing the artificially introduced entangled spins in the thought experiment of section 5.5.3 with the naturally created Hawking quanta, the authors of [149] stumbled upon a fundamental shortcoming of the principle of complementarity. The equivalence principle, unitarity and low energy effective field theory are incompatible.

In this chapter we’ve never doubted the unitary evolution of a quantum black hole. This is because we already discussed the unitary or non-unitary nature of black holes in the context of Hawking’s original formulation of the information paradox in chapter 4. Systematic elimination of unphysical behavior naturally lead us to the conclusion that black holes must evaporate according to the usual laws of quantum mechanics. The AdS/CFT correspondence greatly favors this conclusion. And of course, whether the reason to doubt unitarity is the original prediction by Hawking or the AMPS argument, the consequences of abandoning this foundation of quantum mechanics remain the same.

The second basic principle which could be given up is the equivalence principle. This is the solution proposed by AMPS themselves. In practice, this means a firewall would replace the horizon. It can either be interpreted as a singularity, destroying the entanglement of the Hawking quanta and their interior partner modes, or as the Hawking quanta themselves, who have exponentially blueshifted energies at the horizon. The firewall is the end of the geometry and an infalling observer is terminated before he can enter the black hole.

If one wants to avoid non-unitary evolution and refuses to give up the equivalence principle, then effective field theory must be modified. All published alternatives to the firewall address this possibility. At the present time there are two important options that have been considered. The first is to use non-local dynamics and the second is to use a second generation form of complementarity.

The models of nonlocal effective field theory allow for information to jump over the thermal zone such that the information transfer becomes harmless to an infalling observer. Although the observer will not see the vacuum on the outside of the thermal zone, this doesn’t alarm him since there is no violent blueshift in the energy of the Hawking quanta. So these models predict a deviation from the conventional observations for a freely falling observer, but in such a way that he doesn’t burn up. However, there is a conflict between these models and a mining experiment. A conspiracy between Planckian physics at the stretched horizon and low energy dynamics outside the thermal zone is required to build a consistent model. This presents a severe difficulty for any model trying to add nonlocality to black hole evaporation. At the moment, no mechanism is found to circumvent this problem.

Other modifications of effective field theory can be combined as ‘next generation complementarity’. The first goes under the name of standard complementarity and is an extension of the complementarity principle of section 5, based on the Harlow-Hayden conjecture which states that it impossible to extract an entangled bit out of the Hawking radiation in a time shorter than the black hole evaporation time. Unitarity and the equivalence principle require a bit in the thermal zone after the Page time to be maximally entangled with a bit in the Hawking radiation and a bit in the interior of the black hole. Because no observer can encounter both bits,
no detectable violation of the laws of nature occurs when they are identified as the same bit. This operational point of view removes the double maximal entanglement and therefore evades the need for a firewall. The identification of the two bits effectively comes down to embedding the interior Hilbert space into the radiation Hilbert space. But this embedding has been seen to lead to dramatic conflicts with low energy effective field theory. In particular, there is a dependence of the evolution on the initial state of the black hole which in contrast with the usual rules of quantum mechanics. Also, because of the scrambling process taking place at the stretched horizon a measurement done by an outside observer completely destroys the vacuum for an infalling observer. So the embedding is in conflict with the usual quantum mechanical evolution, is highly unstable and is completely arbitrary on top. It is very unlikely that this scenario could be made viable after all.

The second form of next generation complementarity is called strong complementarity. The difference with standard complementarity is that it doesn’t use an embedding of the interior Hilbert space. Instead, each causal patch has its own quantum mechanics and therefore its own Hilbert space. Coarse grained observables corresponding to different observers must be the same in the overlapping parts of their patches. In the context of an evaporating black hole strong complementarity also relies directly on the Harlow-Hayden conjecture to evade quantum cloning. Although there are no direct inconsistencies of this model, it has been critized by a number of authors because it is very vague and seems to be 'made up' [160, 167]. Each observer having its own description of the universe, approximate or not according to taste seems like a rather inelegant framework. It also remains to provide a concrete working example of strong complementarity that succeeds in evading the need for a firewall.

Although we’ve not considered the fuzzball model explicitely, we can mention for completeness that also fuzzball complementarity as proposed in [176] suffers from the same fatal problems as normal complementarity [167].

So at the end, very few authors accept the existence of the firewall but none of them have succeeded in providing a consistent and working alternative. It becomes increasingly more unlikely that there is just some basic feature that has been overlooked. The firewall paradox poses a real challenge to those trying to reconcile unitarity with black holes. The Higgs particle may be found this year, but maybe black holes can cause the necessary and fruitful commotion to clear the way to a deeper understanding of the quantum world.

6.13 * Personal view

In this concluding section I will take the liberty to express my personal view on the firewall controversy. It should be noted that the ideas presented here are entirely my own and do not by any means represent conceptions accepted by the scientific community.
6.13.1 A firewall?

Although I strongly believe that AMPS make a valid point, I don’t think there is a firewall. Instead of being wrong, I think black hole complementarity is incomplete. The semiclassical framework predicts evolution from pure states to mixed states, but nowadays this does not convince anyone anymore that quantum gravity is non-unitary. I think the firewall paradox is a result of the same insufficient, semiclassical description of the evaporation process.

The firewall paradox leads to the revival of an old question: how do we get the information out of a black hole in the semiclassical picture? Black hole complementarity developed a consistent quantum description for an outside observer based on the membrane paradigm in general relativity, and then simply added the equivalence principle. However, placing a stretched horizon at the Planck scale and then allowing conventional quantum field theory beyond this distance from the horizon appears to be too simple. Things are more subtle than that. In my opinion the firewall paradox arises because one simply imposes unitarity on a framework that predicts there isn’t. This leads to an internal inconsistency as shown by the AMPS argument.

Another reason to doubt the physical existence of the firewall is that it is in sharp contrast with the usual expectation that a proper quantum treatment of gravity will remove the black hole singularity. But instead of making the theory singularity-free, quantum mechanics would seem to present a singularity right in our face at the horizon under the form of a firewall. So contrary to the black body spectrum, we can not rely on quantum mechanics to remove what we regard as unphysical behavior. I find this very unsatisfying.

Furthermore, it was shown in chapter 1 that classical black holes seem to anticipate rather accurately on what will happen if one starts to do quantum mechanics around them. The thermal character of the Hawking radiation is already present in general relativity. Also the stretched horizon has deep foundations in the classical theory. So before the firewall there was a ‘smooth’ transition from the classical description to the quantum description. This perfect match of the two theories would be violently interrupted by the sudden pop-up of a singularity at the horizon in the quantum description, while general relativity predicts it should be a smooth region of spacetime.

And to make things even worse, accepting the firewall as physical reality has the profound consequence that we lose the cosmic censorship hypothesis. One could try to avoid this claim by saying that the firewall actually lies at the apparent horizon, but in any way its existence is against the spirit that all singularities are well hidden behind a horizon. A naked singularity would make a black hole spacetime non-predictable. And as seen in section 1.11.3, predictability is a necessary condition to prove the area theorem. So in some way, a firewall would undermine the thermal framework which lead to its existence in the first place.

The AMPS argument is an inevitable failure of the semiclassical framework. But solving it by keeping unitarity for an outside observer and letting an infalling observer burn up leads to an unequal treatment of different observers. In fact, one could even say it introduces a preferential class of observers, namely the outside observers. But this runs against the very heart of general relativity. In some sense, this would put the clock back to an ether-scenario. Preskill
stated that the firewall puts us 40 years back in time, right to where we were when Hawking first proposed the information paradox. I would even say that accepting a firewall and its associated preferential observers would take us back to the time of Maxwell.

I think the firewall paradox simply repeats what we already knew for a long time: reconciling general relativity with quantum mechanics is difficult. There is no straightforward unification of the two theories and progress in this area requires new principles and insights.

### 6.13.2 Backreaction

In section 4.9 it was shown that every horizon can be approximated by a Rindler horizon. But a true Rindler horizon is not a special place, there is no firewall paradox for Rindler spacetime. And although the Unruh effect takes place in flat spacetime, it is in fact entirely the same mechanism which produces the Hawking radiation in a black hole spacetime. But a black hole horizon does have problems with non-unitarity and firewalls. So maybe we can learn something by really pointing out the difference between the two situations?

The obvious difference is of course that the black hole geometry suffers from backreaction effects. The original AMPS paper contained one short paragraph about general horizons. Their claim was that Rindler horizons represent a black hole of infinite mass and therefore do not possess a firewall because they never get old. But saying something has infinite mass is of course equivalent to saying that it is not influenced by backreaction.

Through the years, many people started to believe that backreaction is the necessary ingredient to make a black hole return the information about the collapsed state. So by the reasoning of the previous section one may then think that an appropriate treatment of backreaction effects can also remove the need for a firewall? In any case, it is clear from section 5.8 that backreaction effects have the potential to drastically change the semiclassical viewpoint, which may appear to be very misleading.

The usual argument for the validity of results from quantum field theory in curved spacetime is that it is only being used in regions with low curvature. There is no reason to doubt this argument, but it may not be the full story. It only involves the effect of the geometry on the behavior of the quantum field. To close the loop, one should also consider the influence of the quantum field on the geometry. This is much more speculation since there is no known quantum source of gravity. The conventional procedure is to take the expectation value of the energy-momentum tensor of the field and put this quantity in the Einstein equations. But it may well be that one fails to capture an essential physical feature by this procedure. Of course, the backreaction in black hole evaporation takes place on a timescale much larger than the time needed for the evaporation of a single Hawking particle. But the information paradox is phrased on timescales of the black hole lifetime, so it can no longer be neglected.

In other words, I think the absence of backreaction is what causes most of the trouble. However, many authors think the resolution of the firewall controversy will come from operational constraints in terms of computational complexity. Especially for strong complementarity, seemingly
Chapter 6. *The Firewall*

the only viable remainder of next generation complementarity, this is an indispensable feature.

I have mixed feelings towards the relevance of the Harlow-Hayden conjecture in the firewall paradox. I favor the operational point of view because it is very intimately related to gravity. It has been very useful in the past to find a proper treatment of gravity. As well known, it was an operational point of view which lead Einstein to the equivalence principle: there is no experiment an observer can do to distinguish an accelerating frame and a gravitational field. Of course this does not imply the operational point of view will again deliver the solution here, but it does show that it should be taken seriously.

On the other hand, I feel that the operational statement made by the Harlow-Hayden conjecture is of a completely different nature than the one used by Einstein. In general relativity, the operational point of view lead to a fundamental equality, a founding principle of the theory: the equivalence principle. However, in strong complementarity the operational constraint does not lead to such a fundamental equality. Different observers have different fine-grained quantum states in their separate quantum descriptions, only no single observer will ever notice that. In this sense I consider the use of the Harlow-Hayden conjecture as an act of despair. It says something like: out theory is inconsistent but since we will never notice that we shouldn’t worry about it. I think the theory should be consistent, observable or not.

Another problem I have with assigning each patch its own quantum description is that the state of the Hawking radiation depends on what an observer will do in the future. The fine-grained properties are determined by wheter or not the observer decides to jump into the black hole at a later time. Also, strong complementarity seems such a wasteful and inelegant solution. I think it would be surprising if nature had chosen for such an inefficient model. I agree with D. Stanford who states in [160] that strong complementarity seems to be made up.

6.13.3  *Freely falling vs. hovering*

The AMPS argument does not state that the the description of an outside observer in black hole complementarity is inconsistent. The problem is in the connection between an outside, hovering observer and an infalling observer. In every thought experiment of black hole complementarity and the firewall paradox, there is a need for both a hovering and a freely falling observer. All thought experiments involve comparing information in the Hawking radiation and the interior modes. Since only a hovering observer detects Hawking radiation it is immediately clear why their role is indispensible. On the other hand, we know that the proper acceleration of a hovering observer becomes infinite at the horizon. So for an observer to hover at the horizon it requires an infinite force to hold him in place. Therefore, each thought experiment which involves information inside the black hole also needs a freely falling observer.

Now consider an observer on a spherical shell of mass with radius $L > 2M^3$. The spherical shell starts to contract. As long as the shell does not reach its Schwarzschild radius, the observer hovers at constant spatial position. When the horizon forms after finite proper time, he starts to fall freely. In this way, he never detects a single Hawking particle emitted by the black hole. But because the evaporation time of a black hole is $\sim M^3$, the black hole will
be completely evaporated at the time he reaches its former center. He will continue to keep falling but never detect one remainder of the shell he witnessed to collapse. So not only is there a loss of information, there is also a loss of energy. The freely falling observer will conclude, based on the lack of Hawking particles and the absence of a singularity, that no black hole has formed. However, if he would have decided to keep hovering at the initial position, he would have concluded a black hole has formed based on the detected Hawking particles. One could argue that the thought experiment described here is undermined by the fact that the horizon forms before the shell reaches its Schwarzschild radius, but it is obvious that this problem can easily be circumvented by simply taking $L$ bigger.

The irony of the situation is that the absence of the singularity which causes the trouble. The singularity is an obstacle in the sense that it indicates a failure of the classical theory, but it actually also is very convenient since it makes most of the semiclassical descriptions consistent. Because if one neglects backreaction for a moment, the black hole would never disappear and the freely falling observer would hit the singularity, thereby knowing that a black hole was formed. He would be destroyed by tidal forces and so the problem that he could detect a loss of information or energy is also resolved.

Different possible outcomes of a same event are inherent to quantum theory. But here the situation is different in the sense that the different outcomes (a black hole or no black hole) are related to a particular kind of observer. Normally, nature just rolls a dice, assigning each outcome a certain probability, and an observer simply has to see what the outcome of his measurement will be. Here however, there is a one-to-one mapping between the different outcomes and the type of observer. This a completely different kind of indeterminacy of measurement-result. In my opinion, this again underlines the fundamental difference between a freely falling and a hovering observer.

An interesting variant of the thought experiment above is to take an $s$-wave electron instead of a spherical shell of matter. Electrons are point particles and have a mass, so by definition they would be able to form a black hole. The probability for this to happen is course tremendously small, but we just want to know what happens next in the rare cases it does. If one does not believe a single particle can form a black hole, just consider the minimum number of particles one thinks it does take to form a black hole. Then to the hovering observer, the electron will have formed a black hole and will subsequently have evaporated. To the freely falling observer, the electron will simply have disappeared.

As already mentioned, all problems of the firewall-paradox arise when one tries to combine the experiences of an infalling observer and a hovering observer in one global picture. In some way, I think the paradox should not come as a surprise. In my opinion, the existence of a physical difference between freely falling and hovering observers is one of the defining features of gravity. Knowing all the trouble people have had in the past (and even today) with reconciling gravity with quantum theory, I would find it surprising if this fundamental aspect of gravity, I would go even further and call it the very heart of general relativity, could be implemented by simply putting a quantum field in a curved background. I think if it would really be that simple, we would already have had a quantum theory of gravity in a very long time. Therefore, I feel that the equivalence principle, one of the founding principles of general relativity, cannot
be realized in quantum theory without first obtaining some new insight about the structure of quantum gravity.

Maybe it could be instructive to look at particle complementarity. There, a particle can possess only a limited amount of information. One has to make a choice between momentum and position, it is impossible to know both to arbitrarily high precision. In the mathematical framework of quantum mechanics this is realized by the fact that position and momentum operators do not commute. This seems to suggest a possible way out for the firewall paradox. Assume for a moment that field operators assigned to freely falling and hovering observers do not commute. That would make it impossible to simultaneously know the outcomes of measurements performed by both types of observers. Since the whole firewall controversy results from comparing measurements of freely falling and hovering observers my first idea was that this could resolve the firewall paradox.

But of course it isn’t that simple. When assigning non-commuting field operators to freely falling and hovering observers, one again encounters the problem that measurements by a hovering observer create a firewall. If a freely falling observer detects the vacuum state and another observer which is hovering outside makes a measurement, the freely falling observer would lose his vacuum and burn up at the horizon. This situation is completely analogous to the Stern-Gerlach experiment where a measurement of the $x$-component of a spin destroys the information of an earlier measurement about the state of its $z$-component. So if the freely falling observer would correspond to the $\hat{S}_z$ operator and its vacuum to $|\uparrow\rangle_z$, then a hovering observer’s measurement corresponding with $\hat{S}_x$ destroys the state $|\uparrow\rangle_z$.

Based on this reasoning and the arguments of section 6.9 it seems to me that it is impossible to achieve a consistent description, explaining the equivalence principle and unitarity, in a single Hilbert space. In this way one is lead naturally to strong complementarity which assigns a different Hilbert space to each causal patch. However, as already argued in the previous section I don’t think strong complementarity is a good resolution of the firewall paradox. So based on the arguments above, which stress the fundamental difference between freely falling and hovering observers, I suggest to assign different Hilbert spaces to both types of observers. Although this may seem a quite logical option to consider, it actually is not done so far because most authors argue that it would undermine the derivation of the Hawking radiation. It is true that the derivation of the Hawking radiation is based on relating annihilation operators of different observers which work on the same vacuum state by a Bogoliubov transformation. But it is important to realize that this only happens at spatial infinity. Because black hole spacetimes are asymptotically flat the comparison of freely falling and hovering measurements in the same Hilbert space happens only in flat spacetime. So strictly speaking, the Hawking derivation does not exclude a different Hilbert space for freely falling and hovering observers in curved regions where gravity is present.

With my present state of knowledge I can’t make the difference between the freely falling description and the hovering description more concrete. But any way, I like the simplicity of the idea. I find it less random than the original formulation of strong complementarity which assigns a different Hilbert space to each causal patch.
6.13.4 **Global vs. local**

In this section I will consider another possibility to use different Hilbert spaces for gravity. I think there is something to be learned from the fact that all thermodynamic properties of a black hole spacetime manifestate themselves in global descriptions. First of all, in the classical description of black holes, entropy is associated with the event horizon which is a truly global object and has no local physical significance. Also, for the derivation of the Hawking radiation in section 2.3.1, the structure of the entire spacetime was important. It used radial null geodesics which extend from $I^+$ to $I^-$. So the origin of the Hawking radiation cannot be traced back to some local mechanism, contrary to the emission of light by atoms for example. And finally, in section 4.9, the thermodynamic nature of horizons is exposed by tracing out the unobservable modes behind the horizon. But I don’t see how this could have any local significance to a distant observer. The point I’m trying to make is that all the conventional thermal properties of black holes associated with outside/hovering observers are based on global descriptions. On the other hand, the Minkowski-vacuum experienced by a freely falling observer is something very local. These two descriptions use a very different approach to describe the same reality. Therefore, a possibility to consider is that they are dual and are realized in different Hilbert spaces. In the remainder of this section I will try to examine a concrete idea of how to realize such a duality between local and global descriptions. I will argue it very extensively. This does not mean I am convinced it is true, the only reason for the arguments below is to motivate my line of thought.

The conventional interpretation of the Einstein equations is that they allow extrapolation in time. Suppose one has a matter distribution on some space-like 3-space. Using these initial data one can then integrate the general relativistic differential equations to obtain the entire 4-manifold. This is usual viewpoint of Cauchy surfaces being evolved forward in time. But why give the time-dimension a special treatment? It’s intuitive of course, but maybe it limits our viewpoint on gravity. One could also take the initial data on a spatial 2-space, but over the entire time-range. In this case a better name for initial data is boundary data. Integrating the Einstein equations then extends these data in a spatial direction. Of course this will not work for every boundary-data space. For example, a logical necessary condition would be that every member of a complete family of causal curves has to intersect the boundary-data slice just once. This way of using the Einstein equations is manifestly global.

To proceed I will first return to the concept of a hovering observer to make the link to the conventional interpretation of black hole complementarity. There, the hovering observers are associated with the thermal properties of a black hole spacetime. A hovering observer follows an orbit of the time-like Killing vector field $\partial/\partial t$, where $t$ is the Schwarzschild time. So his worldline is given by $x, y, z = \text{constant}$. Because the metric has additional rotational symmetry, this introduces an equivalence class of hovering observers given by $r = \text{constant}$. So each equivalence class lives in a space which effectively has $d - 1$ dimensions. Using the reasoning above, one could now define boundary data on this $d - 1$ dimensional space and then integrate the Einstein equations to obtain the entire manifold. Now what if one defines a gravitationless classical field theory on this $d - 1$-dimensional space? This would provide us the boundary data which can be spatially extended by the Einstein equations.

This is all classical reasoning. But in the end we would like to learn more about quantum
black holes. Although I’m not aware of the precise technical details, I know there exists something like the AdS/CFT correspondence, which states there is a duality between a gravity theory in the $d$-dimensional AdS bulk and a theory without gravity at the $d - 1$ dimensional AdS boundary. So based on the AdS/CFT correspondence, a possibility to extend the global view on the Einstein equations to quantum mechanics is to assign each boundary space its own Hilbert space, with its own operators. These operators can be mapped to the operators outside the boundary data space, which can be seen as the 'bulk'. So both set of operators describe the same physics with different variables. In this way, one obtains a natural construction of a 'boundary' and a 'bulk', which have dual descriptions. There is no need for an artificial AdS boundary. Of course the $d - 1$ dimensional boundary space is not a true boundary of the manifold because there is an inner and an outer 'bulk' region. However, in the Schwarzschild case this should not form a drastic revision of the concepts because the all the mass is located at the center.

The spacetime of the global theory effectively has $d - 1$ dimensions, so it’s holographic. Because the thermodynamical properties reveal themselves in the global description of Schwarzschild spacetime it is the holographic theory which is thermal. The spacetimes of the different holographic boundary-data theories are all $S^2 \otimes \mathbb{R}$. The smaller the radius $R$ of $S^2$ in the larger 4-dimensional spacetime, the higher the temperature in the corresponding thermal description must be. This follows from the standard expression for the proper temperature

$$T = \frac{\kappa}{2\pi|\xi|} = \frac{\kappa}{2\pi} \left( 1 - \frac{2GM}{R} \right)^{-1},$$

where the second equality is valid for Schwarzschild spacetime. Because of the higher temperature, the entropy density also increases when $R$ decreases. One could argue that the total entropy in each space $S^2 \otimes R$ should be the same since every holographic theory describes the same black hole, with the same energy $M$. The critical radius is the Schwarzschild radius. At that point the critical entropy density $1/4G$ is reached, i.e. the entropy bound is saturated. For $R$ smaller than $R_s$, there can be no dual global description constructed, i.e. the mapping from the bulk Hilbert space to the boundary data Hilbert space breaks down. This reasoning suggests some reversed logic; the radius at which the entropy bound is saturated implies a breakdown of the global description and it thereby defines a lightsheet which we know as the horizon. This description naturally explains why a horizon is a global object which has no local physical significance for a freely falling observer. And moreover, it explains why entropy can be associated with such a global object. It also keeps the idea of black hole complementarity where the horizon is interpreted as a full hologram of the interior region.

A possible alternative interpretation for the breakdown of the thermal description and the associated existence of a horizon is the following. From (6.110) it follows that the proper temperature becomes infinite at the horizon. It was shown in section 4.1.2 that the $\beta \to 0$ limit of a thermal ensemble lead to a maximally random system. So at the horizon, the thermal description cannot 'get more thermal'. At the point it becomes maximally random, the holographic description must break down.
So to summarize, the two dual descriptions are characterized as follows. The local description is the most intuitive one when it comes down to considering the observations made by a single observer. It is also the most natural one for freely falling observers since locally they experience the Minkowski vacuum and their theory contains no gravity. This construction is consistent with the expected low energy limit in the sense that each freely falling observer will be able to describe the measurements in his local Minkowski space via conventional effective field theory. The other, dual description is global and is the natural framework for the conventional thermal properties of a black hole. It is inspired on classical general relativity and AdS/CFT.

One could nevertheless argue that the thermal properties of a black hole do have a local meaning. Because a hovering observer has a proper acceleration he will detect a thermal bath according to the Unruh effect. That is true, but I don’t think this thermal bath has any relation to black hole thermodynamics. The Unruh effect will take place locally in every spacetime, not only in black hole spacetimes. And as far as I know, only a black hole spacetime has thermal properties at the classical level. Another way to argue that black hole thermal effects have no local significance is to adopt the natural frame of the local description: that of a freely falling observer. A freely falling observer can decide to start accelerating in any direction of his local Minkowski spacetime, and with arbitrary magnitude. The Unruh effect will take place at all times. But it is only when the freely falling observer decides to accelerate in exactly the right direction (radially away from the mass) and with exactly the right magnitude that he becomes a hovering observer following an orbit of $\partial/\partial t$. At the point he does this, why would he suddenly no longer detect random thermal radiation, but the Hawking radiation which contains subtle correlations revealing information about the matter that collapsed to form the black hole? And also, from a local viewpoint, what does it mean to ‘stay in place’? Doesn’t it also require some knowledge about the global spacetime to stay at constant Schwarzschild coordinates?

The use of two dual descriptions which distinguish local and global properties has some benefits which I will explain in the following paragraphs. As explained in section 2.5.2, an important ingredient in the interpretation of entropy is the ‘ergodic principle’ which states the equivalence of time averages and phase space averages. Because of the ambiguous meaning of ‘time’ in general relativity, this presents a severe difficulty for the interpretation of black hole entropy as conventional thermodynamical entropy. But the thermal theory referred to above is defined only on $S^2 \otimes \mathbb{R}$, so there is a natural definition of time which allows for a conventional interpretation of black hole entropy.

The global description could also provide a natural connection between the uniqueness theorems of section 1.8.1 and the thermodynamical properties of black hole spacetimes. The holographic theory is defined on a manifold which is spatially compact because of the rotational symmetry of the Schwarzschild black hole. This is because the orbits of the two Killing vector fields $\partial/\partial \phi$ and $\partial/\partial \theta$ define the sphere $S^2$. In some way it is very natural to assume that the thermality of the holographic theory is a result of this spatial compactness. Because no perturbation can escape to infinity repeated internal interaction will cause the system to equilibrate at some thermal state. So for the holographic description of a general stationary black hole to be thermal, it is no wild assumption that at least one of the two Killing vector fields $\partial/\partial \phi$ and $\partial/\partial \theta$ should remain. This implies that for a stationary black hole to have a dual thermal description, it should be spherically symmetric (Schwarzschild) or axi-symmetric (Kerr). So in this
way we automatically get a link between the no-hair conjecture and black hole thermodynamics.

Another advantage of this global/holographic description is that it removes the counting problem of section 6.11. Because the mapping of the holographic Hilbert space to the one of the local description breaks down when $R$ is smaller than $R_s$, the annihilation operator of an interior field mode has no dual which acts on the states in any holographic Hilbert space. The only thing which will happen when interior negative energy quanta are created is that the entropy bound at the former horizon will no longer be satisfied so that the mapping can be extended a little bit more towards the center.

There is of course still the problem of what exactly happens to a freely falling observer who enters the black hole. Based on the reasoning above I think the horizon is inherent to the global description, it has no relevance in the local description of a freely falling observer. In the same line of reasoning I consider the singularity to have no relevance in the global description since it breaks down at the horizon. In my opinion the singularity is what represents the mysterious backreaction process. The extremely high density of the collapsed mass requires theories beyond the standard model like string theory to describe its evolution. In the viewpoint of an infalling frame, the collapse would create a highly excited state which subsequently decays via gravitational interactions. A possibility would be that this leads to some non-local dynamics, mixing the interior degrees of freedom with the horizon degrees of freedom, as suggested in [177]. This non-local dynamics would also explain the fast-scrambling behavior of the stretched horizon and therefore again provide us with a quantum mechanical origin of the no-hair conjecture, this time from the local point of view.

To conclude I will shortly summarize the two ideas presented above. To preserve the equivalence principle and unitarity, two modifications of the conventional picture of time evolution of effective quantum field theory along a foliation of Cauchy surfaces is presented. The underlying reason for this is to disentangle the freely falling vacuum-observation and the hovering thermal description. Because it is the combination of these two features in one global picture that leads to a conflict with unitarity. The arguments above and those of section 6.9 show that it is almost certainly impossible obtain a consistent description in one Hilbert space. Therefore, the two ideas discussed above use different Hilbert spaces. In the first, the different Hilbert spaces were assigned to freely falling and hovering observers. This is a modification of strong complementarity which in my opinion is physically more plausible. In the second, different Hilbert spaces were used to construct a local and a global description. In both cases the same reality is described by two different sets of operators.

So this is what I think with my present, and of course severely limited, state of knowledge. I am well aware of the fact that there are a lot of words in this section but not much mathematics. It is very well possible that my reasoning contradicts some principle or result which I have not yet encountered in my short period of studying this matter.
Appendix A

Frobenius’s theorem

In this appendix, a very powerful theorem of differential geometry, which concerns foliations of the manifold under consideration, is formulated.

The set-up is as follows. At each point $m$ of a $n$-dimensional manifold $M$, we specify a subspace $W_m \subset T_m M$ of the tangent space $T_m M$ in the point $m$. The dimension of $W_m$ is $r < n$. The collection of all $W_m$ is denoted by $W$ and the map $D : M \rightarrow W, m \mapsto W_m$ is called a distribution. In the following we only consider differentiable distributions which means that $W_m$ has to vary smoothly with $m$ in the sense that for each $m \in M$ one can find an open neighborhood of $m$ such that in this neighborhood, $W$ is spanned by $C^\infty$ vector fields.

A differentiable distribution is said to be integrable if in every $m$ there exists an embedded $r$-dimensional submanifold $S \subset M$ such that the $r$-dimensional tangent space to this submanifold in each point $s \in S$ coincides with $W_s$. So actually, stating the existence of an integrable distribution comes down to stating the existence of a smooth foliation of the manifold in terms of disjoint submanifolds (= hypersurfaces). If the subspaces $W$ are one-dimensional, this problem reduces to that of finding integral curves of a smooth vector field.

A differentiable distribution is involutive if the $C^\infty$ vector fields $X^{(1)}, X^{(2)},..X^{(r)}$ spanning $W$ in an open neighborhood of $m$ have the property that

$$\left[ X^{(i)}, X^{(j)} \right] = \sum_{k=1}^{r} c^{ij}_{k} X^{(k)},$$

(A.1)

where $\left[ , \right]$ denotes the Lie bracket and $c^{ij}_{k}$ are some constants. This effectively implies that for every $X^{(i)}$ and $X^{(j)}$ it holds that $\left[ X^{(i)}, X^{(j)} \right] \in W$.

Now Frobenius’s theorem states [178]:

**Frobenius’s theorem**  A differentiable distribution is integrable if and only if it is involutive.

Frobenius’s theorem also has a dual formulation in terms of one-forms, which are the dual
Appendix A. Frobenius’s theorem

Frobenius’s theorem

278

elements of vector fields. In the case of a (pseudo-) Riemannian manifold this are just the
covariant vector fields. From now on, latin indices will be used to label a vector field or a
one-form, while greek indices will be used to denote the components. Consider the one-forms
\( \alpha \in T^*_mM \) which satisfy

\[
\alpha(X^{(j)}) = \alpha_{\mu} X^{(j)\mu} = 0 \quad \text{for all } j \in \{0, 1, \ldots, r\},
\]

where the \( r \) vector fields \( X^{(i)} \) again span \( W \) in an open neighborhood of \( m \). It is clear that these
one-forms span a \((n-r)\)-dimensional subspace \( V^*_m \subset T^*_mM \) of the dual tangent space in \( m \).
Conversely, an \((n-r)\)-dimensional subspace \( V^*_m \) of \( T^*_mM \) defines an \( r \)-dimensional subspace \( W_m \)
of \( T_mM \) via equation (A.2). Thus, the question of integrability can be reformulated in terms of
\( V^*_m \): Under what conditions does a smooth map of \( M \) to \( V^*_m \), associating with each point of the
manifold a \((n-r)\)-dimensional subspace of one-forms, have the property that the via equation
(A.2) associated tangent subspaces \( W \) admit integrable submanifolds?

According to Frobenius’s theorem, integrable submanifolds will exist if and only if for all \( \alpha \in V^*_m \)
and all \( Y, Z \in W \) so that \( \alpha(Y) = \alpha(Z) = 0 \), one has

\[
\alpha([Y, Z]) = \alpha_{\mu}[Y, Z]^\mu = 0.
\]

To see what this implies for \( \alpha \), one uses the expression for the Lie bracket in terms of an arbitrary
derivation operator \( \nabla_\nu \) to write (A.3) as \([11]\)

\[
0 = \alpha_{\mu}(Y^\nu \nabla_\nu Z^\mu - Z^\nu \nabla_\nu Y^\mu) = -Z^\nu Y^\mu \nabla_\nu \alpha_{\mu} + Y^\mu Z^\nu \nabla_\nu \alpha_{\mu} = 2Y^\mu Z^\nu \nabla_{[\nu, \alpha_{\mu}]},
\]

where the brackets denote the anti-symmetric part. Because \( Y \) and \( Z \) are in the subspace
of \( T_mM \) annihilated by the elements of \( V^*_m \), expression (A.4) can hold only if \( \nabla_{[\nu, \alpha_{\mu}]} \) can be
expressed as

\[
\nabla_{[\nu, \alpha_{\mu}]} = \sum_{i=1}^{n-r} \omega_{[\nu, \beta^{(i)}]} \alpha_{\mu},
\]

where each \( \beta^{(i)} \) is an arbitrary one-form and each \( \omega^{(i)} \in V^*_m \). Thus, Frobenius’ theorem can be
reforomulated in terms of differential forms as follows:

\textbf{Frobenius’s theorem (dual formulation)} Let \( D^* : M \to V^*, m \mapsto V^*_m \) be a differentiable map which associates with each point of the manifold a \((n-r)\)-dimensional subspace \( V^*_m \)
of the dual tangent space \( T^*_mM \). Then the associated distribution which maps every point \( m \) to
the \( r \)-dimensional subspace \( W_m \) of \( T_mM \) defined by \( \forall X \in W_m : \alpha(X) = 0, \forall \alpha \in V^*_m \) is integrable
if and only if for all \( \alpha \in V^* \) it holds that \( d\alpha = \sum_i \omega^{(i)} \wedge \beta^{(i)} \), where each \( \omega^{(i)} \in V^*_m \).

where \( d\alpha \) is the exterior derivative of \( \alpha \), given by the left hand side of (A.5), and \( \wedge \) denotes the
anti-symmetric (or wedge) product.

The dual formulation of Frobenius’s theorem gives a useful criterion for when a vector field
\( \xi \) is orthogonal to a hypersurface. Let \( V^* \) be the one-dimensional subspace spanned by the one-form \( \xi^\mu = g_{\mu\nu} \xi^\nu \). Intuitively, one can look at the situation as follows. Consider \( \xi \) as defining a certain 'direction'. Then \( W \) is defined by all the vector fields \( X \) satisfying \( \xi^\mu X^\mu = 0 \), so \( W \) can be seen as a 'plane' orthogonal to the direction of \( \xi \). This 'plane' is the tangent space of a \((n-1)\)-dimensional submanifold in every point of \( M \). If these submanifolds form a smooth and disjoint foliation of the manifold \( M \), as is the case for a one-parameter family of hypersurfaces, then Frobenius’s theorem implies it should hold that

\[
\nabla_{[\mu} \xi_{\nu]} = \xi_{[\mu} v_{\nu]},
\]

where \( v \) is some covariant vector field. Multiplying both sides of (A.6) with \( \xi_\sigma \) and antisymmetrizing in the indices leads to the equivalent result

\[
\xi_{[\mu} \nabla_{\nu} \xi_{\sigma]} = 0.
\]

Frobenius’s theorem can also be used in the reversed direction, so it follows that if (A.7) holds for a certain vector field, then it is orthogonal to a family of hypersurfaces.

Finally, it should be noted that the results above were derived locally, i.e. within a given chart. So the conclusions of this section are also valid if the manifold obeys the less restrictive condition that it can be foliated smoothly into disjoint submanifolds in a certain open neighborhood.
Appendix B

Surface gravity of a Kerr black hole

The surface gravity is calculated from the formula

\[ \xi^\nu \nabla_\nu \xi^\mu = \kappa \xi^\mu \quad \text{on} \quad r = r_\pm, \quad (B.1) \]

where \( \xi \) is the Killing vector field of the Kerr spacetime

\[ \xi = \frac{\partial}{\partial v} + \frac{a}{r^2 + a^2} \frac{\partial}{\partial \chi}. \quad (B.2) \]

So we get

\[ \left( \nabla_v + \frac{a}{r^2 + a^2} \nabla_{\chi} \right) \xi^\nu = \kappa \xi^\nu. \quad (B.3) \]

In this calculation \( \xi^v = 1 \) will be chosen for \( \xi^\nu \). Because the partial derivative part of the covariant derivatives vanishes, \( (B.3) \) becomes

\[ \kappa = \left( \Gamma^{v}_{v \nu} + \frac{a}{r^2 + a^2} \Gamma^{v}_{\chi \nu} \right) \xi^\nu \big|_{r = r_\pm}, \quad (B.4) \]

so because of \( (B.2) \) this is

\[ \kappa = \Gamma^{v}_{v v} + \frac{2a}{r^2 + a^2} \Gamma^{v}_{v \chi} + \frac{a^2}{(r^2 + a^2)^2} \Gamma^{v}_{\chi \chi} \quad (B.5) \]

evaluated at \( r = r_\pm \), where \( \Gamma \) are the Christoffel symbols. They are given by

\[ \Gamma^\mu_{\nu \rho} = \frac{1}{2} g^{\mu \lambda} (\partial_\nu g_{\lambda \rho} + \partial_\rho g_{\lambda \nu} - \partial_\lambda g_{\nu \rho}). \quad (B.6) \]

So using the fact that the metric coefficients of the Kerr black hole are independent of \( v \) and \( \chi \), \( \Gamma^v_{v v} \) becomes

\[ \Gamma^v_{v v} = -\frac{1}{2} g^{v \lambda} \partial_\lambda g_{vv} \quad (B.7) \]

\[ = -\frac{1}{2} \left( g^{v \theta} \partial_v g_{\theta v} + g^{v \theta} \partial_\theta g_{\theta v} \right). \quad (B.8) \]
Because the metric is symmetric and the minor of $g_{r\theta}$ is zero, $g^{r\theta}$ vanishes. So one gets

$$\Gamma^v_{vv} = -\frac{1}{2}g^{vr}\partial_r g_{vv}, \quad (B.9)$$

and analogously for the other necessary Christoffel symbols

$$\Gamma^v_{v\chi} = -\frac{1}{2}g^{vr}\partial_r g_{v\chi}, \quad (B.10)$$
$$\Gamma^v_{\chi\chi} = -\frac{1}{2}g^{vr}\partial_r g_{\chi\chi}. \quad (B.11)$$

So the expression for the surface gravity in (B.4) becomes

$$\kappa = -\frac{1}{2}g^{vr}\left(\partial_r g_{vv} + 2\frac{a}{r^2 + a^2}\partial_r g_{v\chi} + \frac{a^2}{(r^2 + a^2)^2}\partial_r g_{\chi\chi}\right) \quad \text{on } r = r_{\pm}. \quad (B.12)$$

Now the derivatives of the metric coefficients of the Kerr black hole (1.125) are evaluated at the hypersurfaces $r = r_{\pm}$. First, start with

$$\frac{\partial}{\partial r} g_{vv} = \frac{(2r - 2GM)\rho^2 - 2r(\Delta - a^2 \sin^2 \theta)}{r^4}, \quad (B.13)$$

which on the hypersurface $r = r_{\pm}$ becomes

$$\left. \frac{\partial g_{vv}}{\partial r} \right|_{r=r_{\pm}} = \frac{2(r - GM)\rho^2 + ra^2 \sin^2 \theta}{\rho^4}. \quad (B.14)$$

Then, one gets analogously for the other metric coefficients

$$\left. \frac{\partial g_{v\chi}}{\partial r} \right|_{r=r_{\pm}} = 2a \sin^2 \theta \frac{GM \rho^2 - r(r^2 + a^2)}{\rho^4} \quad (B.15)$$
$$\left. \frac{\partial g_{\chi\chi}}{\partial r} \right|_{r=r_{\pm}} = -2 \sin^2 \theta \rho^2 [2r(r^2 + a^2) - (r - GM)a^2 \sin^2 \theta] - r(r^2 + a^2)^2. \quad (B.16)$$

Using the fact that

$$g^{vr}|_{r=r_{\pm}} = \frac{r^2_{\pm} + a^2}{\rho^2} \quad (B.17)$$

and expression (B.12) for the surface gravity becomes

$$\kappa = -\frac{1}{\rho^4(r^2 + a^2)} \left[ ((r - GM)\rho^2 + ra^2 \sin^2 \theta)(r^2 + a^2)^2 \
+ 2a(r^2 + a^2)a \sin^2 \theta (GM \rho^2 + r(r^2 + a^2)) \\n- a^2 \sin^2 \theta (\rho^2(2r^3 + 2ra^2 - (r - GM)a^2 \sin^2 \theta) - r(r^2 + a^2)^2) \right], \quad (B.18)$$

where the index of $r$ has been dropped for simplicity. From now on it is understood that $r$ is taken at $r_{\pm}$. Putting $\alpha = r - GM$, take all the terms of (B.18) which contain $\rho^2$ and simplify.
them to

\[
\rho^2(\alpha r^4 + 2\alpha r^2 a^2 + \alpha a^4 + 2GM(a^2 r^2 + a^4) \sin^2 \theta - a^2 \sin^2 \theta (2r^3 + 2ra^2 - \alpha a^2 \sin^2 \theta))
\]

\[
= \rho^2 \alpha (r^4 + 2r^2 a^2 + a^4 + a^4 \sin^4 \theta - 2 \sin^2 \theta a^2 r^2 - 2 \sin^2 \theta a^4)
\]

\[
= \rho^2 \alpha (r^4 + 2r^2 a^2 \cos^2 \theta + a^4 \cos^4 \theta)
\]

\[
= \rho^6 \alpha
\]

\[
= \rho^6 (r - GM).
\] (B.19)

One can do the same thing for the terms without \( \rho^2 \)

\[
ra^2 \sin^2 \theta (r^2 + a^2)^2 - r(r^2 + a^2)^2 a (r^2 + a^2) a \sin^2 \theta + a^2 \sin^2 \theta r (r^2 + a^2)^2
\]

\[
= (r^2 + a^2)^2 (ra^2 \sin^2 \theta - 2ra^2 \sin^2 \theta + ra^2 \sin^2 \theta)
\]

\[
= 0.
\] (B.20)

So the surface gravity (B.18) finally becomes

\[
\kappa = \frac{\rho^6 (r_\pm - GM)}{\rho^6 (r_\pm^2 + a^2)}
\]

\[
= \frac{r_\pm - r_\mp}{2(r_\pm^2 + a^2)},
\] (B.21)

where the index \( \pm \) has been reintroduced. It follows that the surface gravity is constant, as expected.
Appendix C

The zeroth law

In general relativity, it holds that for a hypersurface $\mathcal{N}$ to be the horizon of a black hole it has to be the Killing horizon of a Killing vector field $\xi$. So it follows from its definition that $\xi$ is orthogonal to the horizon $\mathcal{N}$. Using Frobenius’s theorem (see appendix A), this implies that

$$\xi_{\mu} \nabla_\nu \xi_{\sigma} = 0$$

(C.1)
on $\mathcal{N}$.

As a consequence of (C.1), the contraction of $\xi_{\mu} \nabla_\nu \xi_{\sigma}$ with a third-rank totally anti-symmetric tensor $A^{\rho\sigma\mu}$ vanishes. This can be seen as follows. Take all the terms in the contraction where the indices are permutations of 0, 1 and 2. These are

$$A^{012} \xi_0 \nabla_1 \xi_2 + A^{102} \xi_1 \nabla_0 \xi_2 + A^{120} \xi_1 \nabla_2 \xi_0$$

+C.2

$$+ A^{210} \xi_2 \nabla_1 \xi_0 + A^{201} \xi_2 \nabla_0 \xi_1 + A^{021} \xi_0 \nabla_2 \xi_1.$$ (C.2)

Because of the anti-symmetry of $A$, this can be rewritten as

$$A^{01234}(\xi_0 \nabla_1 \xi_2 \nabla_3 \xi_4),$$ (C.3)

which is equal to zero because of (C.1) because it should hold for all $A$. The same reasoning can be applied to any other combination of index values.

Because it holds on the horizon that $\Psi = A^{\rho\sigma\mu} \xi_\rho \nabla_\sigma \xi_\mu = 0$, we also have

$$0 = \xi_\mu \nabla_\nu (A^{\rho\sigma\mu} \xi_\rho \nabla_\sigma \xi_\mu)$$

$$= A^{\rho\sigma\mu} (\xi_\nu \nabla_\rho \xi_\mu) \nabla_\sigma \xi_\mu + A^{\rho\mu\nu} \xi_\rho (\xi_\nu \nabla_\sigma \xi_\mu)$$

$$+ A^{\rho\sigma\mu} \kappa \xi_\rho \nabla_\sigma \xi_\mu + A^{\rho\sigma\mu} \xi_\rho \nabla_\sigma (\kappa \xi_\mu) - A^{\rho\sigma\mu} (\xi_\rho \nabla_\sigma \xi_\mu)(\nabla_\nu \xi_\mu)$$

$$+ A^{\rho\sigma\mu} \kappa \xi_\rho \nabla_\sigma \xi_\mu + A^{\rho\sigma\mu} \xi_\rho \nabla_\sigma (\kappa \xi_\mu) - A^{\rho\sigma\mu} (\xi_\rho \nabla_\sigma \xi_\mu)(\nabla_\nu \xi_\mu).$$ (C.4)
So we get

\[ A^{\rho\sigma\mu}(\xi_{\rho} \nabla_{\sigma} \xi_{\mu})(\nabla^{\nu} \xi_{\nu}) = A^{\rho\sigma\mu} \kappa \xi_{\rho} \nabla_{\sigma} \xi_{\mu} + A^{\rho\sigma\mu} \kappa \xi_{\rho} \nabla_{\sigma} \xi_{\mu} + A^{\rho\sigma\mu} \xi_{\rho} \nabla_{\sigma} \kappa \]

\[ = 2A^{\rho\sigma\mu} \kappa \xi_{\rho} \nabla_{\sigma} \xi_{\mu}. \]  

(C.5)

Again choosing a particular set of indices, for example 0, 1 and 2, we can write the contribution of all permutations of these indices in the summation in (C.5) as

\[ 12A^{012} \kappa \xi_{[0} \nabla_{1]} \xi_{2} = 12A^{012}(\xi_{[0} \nabla_{1]} \xi_{1})(\nabla^{\nu} \xi_{[2]}). \]  

(C.6)

Because (C.5) should hold for all \( A \), we can write

\[ \kappa \xi_{[\mu} \nabla_{\sigma]} \xi_{\lambda} = (\xi_{[\mu} \nabla_{\sigma]} \xi_{\nu})(\nabla^{\nu} \xi_{[\lambda]}), \]  

(C.7)

which after contraction with \( g^{\lambda\mu} \) becomes

\[ \kappa \xi_{[\mu} \nabla_{\sigma]} \xi_{\mu}^{\lambda} = (\xi_{[\mu} \nabla_{\sigma]} \xi_{\nu})(\nabla^{\nu} \xi_{\lambda}^{\mu}), \]  

(C.8)

an expression we will have to use further on.

From the fact that \( \Psi \) vanishes on \( N \) it follows that its derivative is normal to \( N \). This implies that \( \partial_{\mu} \Psi \) is propotional to \( \xi_{\mu} \), and hence that \( \xi_{[\alpha} \nabla_{\beta]} \Psi = 0 \) on \( N \). Thus

\[ 0 = \xi_{[\alpha} \nabla_{\beta]} (A^{\nu\rho\sigma} \xi_{\nu} \nabla_{\rho} \xi_{\sigma}) \]

\[ = (\xi_{[\alpha} \nabla_{\beta]} A^{\nu\rho\sigma}) \xi_{\nu} \nabla_{\rho} \xi_{\sigma} + A^{\nu\rho\sigma}(\xi_{[\alpha} \nabla_{\beta]} \xi_{\nu}) \nabla_{\rho} \xi_{\sigma} + A^{\nu\rho\sigma}(\xi_{[\alpha} \nabla_{\beta]} \nabla_{\rho} \xi_{\sigma}). \]  

(C.9)

Since this should hold for all totally antisymmetric tensors \( A^{\nu\rho\sigma} \), it follows that all terms in (C.9) should vanish individually.

Before continuing we first derive an important identity. Consider the definition of the Riemann curvature tensor

\[ \nabla_{\mu} \nabla_{\nu} \xi_{\sigma} - \nabla_{\mu} \nabla_{\nu} \xi_{\sigma} = R_{\mu\nu\sigma}^{\rho} \xi_{\rho}. \]  

(C.10)

Using the Killing vector lemma one gets

\[ \nabla_{\mu} \nabla_{\nu} \xi_{\sigma} + \nabla_{\mu} \nabla_{\sigma} \xi_{\nu} = R_{\mu\nu\sigma}^{\rho} \xi_{\rho}. \]  

(C.11)

If one now writes the same equations with cyclic permutated indices and adds the \((\mu\nu\sigma)\) equation to the \((\nu\sigma\mu)\) equation and subtracts the \((\sigma\mu\nu)\) equation, one obtains

\[ 2\nabla_{\nu} \nabla_{\mu} \xi_{\sigma} = (R_{\mu\nu\sigma}^{\rho} + R_{\nu\sigma\mu}^{\rho} - R_{\sigma\mu\nu}^{\rho}) \xi_{\rho}, \]  

(C.12)

\[ = -2R_{\sigma\mu\nu}^{\rho} \xi_{\rho}, \]  

(C.13)

where the Jacobi identity was used in the second step.

Taking the third term of (C.9) and using (C.13) we get

\[ A^{\nu\rho\sigma} \xi_{\nu} R_{\sigma\rho}[\lambda] \xi_{\lambda} = 0. \]  

(C.14)
Now using the anti-symmetry of the Riemann curvature tensor under interchange of the first two indices, the same reasoning which lead to (C.3) now gives
\[
(\xi_{\nu R_{\sigma\rho}}^{\lambda})_{\xi_{\sigma}} + \xi_{\rho R_{\sigma\nu}}^{\lambda} + \xi_{\sigma R_{\rho\nu}}^{\lambda} \xi_{\lambda} = 0. \tag{C.15}
\]
Now we contract (C.20) on \(\rho\) and \(\alpha\). To do so, rewrite the first term
\[
\xi_{\nu} R_{\sigma\rho}^{\lambda} \xi_{\sigma} - \xi_{\nu} R_{\sigma\rho}^{\lambda} \xi_{\lambda}. \tag{C.16}
\]
The contraction on \(\rho\) and \(\alpha\) gives
\[
\frac{1}{2} (\xi_{\nu} R_{\sigma\rho}^{\lambda} \xi_{\sigma} + \xi_{\nu} R_{\rho\sigma}^{\lambda} \xi_{\lambda}), \tag{C.17}
\]
where \(R_{\sigma\lambda} = R_{\alpha\sigma\lambda}\) is the Ricci tensor. The second term can be rewritten as
\[
\xi_{\rho} R_{\nu\sigma}^{\lambda} \xi_{\sigma} = \frac{1}{2} (\xi_{\rho} R_{\nu\sigma}^{\lambda} \xi_{\sigma} - \xi_{\rho} R_{\nu\sigma}^{\lambda} \xi_{\lambda}). \tag{C.18}
\]
Contraction on \(\rho\) and \(\alpha\) in the first term of (C.18) gives zero because \(\xi^{2} = 0\) on \(\mathcal{N}\). The second term also becomes zero after contraction because the Riemann curvature tensor is anti-symmetric in the last two indices. The third term of (C.20) can be written as
\[
\xi_{\sigma} R_{\rho\nu}^{\lambda} \xi_{\sigma} = \frac{1}{2} (\xi_{\sigma} R_{\rho\nu}^{\lambda} \xi_{\sigma} - \xi_{\sigma} R_{\nu\rho}^{\lambda} \xi_{\lambda}). \tag{C.19}
\]
Which after contraction on \(\rho\) and \(\alpha\) becomes
\[
\frac{1}{2} (\xi_{\sigma} R_{\rho\nu}^{\lambda} \xi_{\sigma} - \xi_{\sigma} R_{\nu\rho}^{\lambda} \xi_{\lambda}). \tag{C.20}
\]
Now combining the first terms of (C.17) and (C.20) gives
\[
\xi_{\alpha} \xi_{\nu} R_{\sigma\rho}^{\lambda} \xi_{\sigma} \lambda. \tag{C.21}
\]
And combining the second term of (C.17) and (C.20) results in
\[
\xi_{\beta} \xi_{\nu} R_{\sigma\rho}^{\lambda} \xi_{\lambda}. \tag{C.22}
\]
Finally, using (C.21) and (C.22), one gets for the contraction of (C.20) on \(\rho\) and \(\alpha\)
\[
\xi_{\alpha} \xi_{\nu} R_{\sigma\rho}^{\lambda} \xi_{\sigma} \lambda = -\xi_{\beta} \xi_{\nu} R_{\sigma\rho}^{\lambda} \xi_{\lambda}. \tag{C.23}
\]
Because on \(\mathcal{N}\), it holds that
\[
\xi \cdot \nabla \xi = \kappa \xi^{\mu}, \tag{C.24}
\]
the scalar \(\Phi = (\xi \cdot \nabla \xi - \kappa \xi) \cdot v\), with \(v\) an arbitrary vector, vanishes on \(\mathcal{N}\). So by the same reasoning of above, it follows that \(\xi_{\mu} \partial_{\nu} \Phi_{\mathcal{N}} = 0\). Because \(v\) is arbitrary, this implies
\[
0 \quad = \quad \xi_{\mu} \nabla_{\nu}(\xi_{\alpha} \xi_{\mu} - \kappa \xi_{\alpha})
= \quad \xi_{\alpha} (\xi_{\mu} \nabla_{\nu}) + (\nabla_{\alpha} \xi_{\nu})(\xi_{\mu} \nabla_{\nu} \xi_{\alpha}) - \xi_{\alpha} (\xi_{\mu} \nabla_{\nu} \xi_{\nu}) - \kappa (\xi_{\mu} \nabla_{\nu} \xi_{\alpha}). \quad \tag{C.25}
\]
Appendix C. The zeroth law

The second term in (C.25) can be rewritten as \( \kappa \xi_\mu \nabla_\nu \xi_\sigma \) by using (C.8). So one gets

\[
\xi_\sigma \xi_\mu \nabla_\nu \kappa = \xi^\alpha \xi_\mu \nabla_\nu \nabla_\alpha \xi_\sigma.
\] (C.26)

Again using the important identity (C.13), one can write (C.26) as

\[
\xi_\sigma \xi_\mu \nabla_\nu \kappa = -\xi^\alpha \xi_\mu R_{[\sigma \alpha \nu]}^{\lambda} \xi_\lambda
= \xi^\alpha R_{\sigma \alpha [\mu}^{\lambda} \xi_{\nu]\xi_\lambda}.
\] (C.27)

So to get the desired expression, rename some of the indices to get

\[
\xi_\mu \xi_\nu \partial_\sigma \kappa = -\xi^\nu R_{\mu [\nu}^{\lambda} \xi_{\rho]} \xi_\lambda.
\] (C.28)

Because of the cyclic identity of the Riemann curvature tensor

\[
R_{\mu \nu \rho \sigma} + R_{\mu \rho \sigma \nu} + R_{\mu \sigma \nu \rho} = 0,
\] (C.29)

the right hand side of (C.28) becomes

\[
-\xi^\nu R_{\mu [\nu}^{\lambda} \xi_{\rho]} \xi_\lambda
= \xi^\nu R_{\mu \alpha [\sigma \rho]}^{\lambda} \xi_{\lambda}
= -\xi^\nu (R_{\mu \alpha [\sigma \rho]}^{\lambda} + R_{\mu [\sigma \rho \alpha]}^{\lambda} \xi_{\rho]} \xi_\alpha
\] (C.30)

The second term in the last line is zero because the Riemann curvature tensor is anti-symmetric in its last two indices. So it follows that

\[
-\xi^\nu R_{\mu [\nu}^{\lambda} \xi_{\rho]} \xi_\lambda
= -\xi^\nu R_{[\sigma \nu \rho \alpha]}^{\lambda} \xi_{\lambda}
= -\xi^\nu \xi_\rho [R_{\sigma \nu \rho \alpha}^{\lambda} \xi_{\lambda}.
\] (C.31)

By using (C.23) and (C.31), (C.28) becomes

\[
\xi_\mu \xi_\nu \partial_\sigma \kappa = -\xi_\rho R_{\mu \rho \sigma}^{\lambda} \xi_\lambda.
\] (C.32)

This is the desired relation. Because it is shown in section 1.11.2 that \( \xi_\sigma R_{\rho \sigma}^{\lambda} \xi_\lambda = 0 \), it follows from (C.32) that

\[
\xi_\rho R_{\mu \rho \sigma}^{\lambda} \xi_\lambda|_N = 0,
\] (C.33)

which implies that \( \kappa \) is constant on the horizon \( N = H^+ \).
Appendix D

The Hamiltonian formulation of general relativity

The Hamiltonian formulation of general relativity is the foundation of canonical quantum gravity. In this thesis it is used multiple times when the spacetime is 'sliced' and evolution between slices is considered. For these two reasons, the main principles of this approach are presented here [11]. The Hamiltonian framework of general relativity is sometimes referred to as 'the ADM formalism'.

The conventional Lagrangian formulation of general relativity via the Einstein-Hilbert action is spacetime covariant. A Hamiltonian formulation, however, requires a breakup of spacetime into space and time. Indeed, the first step in producing a Hamiltonian formulation of a field theory consists of choosing a time function \( t \) and a vector field \( t^\mu \) on a spacetime such that the surface \( \Sigma_t \) of constant \( t \) are space-like Cauchy surfaces and such that \( t^\mu \nabla_\mu = 1 \). The vector field \( t^\mu \) may be interpreted as describing the 'flow of time' in the spacetime and can be used to identify each \( \Sigma_t \) with the initial surface \( \Sigma_0 \). In Minkowski spacetime the choice of \( t \) and \( t^\mu \) is usually made via a global inertial coordinate system, but in curved spacetime there may not be any preferred choice.

In performing integrals of functions over the spacetime \( M \) it would be natural for most purposes to use the volume element \( \epsilon_{\mu\nu\rho\sigma} = \sqrt{g} dx^1 \wedge ... \wedge dx^n \) associated with the spacetime metric. Similarly, in performing integrals over \( \Sigma_t \), it would be natural in most cases to use the volume element \( \epsilon^{(3)}_{\mu\nu\rho} = \epsilon_{\sigma\mu\nu\rho} n^\sigma \), where \( n^\sigma \) is the unit normal to \( \Sigma_t \). However, these volume elements will, in general, depend on \( t \) in the sense that \( \mathcal{L}_t \epsilon_{\mu\nu\rho\sigma} \neq 0 \) and \( \mathcal{L}_t^{(3)} \epsilon^{(3)}_{\mu\nu\rho} \neq 0 \). The use of a time dependent volume element on \( \Sigma_t \) is particularly inconvenient if one wishes to identify \( \Sigma_t \) with \( \Sigma_0 \) in order to view dynamical evolution as the change of fields on the fixed manifold \( \Sigma_0 \). Therefore, we shall introduce a fixed volume element \( \epsilon_{\mu\nu\rho\sigma} \) on \( M \) satisfying \( \mathcal{L}_t \epsilon_{\mu\nu\rho\sigma} = 0 \). One way to do this, at least locally, is to introduce coordinates \( x^1, x^2, x^3 \), and to take \( e \) to be the coordinate volume element \( dt \wedge dx^1 \wedge dx^2 \wedge dx^3 \). On each \( \Sigma_t \), one then defines \( \epsilon^{(3)}_{\mu\nu\rho} = \epsilon_{\sigma\mu\nu\rho} t^\sigma \). Unless stated otherwise, all integrals over \( M \) will be performed using the volume element \( \epsilon_{\mu\nu\rho\sigma} \) and all integrals over \( \Sigma_t \) will be with respect to the volume element \( \epsilon^{(3)}_{\mu\nu\rho} \).
The next step in giving a Hamiltonian formulation is to define a configuration space for the field by specifying what tensor field (or fields) \( q \) on \( \Sigma_t \) physically describes the instantaneous configuration of the field \( \psi \). The space of possible momenta of the field at a given configuration \( q \) then is taken to be the ‘cotangent space’ of the configuration space at \( q \). In the case where the allowed infinitesimal variations, i.e. tangent vectors, \( \delta q \) at \( q \) are represented by tensor fields on \( \Sigma_t \) of type \((k,l)\), the space of momenta consists of tensor fields \( \pi \) of type \((l,k)\) on \( \Sigma_t \) so that \( \pi \) maps \( \delta q \) into \( \mathbb{R} \) via \( \delta q \to \int_{\Sigma_t} \pi \delta q \), where contraction of indices is understood. A prescription must then be given for associating a momentum \( \pi \) to the field \( \psi \) on \( \Sigma_t \).

The final step required for a Hamiltonian formulation of a field theory is the specification of a functional \( H[q,\pi] \) on \( \Sigma_t \), called the Hamiltonian, which is of the form

\[
H = \int_{\Sigma_t} \mathcal{H},
\]

where \( \mathcal{H} \) is the Hamiltonian density. It is a local function of \( q, \pi \) and of their spatial derivatives up to a finite order, such that the pair of equations

\[
\dot{q} \equiv \mathcal{L}_t q = \frac{\delta \mathcal{H}}{\delta \pi} \quad \text{(D.2)}
\]
\[
\dot{\pi} \equiv \mathcal{L}_t \pi = -\frac{\delta \mathcal{H}}{\delta q} \quad \text{(D.3)}
\]

is equivalent to the field equation satisfied by \( \psi \). These equations are called the Hamilton equations.

Given a Lagrangian formulation of a field theory, there is a standard prescription for obtaining a Hamiltonian formulation which is closely analogous to the well known procedure of particle mechanics. First, one takes \( q \) to be simply the field \( \psi \) evaluated on \( \Sigma_t \). Then one views the Lagrangian density as a function of \( q \), its time derivatives and its space derivatives. Assuming that \( \mathcal{L} \) does not depend on time derivatives of \( q \) higher than first order, the momentum \( \pi \) associated with \( \psi \) on \( \Sigma_t \) is

\[
\pi = \frac{\partial \mathcal{L}}{\partial \dot{q}}. \quad \text{(D.4)}
\]

If this equation can be solved for \( \dot{q} \), one defines

\[
\mathcal{H}(q,\pi) = \pi \dot{q} - \mathcal{L}, \quad \text{(D.5)}
\]

where \( \dot{q} = \dot{q}(q,\pi) \) is understood in this equation. Now define

\[
J = \int_{t_1}^{t_2} H \, dt = -I + \int_{t_1}^{t_2} dt \int_{\Sigma_t} \pi \dot{q}. \quad \text{(D.6)}
\]
Then, for a smooth one-parameter variation of $\psi$ which satisfies that $\delta \psi = 0$ at $t = t_1$ and $t = t_2$, one has

\[
\frac{dJ}{d\lambda} = \int_{t_1}^{t_2} dt \int_{\Sigma_t} \left[ \frac{\delta \mathcal{H}}{\delta q} \delta q + \frac{\delta \mathcal{H}}{\delta \pi} \delta \pi \right] = \int_{t_1}^{t_2} dt \int_{\Sigma_t} [\pi \delta \dot{q} + \dot{q} \delta \pi] - \frac{dI}{d\lambda} = \int_{t_1}^{t_2} dt \int_{\Sigma_t} [-\dot{\pi} \delta q + \dot{q} \delta \pi] - \frac{dI}{d\lambda}.
\]

Thus, comparing the first and last line of this equation, it follows that $\delta I/\delta \psi = 0$ if and only if the Hamilton equations (D.2) and (D.3) are satisfied. Thus, $\mathcal{H}$ is indeed a Hamiltonian density for $\psi$. With the construction above, one can readily construct a Hamiltonian formulation for fields in a general spacetime.

Now, we would like to obtain a Hamiltonian formulation of Einstein’s equations. First note that one cannot interpret $t$ and $t^\mu$ in terms of physical measurements using clocks which run a certain proper time until one knows the spacetime metric, which is the unknown field variable in Einstein’s equations.

Given a metric $g_{\mu\nu}$, it is convenient to decompose $t^\mu$ into its normal and tangential parts with respect to the surfaces $\Sigma_t$ of constant $t$. Define the lapse function $N$ by

\[
N = -g_{\mu\nu} t^\mu n^\nu
\]

and the shift vector $N^\mu$ by

\[
N^\mu = h^\mu_\nu t^\nu,
\]

where $n^\mu$ is again the unit normal to $\Sigma_t$ and $h_{\mu\nu} = g_{\mu\nu} + n_\mu n_\nu$ is the induced spatial metric on $\Sigma_t$. Thus, $N$ measures the rate of flow of proper time $\tau$ with respect to coordinate time $t$ as one moves normally to $\Sigma_t$, whereas $N^\mu$ measures the amount of ’shift’ tangential to $\Sigma_t$ contained in the time flow vector field $t^\mu$. This is depicted on figure D.1.

The lapse function can be defined in a more convenient form by considering the definition of $n^\mu$

\[
n^\mu = -f g^{\mu\nu} \nabla_\nu t.
\]
Because $n^\mu$ is normalized, it follows that

$$f^2 = -\frac{1}{g^{\mu\nu}\nabla_\mu t \nabla_\nu t} \Rightarrow f = \frac{1}{n^\mu \nabla_\mu t}.$$  

(D.11)

With this, (D.8) can be rewritten as

$$N = g_{\mu\nu} t^\nu (g^{\rho\sigma} \nabla_\rho t)^{-1} (n^\sigma \nabla_\sigma t)^{-1}$$

$$= (t^\rho \nabla_\rho t)^{-1} (n^\sigma \nabla_\sigma t)^{-1}.$$  

(D.12)

In terms of $N$, $N^\mu$ and $t^\mu$, one has

$$n^\mu = \frac{1}{N} (t^\mu - N^\mu),$$  

(D.13)

and hence the inverse spacetime metric can be written as

$$g^{\mu\nu} = h_{\mu\nu} - n^\mu n^\nu = h^{\mu\nu} - N^{-2} (t^\mu - N^\mu)(t^\nu - N^\nu).$$  

(D.14)

It is convenient to choose as our field variables the spatial metric $h_{\mu\nu}$, the lapse function $N$ and the covariant form of the shift vector $N_\mu = h_{\mu\nu} N^\nu$ rather than the inverse metric $g^{\mu\nu}$. The requirements that $h^{\mu\nu} h_{\nu\rho}$ is the identity operator on the tangent space to $\Sigma_t$ and that $h^{\mu\nu} \nabla_\nu t = 0$ allow us to compute $h^{\mu\nu}$ from $h_{\mu\nu}$ and thence obtain $N^\mu = h^{\mu\nu} N_\nu$. Thus, from (D.14) one sees that the information contained in $(h_{\mu\nu}, N, N_\mu)$ is equivalent to that contained in $g^{\mu\nu}$.

To simplify the discussion here, we will restrict the analysis to the situation without boundaries so we can use the conventional Einstein-Hilbert action. In the following, the main results of the Hamiltonian formulation are simply given, for a detailed derivation we refer to [11].

The first step in obtaining a Hamiltonian functional for general relativity is to express the gravitational action in terms of $(h_{\mu\nu}, N, N_\mu)$ and their time and space derivatives. The correct form is

$$\mathcal{L}_G = \sqrt{h} N [R^{(3)} + K_{\mu\nu} K^{\mu\nu} - K^2],$$  

(D.15)

where $K_{\mu\nu} = h_{\rho}^\sigma \nabla_\sigma n_\nu$ is the extrinsic curvature of $\Sigma_t$ and $K = K_\mu^\mu$. The extrinsic curvature can be related to the 'time derivative' $\dot{h}_{\mu\nu} \equiv h_{\rho}^\sigma \partial_t h_{\mu\sigma}$ of $h_{\mu\nu}$ by

$$K_{\mu\nu} = \frac{1}{2N} [\dot{h}_{\mu\nu} - D_\mu N_\nu - D_\nu N_\mu],$$  

(D.16)

where $D_\mu$ is the derivative operator on $\Sigma_t$ associated with $h_{\mu\nu}$.

The momentum canonically conjugate to $h_{\mu\nu}$ then is

$$\pi^{\mu\nu} = \partial \mathcal{L}_G / \partial \dot{h}_{\mu\nu} = \sqrt{h} (K^{\mu\nu} - K h^{\mu\nu}).$$  

(D.17)

Note that the Lagrangian density does not contain any time derivatives of $N$ or $N_\mu$, so their conjugate momenta vanish identically. This is interpreted as telling us that $N$ and $N_\mu$ should
not be viewed as dynamical variables. They will only give constraints and will not give rise to dynamical equations. Hence, the configuration space is defined as to consist of Riemannian metrics $h_{\mu \nu}$ on $\Sigma_t$. The Hamiltonian density is then defined in the standard way

$$\mathcal{H}_G = \pi^{\mu \nu} \dot{h}_{\mu \nu} - L_G,$$

leading to the Hamilton equations by the conventional formulas.

To conclude, we give a few remarks on the constraints that appear due to the independence of $L_G$ on $N$ and $N^\mu$. By variation of $H_G$ with respect to $N$ and $N^\mu$, one can identify these constraints as

$$h R^{(3)} - \pi^{\mu \nu} \pi_{\mu \nu} + \frac{1}{2} \pi^2 = 0 \quad (D.19)$$

$$D_\mu (\sqrt{h} \pi^{\mu \nu}) = 0.$$  \quad (D.20)

The situation is very similar to the one electromagnetism, where the Maxwell equations split up into a gauge condition and the true dynamical equations. In that sense, (D.19) and (D.20) could be viewed as the analogons of $\vec{\nabla} \cdot \vec{E} = 0$. The gauge freedom in general relativity is the covariance under general coordinate transformations. So the constraints are in some way ‘gauge fixing terms’. This is confirmed by the fact that most of them disappear when one uses equivalence classes of metrics $\tilde{h}_{\mu \nu}$, where two metrics who are related by a diffeomorphism are in the same equivalence class, instead of individual metrics. The configuration space of these equivalence classes is known as a superspace. The constraints that not disappear after the use of superspace are due to the gauge arbitrariness of how to slice spacetime into space and time.

It does not appear possible to find a choice of configuration space for general relativity such that only the ‘true dynamical degrees of freedom’ are present in this phase space. The presence of the constraints appears to be an unavoidable feature of the Hamiltonian formulation of general relativity. This provides a serious obstacle to the formulation of a quantum theory of gravity by the canonical quantization approach.
Appendix E

Scrambled entanglement

There are two situations in which large amounts of entanglement are known to occur [164]. The first has to do with the properties of the ordered ground states of quantum field theories including condensed matter systems. The second is almost the complete opposite; it involves entanglement that occurs as a result of complete randomness.

In the first case nearby subsystems tend to be highly entangled as a result of energy considerations. This type of entanglement entropy leads to the area law for entanglement entropy, the reason being elementary; the number of lattice points adjacent to a given region is proportional to the surface area of the region. If one thinks of entanglement as the sharing of Bell pairs, then the Bell pairs in this first type of entanglement are well localized and the components of a pair are not distantly separated.

E.1 Ordered ground states

A typical example of an ordered ground state is the vacuum of a conformal field theory. If one divides space into a left and right half, the two halves will be entangled with an divergent entropy proportional to the area of the dividing plane

\[ S = \frac{A}{\epsilon^2}, \]  

(E.1)

where \( \epsilon \) is a UV cutoff. A heuristic picture of the entanglement can be provided by dividing the space on either side into cells in a scale-invariant way. This is shown on figure E.1. The vertical line in the figure representing the boundary between the entangled regions has been drawn thickened to represent the cutoff length \( \epsilon \).

In each cell a degree of freedom can be defined by averaging the field over the cell. The degree of freedom in a cell at a distance \( l \) from the dividing-surface are therefore field-modes with wavelength of order \( l \). The entanglement across the surface can be approximated by saying that mirror image cells are entangled. The locality and scale-invariant character of the entanglement can be roughly modeled by thinking of the cells as qubits which are entangled in Bell pairs, \( A_i \).
being entangled with $B_i$, as in figure E.1. Each entangled Bell pair contributes a single bit of entanglement entropy.

E.2 Scrambled systems

The second situation is entirely different in character. It occurs when energy is not a consideration at all. It is the entanglement entropy of a scrambled system. The shared Bell pairs in this type of entanglement are extremely delocalized, they are diffused over the entire system.

A good example is based on a random system of a large number $N$ qubits. In the computational basis each qubit has two basis states labeled 0 or 1. One begins with a highly non-typical state such as

$$|\Psi_0\rangle = |0000000...00\rangle .$$

To scramble this system, randomly pick an operator $U$ from some ensemble of $2^N \times 2^N$ unitary matrices. A simple ensemble is the maximally random Haar ensemble. The scrambled state is then defined by

$$|\Psi\rangle = U|\Psi_0\rangle$$

With overwhelming probability $|\Psi\rangle$ has the scrambled property which means that any subsystem has essentially no information. A small subsystem means any subset of qubits fewer than half the total number. If $M < N/2$, then a subsystem of $M$ qubits is small. On the left of figure E.2 an $N$-qubit system is divided into an $M$-qubit subsystem and an $(N - M)$-qubit subsystem.

The statement that the subsystem contains no information with overwhelming probability means that for almost all matrices $U$ its entanglement entropy is very close to maximal

$$S_M = M \log 2.$$ 

(E.4)

Throughout this thesis, the factor $\log 2$ is dropped and the entropy is measured in bits. The equality sign in (E.4) is not exact but the error is less than a single bit, and generally much less than that. This small discrepancy will be ignored.
Another way to say the same thing is that the density matrix of the small subsystem $M$ is extremely close to the maximally mixed density matrix

$$\rho_M = \frac{1}{2^M} I.$$  \hfill (E.5)

Again, the equality sign is correct up to negligible errors in the large $N$ limit. It follows that the scrambled state $|\Psi\rangle$ can be written in the form

$$|\Psi\rangle = \sum_i |i\rangle_s |\phi_i\rangle_b,$$  \hfill (E.6)

where the states $|i\rangle_s$ represent a basis for the small $M$-qubit subsystem, and the $|\phi_i\rangle_b$ represent states in the big subsystem of $(N - M)$ qubits. Moreover, the the fact that the density matrix of the small subsystem is maximally mixed implies that the $|\phi_i\rangle_b$ are orthonormal. However, the $|\phi_i\rangle_b$ cannot be a complete basis for the big system. They only span a subsystem of $M$ qubits that lives in the larger $(N - M)$ qubit subsystem. This subsystem is most certainly not a collection of the original defining qubits that make up the computational basis. However, it is unitarily equivalent to such a subsystem. To make this precise, one can take any $M$-qubit subsystem from the $(N - M)$ system. This is shown on the right side of figure E.2. The point now is that any state of the form (E.6) is close to a state that can be expressed by the following two step process. First, define a state in which the two small subsystems of both $M$ qubits are maximally entangled, and the third $(N - 2M)$ subsystem factors off

$$|\Phi\rangle = \sum_i |i\rangle_s |i\rangle_{s'} |00000\ldots0\rangle,$$  \hfill (E.7)

where $s'$ refers to the second small subsystem and $|00000\ldots0\rangle$ denotes the state of the remaining $(N - 2M)$ qubits. In such a state the small subsystem is manifestly maximally entangled with a subsystem of the big big subsystem. As the second step, one obtains $|\Psi\rangle$ from $|\Phi\rangle$ by applying a unitary scrambling operator $V$ on the big $(N - M)$ qubit subsystem

$$|\Psi\rangle = V|\Phi\rangle.$$  \hfill (E.8)

The operator $V$ is the product of a scrambling operator on the big subsystem and the unit operator in the small subsystem. What $V$ does is to scramble the $M$ qubits that are entangled with the small subsystem and hide them among the larger $(N - M)$ qubits of the big subsystem.
An important point to bear in mind is that the matrix $V$ depends on the state $|\Psi\rangle$. In other words, $V$ is a function of $U$.

The two situations described above seem to have very little in common, but in black hole physics they both fulfill a crucial role. The ordered ground state describes the vacuum as seen by an infalling observer while the scrambled, chaotic state is perceived by the outside observer.
Bibliography


[77] W. Zurek and K. Thorne. Statistical mechanical origin of the entropy of rotating, charged

    2011.


[80] V.P. Frolov and D.N. Page. Proof of the generalized second law for quasistationary semi-


[83] Ming-Sheng Zhan Qing-yu Cai, Boacheng Zhang and Li You. Comment on ”what the


