Lectures on black hole thermodynamics

Marija Tomašević

CPHT, CNRS, Ecole Polytechnique, IP Paris, F-91128 Palaiseau, France

E-mail: marija.tomasevic@polytechnique.edu

Abstract: We will cover some of the basic aspects of black hole thermodynamics and its current status in the field. We begin with Bekenstein's reasoning, then explain why black holes emit quantum radiation, and show that the laws of thermodynamics hold for black holes, just like they would for any thermal system. We outline some aspects and puzzles regarding black hole evaporation. Finally, we comment on the significance of AdS/CFT in solidifying the laws as truly thermodynamic ones.

Contents

1 Introduction

Black holes were found as solutions to Einstein's equations more than a 100 years ago. Yet, they still have a lot to teach us about the fundamental nature of the universe. When people say that they are working on problems related to quantum gravity, they always include black holes as one of their research directions. But how is it that such simple-seeming objects, which are characterized by only a few parameters, form such a basis for studying the quantum theory of gravity?

Black holes are extreme objects. If a star collapses and forms a white dwarf, not a lot of things happen: the white dwarf is a more compact object, whose core is supported only by electron degeneracy pressure, causing it to be extremely dense. Nevertheless, this object will radiate as any thermodynamic object would, and in principle, one would have no problem entering the core of a white dwarf[1]. If one was to live on the surface of such a dwarf, one would have some trouble when wanting to take-off – but not too much. The surface gravity of a solar-mass white dwarf is 350,000 times that of gravity on Earth, and the escape velocity is $\sim 10^3$ km/s. However, if a star is massive enough, it will collapse to a black hole – a completely different object. A black hole does not have a surface on which one can stand; one will keep falling in until one reaches *a singularity* – a label that indicates our theory broke down. If one does not want to end up reaching the singularity, one would need the escape velocity to be bigger than the speed of light! That includes radiation – photons travel at the speed of light, and if they are trapped behind an *event horizon*, they will never be able to escape; hence the term black hole.

So we see how peculiar black holes are – they seem to be sinks of the universe, not letting anything escape the tight grip of the horizon. That sounds like nothing we know in nature! But even though the theory of General Relativity tells us this behaviour is perfectly

[1] One will be burned though; the average temperature of the inner layers is approximately 10^7 K.

fine (after all, black holes can be found as many solutions to the main equations of gravity), does this behaviour make sense when combined with other laws of physics? If not, then there must be something wrong with either GR or other laws!

One of the laws that *must* hold in nature, simply based on statistics, are the laws of thermodynamics – in particular the second law. There is even a famous quote by Eddington that says:

> "The law that entropy always increases holds, I think, the supreme position among the laws of Nature. If someone points out to you that your pet theory of the universe is in disagreement with Maxwell's equations — then so much the worse for Maxwell's equations. If it is found to be contradicted by observation — well, these experimentalists do bungle things sometimes. But if your theory is found to be against the second law of thermodynamics I can give you no hope; there is nothing for it but to collapse in deepest humiliation."

Why did Eddington proclaim such a bold statement?

Thermodynamics deals with concepts of heat, energy, work, temperature, and, most centrally, entropy. It has the benefit of being summarized in 3-4 fundamental laws. It has applications to phases of matter like gases and liquids, chemical reactions, cell biology, all the way up to the universe at its largest scales. Statistical mechanics has a goal of explaining/deriving thermodynamics from the microscopic laws of physics, with its main mathematical tool the probability theory applied to many degrees of freedom. For instance, a box containing gas has $\sim$ Avogadro's number worth of molecules, and comparable numbers of degrees of freedom. Instead of considering each possible molecule, we can compose an ensemble of macroscopic states and take averages to get key statistical quantities, like energy, volume, etc. This then allows us to predict macroscopic behavior without the precise knowledge of the microscopic state. Moreover, large fluctuations from the average behavior are extraordinarily unlikely – the probability of all air molecules in a box accumulating randomly in one corner of the box is extremely unlikely.

Probability theory arises in statistical mechanics because we can't obtain complete knowledge of the microscopic state of the system. Even if we knew the microscopic state, we would still average over the microscopic details of the state when computing expected observations. The main mathematical tools are probability theory and combinatorics – counting, in other words.

Since the basis of statistical mechanics lies in probability theory, it is evident that any system with large numbers will correspondingly follow the laws of thermodynamics. So, seeing if our black holes obey the second law of thermodynamics looks like a prominent way to test the theory of General Relativity – or at least, its black hole solutions.

2 Testing the second law of thermodynamics

How can we test the second law? We will loosely follow the logic that Jacob Bekenstein outlined in his famous paper in 1972 on exactly this topic.

First, we need to find an isolated system. So, for our purposes, let's take a cup of some good hot Turkish coffee[2] as one part of our system, and let's take a black hole as the other. Together, the coffee cup and the black hole form the whole universe – also known as an isolated system. The claim is now that the entropy must increase in such a system. Let us put the coffee cup on a trajectory towards the black hole – these are our initial conditions. And now, let's see what happens: our cup falls across the horizon, never to be seen again. But a coffee cup carried some entropy with it, say S_c. Once the cup crossed the horizon, the total entropy of the observable universe did not increase – in fact, it decreased! In other words, we lost that entropy to the depths of the black hole.

This was a very simple thought experiment, and yet we immediately came to a radical conclusion: *the laws of thermodynamics do not seem to hold when a black hole is involved.* But hold on a second: are we sure that's the correct conclusion? Why can't we simply say that the entropy is *inside* the black hole: too bad for you that you can't reach it anymore, but it surely exist within the black hole itself. In other words, the cup still exists inside the black hole, presumably with its entropy intact, and so, presumably, the second law of thermodynamics still holds.

Compelling as it may sound, such reasoning verges on taking us out of the realm of science! For stating that somehow the entropy still exists inside the black hole, but we just can't see it, simply puts the laws of thermodynamics as unfalsifiable. And the laws of physics should continue to make sense in our laboratories, whether or not a black hole forms in one of the test tubes. In other words, we could never find a counterexample because we can always declare there was a black hole somewhere in the process. And black holes smaller than a coarse-grained scale will not be detected.

Led by this conviction, Bekenstein proposed a resolution: *we have to assign an entropy to black holes.* And there even exists a perfect candidate for it – the area of a black hole. To be more precise, the entropy of a black hole will be given by

$$S = \frac{A_H}{4G\hbar}, \tag{2.1}$$

where G is the Newton's constant and $\hbar$ is the Planck constant. The horizon area A_H is measured in Planck units, $\ell_p = \sqrt{G\hbar c^{-3}} = 10^{-33}$ cm, where c represents the speed of light (and we will usually set $c = 1$). Let us try to motivate this peculiar choice for an entropy.

The key reason behind the choice of an area as the entropy lies in a theorem proven by Stephen Hawking– he proved a statement known as the *area theorem.* The area theorem says that, under some reasonable conditions (like the fact that energy has to be positive along light ray trajectories and that we do not end up forming an uncontrollable singularity), the area of a black hole can never decrease. The intuitive picture is clear: black holes do not let even photons out, so they cannot shrink in size, and if we throw something inside a black hole, that something can only increase the size – we will see this explicitly shortly. The proof of the theorem is not difficult, although it requires some knowledge of differential geometry within the scope of General Relativity.

[2]Note: the nature of the coffee does not matter for the thought-experiment, nor for a real one.

Area never decreasing is strikingly similar to the statement that the entropy never decreases; indeed, this was the reason why Bekenstein thought the horizon area can serve its purpose and save the second law of thermodynamics. Note also that the area is measured in Planck units: the Planck length is comprised out of the three most fundamental constants of Nature: gravity (G), quantum physics ($\hbar$) and relativity (c)[3]. However, assigning an entropy to a black hole is not a conceptually trivial task – the area theorem is an exact theorem in differential geometry, whereas the second law of thermodynamics is only a statistical law. Therefore, Bekenstein asserted that the area of a geometric surface is more fundamentally a statistical quantity. Nearly half a century later, this remains the single deepest insight we have gained into the fundamental nature of space and time. As we will see, it has proven extraordinarily fruitful and lies at the center of today's most promising avenues for understanding the quantum theory of gravity.

Exercise: Read Bekenstein's original paper. It can be found here.

Note that we immediately come to an interesting conclusion for all gravitating systems. Given that the second law seems to be true, one is led to the conclusion that $\frac{A}{4G\hbar}$ must be the *most entropy* that can be contained in a region surrounded by a surface of area A. To maximize the volume one would take a sphere, and if there were more entropy than $\frac{A}{4G\hbar}$, but no black hole, one could simply add more mass until a black hole formed, at which point the entropy would go down to $\frac{A}{4G\hbar}$, violating the second law. Thus the entropy must have been less than $\frac{A}{4G\hbar}$ to begin with. Putting this statement into an equation, we have

$$S \leq \frac{A}{4G\hbar}, \tag{2.2}$$

which is known as the *entropy bound*. Roughly it says that the maximum amount of entropy in a spacetime region scales with the area of the boundary of the region. And if you try to excite more degrees of freedom, you make a black hole instead.

3 Thermodynamics of black holes

A notable feature of systems with some entropy and energy is that they also have a temperature – a feature seen through the first law of thermodynamics. If we couple our system to a cooler bath, our system will transfer its heat until a thermal equilibrium is achieved; in other words, our system must radiate. This suggests that black holes with a higher temperature than their surroundings will also radiate – this seems to be contradicting the very name of a black hole. How can a system radiate when lightrays – the fastest excitations of the field – cannot escape the black hole?

This line of reasoning is why Hawking thought that Bekenstein's calculation must be wrong. In order to prove that, he set out to examine the behaviour of fields near a black hole and prove that it cannot radiate. Funnily enough, he *did* find that black holes radiate once one includes quantum effects. This radiation is now known as Hawking radiation.

[3]Why is there an $\hbar$ in this equation? We will come back to this point; in the meantime, try to think about it.

Hawking radiation through particle production in an external field

Let us try and argue how black holes might radiate by studying particle production in an external field first. In other words, we want to study the production of particles in a given background quantum field which arises due to the fluctuations of the field. Energy required for this materialization of particles can be provided if there is an external field to which the field couples. Such quantum fluctuations are described by a 2-point function,

$$\langle \phi(x)\phi(0)\rangle \sim e^{-\frac{x}{\lambda_c}}, \tag{3.1}$$

where $\lambda_c = \frac{\hbar}{mc}$ is the (reduced) Compton wavelength, comprised of the mass of the field excitation m, speed of light c and the Planck constant $\hbar$. The Compton wavelength plays a role of a quantum-mechanical cutoff below which quantum fluctuations become important; in other words, a significant amount of entanglement is generated. The 2-point function tells us about the probability to spontaneously nucleate an entangled pair of quanta of the field which are separated by a distance x – it is a probability amplitude. Even though we did not derive it here, we see that it makes sense: if the separation between the particles is great, the probability to nucleate such particles goes down, and vice versa.

Now let us rewrite the nucleation distance by the energy required for this pair production process. In order to estimate such energy, we can imagine two capacitors separated by the same distance x that play the role of the particles produced[4]. The energy that is stored between two capacitors plays the role of the necessary energy required for this popping-up process – it is the potential energy. And we know how to calculate the stored energy ϵ between the capacitors: the field is electric $\vec{E}$ and the coupling is the electromagnetic coupling e, so we have

$$eEx = \epsilon, \tag{3.2}$$

To use this formula more generally, we need to understand what it is telling us. The force field F, which in this case is the electric field, along the region x produces energy ϵ, and the strength of that energy production is indicated by the strength of the field, a.k.a. the coupling g_F (in the electric case e). That is, generically, we have

$$g_F F x = \epsilon. \tag{3.3}$$

This energy ϵ now represents our nucleated pair, and so $\epsilon = 2mc^2$. We can plug instead of x in (3.1) this formula, and calculate the probability – given by the square of the norm of the probability amplitude,

$$\Gamma \sim |\langle \phi(x)\phi(0)\rangle|^2 \sim \exp\left\{-\frac{4m^2c^3}{\hbar g_F F}\right\}. \tag{3.4}$$

This Γ represents the probability for pair creation (per unit volume, per unit time). We obtained this probability through a rough argument, but it does give the correct qualitative behaviour, and more importantly, this argument emphasizes the important physics behind

[4]To refresh your memory on capacitor plates and the energy calculation, a quick recap can be found here.

the process. In the case of the electric field, this calculation was first done, in much more detail, by Schwinger in the 1950s, and is known as the *Schwinger pair production* [1].

A proper calculation in quantum field theory would fix the $\sim$ into a $=$ sign. Such a calculation involves a tunneling process (hence the exponential suppression) and can be evaluated through a WKB approximation to give something like

$$\Gamma = A \exp\left\{-\gamma \frac{4m^2c^3}{\hbar g_F F}\right\}, \tag{3.5}$$

where A is calculated through the 1-loop determinant – basically the first-order perturbative quantum correction – and γ is a constant factor of $\mathcal{O}(1)$ which depends on the field content and the coupling. Let us now apply (3.4) to a gravitational field. Gravitational field is always created by some massive body, say the Earth or a black hole; let us focus on black holes here[5]. The gravitational field strength is given by the surface gravity κ, and the coupling is given by the mass m of our particles that we want to nucleate. This is because gravity couples to all matter through their energy. The probability for a pair production is then given by

$$\Gamma \sim \exp\left\{-\gamma \frac{4m^2c^3}{\hbar m\kappa}\right\} = \exp\left\{-\gamma \frac{4mc^3}{\hbar\kappa}\right\}. \tag{3.6}$$

We see that there is a linear dependence on the energy mc^2 in the exponent, which is indicative of a Boltzmann distribution

$$\Gamma \sim e^{-\frac{E}{T_H}}, \tag{3.7}$$

where we have rewritten

$$T_H = \frac{\hbar\kappa}{4\gamma c}, \tag{3.8}$$

with T_H standing for Hawking temperature. This probability (3.7) gives us a *thermal* spectrum, with the temperature $T_H \propto \kappa$. In other words, our black hole is expected to radiate as a black-body! Notice that this thermality does not appear for the Schwinger process, for instance. But it does appear in the case of a gravitational field since gravity couples universally to the energy/mass. It is very suggestive that the universal character of gravity appears to be related to a universal thermal behavior.

Before we move on, we should note the caveats associated with this argument:

[5]Why don't we see pair production due to the gravitational field of Earth? In that case, the field strength is pretty small compared to the black hole of the same size. So, if we don't want our probability to be completely suppressed, one would need to pair-produce very low energy particles. However, such light particles have a much longer wavelength – in fact, much bigger than Earth's radius! One can see this through the acceleration parameter, $a = \frac{GM}{R^2}$, where M is the mass of the Earth, R its radius and G is the Newton's constant. The characteristic length associated with this acceleration is given by $\lambda_a = a^{-1} = \frac{R^2}{GM}$ and we see that, unless $GM \sim R$, we will have $\lambda_a >> R$. This is to say that the characteristic length is much bigger than the radius of the planet. The importance of this λ_a comes in the probability instead of x in (3.1), so we see that for non-negligible probability, $\lambda_c \gtrsim \lambda_a >> R$. Our approximations break down at this point since one cannot treat the nucleated pair as a pair of particles anymore.

- We only derived a qualitative result. Hawking performed a proper calculation in which he fixed the constants that we could not. For instance, $\gamma = \pi/2$, and so the proper Hawking temperature is given by

$$T_H = \frac{\hbar\kappa}{2\pi}, \tag{3.9}$$

  where we set $c = 1$;

- This argument is valid for massive particles. However, Hawking's calculation works for massless particles as well.

Note another feature of this calculation: since the black hole is responsible for supplying the energy required for the pair production, the black hole will then lose some of its energy once the pair is created. From the point of view of produced particles, one has to have a positive energy with respect to the asymptotic observer, and the other will have negative energy, since $E_1 + E_2 = 0$, where $E_1 > 0$. For consistency reasons, the negative energy particle must fall into the black hole. The change in black hole's mass is exactly then $\delta M = E_2 = -E_1 < 0$.

The near-horizon geometry

Having seen that the surface gravity plays a crucial role in defining the temperature of our black hole, we will make a short detour to derive what κ is in terms of other parameters of the black hole. We will do so by going very close to the horizon of a black hole – also known as the Rindler limit. We will do this example for a Schwarzschild black hole,

$$ds^2 = -(1 - \frac{r_h}{r})dt^2 + \frac{dr^2}{1 - \frac{r_h}{r}} + r^2 d\Omega^2, \tag{3.10}$$

where $r_h = 2MG$. To focus on the near-horizon region, we take the horizon radius and expand around it,

$$r = r_h + \xi, \qquad \xi << r_h. \tag{3.11}$$

This new parameter ξ indicates the distance to the horizon, and as we see, it is very small. Now we can expand our Schwarzschild metric for this new coordinate,

$$\frac{r_h}{r} = \frac{r_h}{r_h + \xi} = 1 - \frac{\xi}{r_h} + \mathcal{O}\left(\frac{\xi^2}{r_h^2}\right), \qquad dr = d\xi. \tag{3.12}$$

The metric now takes the form

$$ds^2 = -\frac{\xi}{r_h}dt^2 + r_h\frac{d\xi^2}{\xi} + r_h^2 d\Omega^2. \tag{3.13}$$

We can change the coordinates now conveniently to

$$r_h\frac{d\xi^2}{\xi} = d\rho^2 \;\longrightarrow\; \xi = \frac{\rho^2}{4r_h}, \tag{3.14}$$

so that

$$ds^2 = -\frac{\rho^2}{4r_h^2}dt^2 + d\rho^2 + r_h^2 d\Omega^2. \tag{3.15}$$

We see that this geometry factorizes: the (t, r) part is not affected by the sphere part (θ, ϕ) since no metric components depend on these parameters. The (t, r) part of the metric is

$$ds^2 = -\frac{\rho^2}{4r_h^2}dt^2 + d\rho^2, \tag{3.16}$$

and it is just a slightly unusual way of writing flat spacetime. One can see that by computing the Ricci scalar of this geometry (it should be zero), or by finding a suitable set of coordinates in which the metric is manifestly flat. If we change

$$X = \rho \cosh\left(\frac{t}{2r_h}\right), \qquad T = \rho \sinh\left(\frac{t}{2r_h}\right), \tag{3.17}$$

we obtain

$$ds^2 = -dT^2 + dX^2, \tag{3.18}$$

that is, we obtained a flat spacetime. Likewise, we recognize the metric form (3.16) as the Rindler metric, with κ

$$\kappa^2 = \frac{1}{4r_h^2} \quad \longrightarrow \quad ds_r^2 = -\rho^2\kappa^2 dt^2 + d\rho^2 \tag{3.19}$$

the surface gravity parameter. In static, asymptotically flat spacetimes, surface gravity is simply the acceleration of a static observer near a black hole as measured by the asymptotic observer. Another way to view is through the tension of a string near the horizon: if one stands far away from the black hole holding a string, and one dangles an object on the string (say, a ball) near so it hovers near the horizon, then one can measure the tension of the string to be κM_{object}. From this simple argument, one cannot see that, but in Chapter 6 ("Killing horizons") of Carroll's book [2], this is nicely laid out. For us, it is important that we obtained the surface gravity in terms of black hole parameters,

$$\kappa = \frac{1}{4GM}. \tag{3.20}$$

The black hole laws

Having confirmed that black holes radiate with some temperature T_H and that they have some entropy $S \propto A$, we can see that these quantities obey the laws of thermodynamics. Let us take the example of the Schwarzschild black hole,

$$ds^2 = -\left(1 - \frac{2MG}{r}\right)dt^2 + \frac{dr^2}{1 - \frac{2MG}{r}} + r^2 d\Omega^2, \tag{3.21}$$

where M stands for the black hole mass, and $d\Omega^2 = d\theta^2 + \sin^2\theta d\phi^2$. The area of this black hole is given by the size of the sphere of radius $r_h = 2MG$. When Bekenstein first proposed the horizon area as the black hole entropy, he could only argue it qualitatively (with correct

units) – he could not get the factor of 4 in the entropy formula for instance. This factor was obtained by Hawking when he plugged in his temperature in the first law; this is what we will do now.

The entropy can be represented as

$$S = \alpha \frac{A}{G\hbar} = \alpha \frac{4\pi r_h^2}{G\hbar} = \alpha 16\pi M^2 G\hbar^{-1}, \tag{3.22}$$

where α is the factor we would like to determine. The energy of the black hole is given by its mass M, so in the first law

$$dM = T_H dS \tag{3.23}$$

we will simply replace the quantities that we obtained through Bekenstein and Hawking,

$$dM = \frac{\hbar\kappa}{2\pi} d\left(\alpha 16\pi M^2 G\hbar^{-1}\right) = 16\alpha\kappa GM dM, \tag{3.24}$$

so

$$16\alpha\kappa GM = 1. \tag{3.25}$$

We can obtain κ through the Rindler limit of black holes, which gives us $\kappa = \frac{1}{4GM}$ from which we can see that

$$\alpha = 1/4, \tag{3.26}$$

just like we needed. Even though we obtained this result for the Schwarzschild black hole, one can show that this relation between the area and the entropy is generically true for all black holes. However, the surface gravity relation is *not* – it depends on the asymptotics of the black hole spacetime, and therefore, the temperature dependence on the mass will change as well; for instance, Anti-de Sitter black holes have temperature that is linearly dependent on the mass.

Exercise: The Schwarzschild black hole can exist in spacetimes with other cosmological constants. For instance, the Schwarzschild-AdS black hole is described by

$$ds^2 = -f(r)dt^2 + \frac{dr^2}{f(r)} + r^2 d\Omega^2,$$

with

$$f(r) = 1 - \frac{2MG}{r} + k^2 r^2,$$

where k is the curvature of the AdS. Calculate the entropy and the temperature of this black hole. Then, calculate the specific heat, $c_v = \frac{dE}{dT}$, for this black hole and also for an asymptotically flat (AF) black hole. The specific heat tells us about the thermodynamic stability of systems; what can you conclude about the stability of Sch-AdS and Sch-AF black holes?

We showed here that Schwarzschild black holes obey the first law, but one can actually *prove* this law for all black holes, including when black holes add additional work terms to the first law. As a matter of fact, Bardeen, Carter and Hawking (BCH) proved all

three[6] laws of 'black hole mechanics' before they were understood to be the actual laws of thermodynamics[7]. We will not try to prove them here, but simply list them.

The 0th law:

- Thermodynamics: The temperature T of body at thermal equilibrium is constant throughout the body. Otherwise heat will flow from hot spots to the cold spots.
- Black holes: Stationary black holes have constant surface gravity κ on the event horizon. Can be proven regardless of spherical symmetry.

The 1st law:

- Thermodynamics: Energy is conserved, $dE = TdS + \mu dQ + \Omega dJ$, where E is the energy, Q is the charge with chemical potential μ and J is the spin with chemical potential Ω.
- Black holes: Energy is conserved, $dM = \frac{\kappa}{8\pi G} dA + \mu dQ + \Omega dJ$. For a Schwarzschild black hole we have $\mu = Q = 0$ because there is no charge or spin.

The 2nd law:

- Thermodynamics: In a physical process the total entropy S never decreases, $\Delta S \geq 0$.
- Black holes: The area theorem tells us that the net area in any process never decreases, $\Delta A \geq 0$. For example, two Schwarzschild black holes with masses M_1 and M_2 can coalesce to form a lighter black hole of mass $M < M_1 + M_2$, due to the gravitational waves that carry out the rest of the mass. However, this lighter black hole is still *bigger* in area since $A \propto M^2$ and so, $(M_1 + M_2)^2 > M_1^2 + M_2^2$.

The 3rd law:

- Thermodynamics: It is impossible by any procedure, no matter how idealized, to reduce the temperature to zero by a finite sequence of operations.
- Black holes: One cannot reduce the surface gravity κ to zero by a finite sequence of operations. This law has not been proven, but it is believed to be true. One can imagine bringing the surface gravity all the way to zero, and then continuing beyond – this would create an uncontrollable, *naked* singularity which is believed to be non-existing; this statement is known as the *cosmic censorship conjecture*.

[6] They also note that the fourth law can be written in an analogous form, but they could not prove it.

[7] Note also that in the original paper by BCH (which was written before Hawking's temperature derivation), the horizon area and the black hole surface gravity were seen only as analogous to the entropy and the temperature – their paper therefore sometimes mentions "...it is clear that the black hole cannot radiate..." among other statements, which we now know is not true.

Exercise: Calculate the entropy and the temperature for a Reissner-Nordstrom black hole,

$$ds^2 = -f(r)dt^2 + \frac{dr^2}{f(r)} + r^2 d\Omega^2,$$

with

$$f(r) = 1 - \frac{2MG}{r} + \frac{Q^2}{r^2}, \qquad A_\mu dx^\mu = -\frac{Q}{r}dt, \qquad F_{tr} = \frac{Q}{r^2},$$

where A_μ is the vector potential for this charged black hole of charge Q and mass M. The component of the field strength F_{tr} is the electric field in the radial direction, so this is exactly the gauge field corresponding to a point source of charge Q at $r = 0$. Write down the first law and see what is the work term.

4 Euclidean methods in gravity

Reading off the thermodynamic quantities from the metric only gives us tree-level answers, and one cannot get quantum corrections via this method. For higher-loop contributions – quantum corrections – one needs to employ a different approach. This approach is known as the Euclidean gravitational path integral, but before we delve into it, we will recall how it works in the case of quantum mechanics.

Quantum mechanics at finite temperature

Let us start with a one-dimensional relativistic particle in some external potential $V(q)$, where q is the position of the particle, with the Hamiltonian $H = \frac{p^2}{2m} + V(q)$. The evolution operator is given by

$$U(t) = \exp\left\{\frac{t}{i\hbar}H\right\}, \qquad H \neq H(t), \tag{4.1}$$

so states evolve as

$$|\psi(t)\rangle = U(t)\,|\psi(t=0)\rangle\,. \tag{4.2}$$

We are used to thinking about canonical quantization when it comes to quantum mechanics: we start with some classical variables, and we promote them to operators acting on a Hilbert space. However, there is a different way of obtaining quantized particles and fields. This method is known as the *path integral* approach, and it was developed by Richard Feynman, who used his intuition to define a probability amplitude – the matrix element of the evolution operator – in a very pictorial way[8]. It replaces the classical notion of a single, unique classical trajectory for a system with a sum, or functional integral, over an infinity of quantum-mechanically possible trajectories to compute the probability amplitude. In quantum mechanics, we cannot speak of the particle taking any well-defined trajectory between two points. Instead, we can only speak of the probability of finding the particle at these locations. That is, all that can be determined is the relative probability of the particle taking one path or another. Feynman's insight was this - the total transition probability

[8] A very detailed derivation can be found here; just ignore the language, and follow the equations.

amplitude can be obtained by summing the amplitudes of the particle having taken any individual path. So, the three things to keep in mind:

- The probability for an event is given by the squared modulus of a complex number called the "probability amplitude";
- The probability amplitude is given by adding together the contributions of all paths in configuration space;
- The contribution of a path is proportional to $\exp\{\frac{i}{\hbar}S\}$, where S is the action given by the time integral of the Lagrangian along the path.

In order to find the overall probability amplitude for a given process one integrates the amplitude over the space of all possible paths of the system in between the initial and final states, including those that are absurd by classical standards. In calculating the probability amplitude for a single particle to go from one space-time point to another, it is correct to include paths in which the particle describes elaborate curves in which the particle shoots off into outer space and flies back again, and so forth. The transition amplitude is then obtained as

$$K(q_2,t_2;q_1,t_1) = \langle q_2|\, U(t_2-t_1)\,|q_1\rangle = \int_{q(t_1)=q_1}^{q(t_2)=q_2} \mathcal{D}[q(s)]\exp\left\{\frac{i}{\hbar}S[q]\right\}, \tag{4.3}$$

where S is the classical action

$$S[q] = \int_{t_1}^{t_2} dt\left(\frac{m}{2}\dot{q}^2 - V(q)\right). \tag{4.4}$$

Now what happens if we take the one-dimensional particles and we put it in a thermal equilibrium with some bath? In this case, we cannot say for sure what the state of the particle is – it is completely mixed with the thermal bath. However, we can describe its properties in a statistical sense. In other words, the particle is in a thermal state and described by a density matrix,

$$\rho = \frac{1}{Z}\exp\{-\beta H\}, \qquad \beta = \frac{1}{k_B T}, \qquad k_B - \text{Boltzmann constant.} \tag{4.5}$$

This state is known as the Gibbs state, and Z is the partition function,

$$Z = tr\left(\exp\{-\beta H\}\right) \tag{4.6}$$

chosen in such a way so as to have $tr\rho = 1$. We see that there is a factor of $\exp\{-\beta H\}$, which looks very much like the evolution operator, with a funny, imaginary time[9]. We can replace

$$\exp\{-\beta H\} = U(-i\tau), \tag{4.8}$$

[9] In order to keep the spectrum of the Hamiltonian bounded, one needs to have $\mathrm{Im}\, t < 0$. One can show this through the energy representation of the evolution operator,

$$U(t) = \sum_n \exp\left\{\frac{t}{i\hbar}E_n\right\}|n\rangle\langle n|, \tag{4.7}$$

where $H|n\rangle = E_n|n\rangle$. We see that the sum converges only when $\mathrm{Im}\, t < 0$.

where

$$\frac{\tau}{\hbar} = \beta = \frac{1}{k_B T}. \tag{4.9}$$

In some sense, one can view a quantum system in thermal equilibrium as a quantum system in imaginary time. And of course, we can now define the Euclidean path integral since we have Euclidean time evolution,

$$U(-i\tau) \equiv U_E(\tau). \tag{4.10}$$

The transition amplitude is simply given by

$$K(q_2, \tau_2; q_1, \tau_1) = \langle q_2 | \, U_E(\tau_2 - \tau_1) \, | q_1 \rangle = \int_{q_1 = q(\tau_1)}^{q_2 = q(\tau_2)} \mathcal{D}[q[\sigma]] \exp\left\{ -\frac{1}{\hbar} S_E[q] \right\}, \tag{4.11}$$

where S_E is the real and positive Euclidean action,

$$S_E[q] = \int d\sigma \left(\frac{m}{2} \dot{q}^2 + V(q) \right), \tag{4.12}$$

and it is equal to the standard action with an imaginary factor, $iS[q] = -S_E[q]$. We see that there are no more oscillations, but now we have a minimal trajectory, with others suppressed. In other words, we associate a Boltzmann factor to each trajectory $q[\sigma]$ and compute the average over a statistical system, where the statistics is over all the trajectories.

To see that Z is now really a partition function, let us write it out in some basis,

$$Z = tr \left(U_E(\beta\hbar) \right) = \sum_n \langle n | \, U_E(\beta\hbar) \, | n \rangle = \sum_n \exp\{-\beta E_n\}. \tag{4.13}$$

Using the decomposition of unity,

$$\mathbb{1} = \sum_n |n\rangle \langle n| = \int dq \;\; |q\rangle \langle q| \,, \tag{4.14}$$

we can rewrite (4.13) as

$$Z = \int dq \;\; \langle q | \, U_E(\beta\hbar) \, | q \rangle \,, \tag{4.15}$$

where now this integral is over all position eigenstates q, and equal to

$$Z = \int \mathcal{D}q[\sigma] \exp\left\{ -\frac{1}{\hbar} S_E[q] \right\}. \tag{4.16}$$

This is the path integral representation of the partition function at finite temperature. Notice that the trace imposes that $q_1 = q_2 = q$; in other words, it imposes a periodicity on the initial and final trajectories, $q(\beta\hbar) = q(0)$, with periodicity

$$\tau_\beta = \beta\hbar = \frac{\hbar}{k_B T}. \tag{4.17}$$

Finally, let us just note why we called this the Euclidean method. Recall the Minkowski metric,

$$ds^2 = -dt^2 + \sum_i dx_i^2. \tag{4.18}$$

If we now replace $t \longrightarrow -i\tau_\beta$, we get

$$ds_E^2 = d\tau_\beta^2 + \sum_i dx_i^2, \tag{4.19}$$

which now has the Euclidean signature $(++++)$. Notice that we can do the same trick in Rindler space,

$$ds^2 = -\rho^2\kappa^2 dt^2 + d\rho^2 \tag{4.20}$$

to obtain

$$ds_E^2 = \rho^2\kappa^2 d\tau_\beta^2 + d\rho^2 = \rho^2 d(\kappa\tau = \theta)^2 + d\rho^2 \tag{4.21}$$

where now the time is periodic with a specific period,

$$\theta \sim \theta + 2\pi. \tag{4.22}$$

One has to have a period of 2π since otherwise we would not have a regular metric at $\rho = 0$; we would get a conical singularity if the period is less than 2π and conical excess if it is bigger than 2π. This imposes a constraint on κ

$$\kappa\tau_\beta = 2\pi \quad \longrightarrow \quad \kappa = \frac{2\pi}{\tau_\beta} = \frac{2\pi}{\beta\hbar}. \tag{4.23}$$

This is exactly the constraint we obtained by thinking about surface gravity! This simple example indicates that something like this Euclidean method should make sense in gravity as well.

Euclidean gravitational path integral

In ordinary quantum field theory, in order to do a path integral we first fix the spacetime manifold $\mathcal{M}$, and then integrate over fields defined on $\mathcal{M}$. In quantum gravity however, we have to integrate over the geometry as well,

$$Z = \int \mathcal{D}g\mathcal{D}\phi \exp\{-S_E[g,\phi]\}, \tag{4.24}$$

where now

$$S_E[g,\phi] = -\frac{1}{16\pi}\int \sqrt{g}(R + \dots) + \text{boundary terms} + S_{\text{matter}}[\phi] \tag{4.25}$$

We have put (for now) $\hbar = G = 1$. But what are these boundary terms? Note that the gravitational action has an integrand that schematically is like $R \sim g^{-1}\partial^2 g$, which is a little bit unusual: if we think about a classical toy model for this, it would have the action as

$$S \sim -\frac{1}{2}\int dt\ q\ddot{q}, \tag{4.26}$$

which differs from the usual action form by a boundary term,

$$\bar{S} = \frac{1}{2}\int dt\ \dot{q}^2 = -\frac{1}{2}\int d\left(q\dot{q}\right) - \frac{1}{2}\int dt\ q\ddot{q}. \tag{4.27}$$

So, we see that if we want to obtain a Hamiltonian contribution, we should add a boundary term to our gravitational action, which we would expect to involve first derivatives of the metric. A term like that is called the *extrinsic scalar* and it is obtained from the *extrinsic curvature* $K_{\mu\nu}$ through

$$K = h^{\mu\nu}K_{\mu\nu}, \qquad K_{\mu\nu} = \frac{1}{2}\mathcal{L}_n h_{\mu\nu}, \tag{4.28}$$

where $h_{\mu\nu}$ is the induced metric on the boundary $\partial\mathcal{M}$, $\mathcal{L}_n$ is the Lie derivative along the unit normal n_μ,

$$\mathcal{L}_n h_{\mu\nu} = n^\alpha \partial_\alpha h_{\mu\nu} + (\partial_\mu n^\alpha)h_{\alpha\nu} + (\partial_\nu n^\alpha)h_{\mu\alpha}, \qquad n_\mu = \pm\frac{\nabla_\mu F}{\sqrt{g^{\alpha\beta}\nabla_\alpha F \nabla_\beta F}}, \tag{4.29}$$

where $F(x)$ is a level function that defines the hypersurface. For instance, if we take a constant r hypersurface, we will have $F = r - const.$ The $\pm$ signs depend on whether one is taking a timelike or a spacelike hypersurface. For timelike ones, we have a minus sign. Notice that K has exactly the form that we need,

$$K \sim h^{-1}\partial h. \tag{4.30}$$

One can find the induced metric by computing

$$h_{\mu\nu} = g_{\mu\nu} - n_\mu n_\nu. \tag{4.31}$$

In order to compute the thermal partition function, we now know what to do: a path integral on a Euclidean manifold with boundary conditions such that Euclidean time is a circle of proper size β,

$$\tau_\beta \sim \tau_\beta + \beta. \tag{4.32}$$

We impose this boundary condition asymptotically, where the field falls off rapidly enough, so as $g_{\tau\tau} \longrightarrow 1$. In order to actually evaluate this partition function, we will expand around its classical saddle (this is the WKB expansion),

$$Z(\beta) \sim \exp\{-S_E[\bar{g}, \bar{\phi}]\} + \ldots, \tag{4.33}$$

where the barred quantities now correspond to classical solutions to the equations of motion, and the dots are higher order quantum corrections. For instance, calculating the first-order correction is what Hawking basically did to get the Hawking radiation. We will now evaluate this classical contribution for a Schwarzschild black hole, but first note that from this partition function, we can use the usual thermodynamic relations to compute the entropy and the energy of the system,

$$S = (1 - \beta\partial_\beta)\log Z, \tag{4.34}$$

$$E = -\partial_\beta \log Z. \tag{4.35}$$

The Schwarzschild solution

Let us set ourselves now in the Euclidean Schwarzschild solution, which satisfies the aforementioned boundary conditions,

$$ds^2 = \left(1 - \frac{2M}{r}\right) d\tau^2 + \frac{dr^2}{1 - \frac{2M}{r}} + r^2 d\Omega^2, \tag{4.36}$$

where $\tau \sim \tau + \beta$. Notice that this solution is now completely smooth and devoid of singularities – the τ-metric component is not allowed to switch signs as in the Lorentzian case, since $r = 2M$ now constitutes the origin of coordinates. In other words, unlike in the Lorentzian geometry, now it is not possible to cross to $r < 2M$. The spheres are still shrinking, so the full geometry looks like a stretched-out bowl in the horizontal direction. People usually refer to this solution as the Euclidean cigar.

Let us now evaluate the action. The Einstein-Hilbert term that is $\sim R$ will be equal to zero since Schwarzschild is a vacuum solution (there are no matter fields ϕ). So, the only term we need to evaluate is the boundary term. One can see easily by using (4.29) that

$$\int_{\partial\mathcal{M}} \sqrt{h} K = 4\pi\beta(2r_0 - 3M), \tag{4.37}$$

where $F = r - r_0 = r - const.$ We see that this integral will diverge as we take $r_0 \longrightarrow \infty$, which is where the boundary is, so we need to regulate it and subtract the divergence in a suitable way. The method we will use is called *vacuum subtraction*, although it is not the most general method. More generally one uses counterterms that can be systematically obtained in some cases but we will not cover this here.

The "vacuum" in this case is simply Minkowski, since it is the only other solution which satisfies the boundary conditions of fixed periodic time. Adding this new contribution will make

$$S_E[g] = -\frac{1}{16\pi}\int_{\mathcal{M}} \sqrt{g} R - \frac{1}{8\pi}\int_{\partial\mathcal{M}} \sqrt{h} K + \frac{1}{8\pi}\int_{\partial\mathcal{M}} \sqrt{h} K_0, \tag{4.38}$$

where K_0 is associated to the Minkowski subtraction[10]. We will see that subtracting this Minkowski piece will change the finite part, besides cancelling out the singular part. To compute the Minkowski part, we will rewrite the flat induced metric in some convenient way: the boundary metrics of the spacetime and the background must match (they must be contributions to a path integral with the same boundary metric), and so

$$ds^2 = \left(1 - \frac{2M}{r_0}\right) d\tau^2 + r_0^2 d\Omega^2. \tag{4.39}$$

One can check that this truly is flat – the metric coefficients are constants and they can be reabsorbed into the coordinates. With this metric, one obtains

$$\int_{\partial\mathcal{M}} \sqrt{h} K_0 = 8\pi\beta(r_0 - M + \mathcal{O}(r_0^{-1})). \tag{4.40}$$

[10] In principle one should also add a bulk term R_0. But for the case at hand we can remove all bulk terms in this equation and only keep the boundary term.

Putting everything together, we obtain for the Euclidean action

$$S_E = \frac{\beta M}{2}, \text{ where } \beta = 8\pi M. \tag{4.41}$$

Thus the thermal partition function, or leading approximation to the path integral, is

$$Z(\beta) \sim \exp\{-4\pi M^2\} = \exp\left\{-\frac{\beta^2}{16\pi}\right\}, \tag{4.42}$$

from which we can compute (and reassure ourselves) that

$$S = (1 - \beta\partial_\beta) \log Z = 4\pi M^2, \tag{4.43}$$

$$E = -\partial_\beta \log Z = M. \tag{4.44}$$

Given that the area of the black hole is $A = 4\pi r_h^2 = 16\pi M^2$, we see that the entropy matches the expected area formula, and of course, the energy is simply given by the mass.

Exercise: Fill in the gaps in this subsection and verify the main results.

5 Evaporation of a black hole

We saw that black holes are characterized by the same parameters as any other thermal system, with their entropy and temperature defined through certain geometric quantities. We also saw that once quantum mechanics is incorporated, they behave as thermal systems since they can radiate a black-body spectrum. But each quanta that is emitted shrinks the black hole by a bit, and therefore, shrinks its area and the mass. How is this compatible with the second law now, where we said that the area must never decrease?

This 'puzzle' is resolved once we realize that the black hole in this case is not an isolated system – it is losing its entropy through radiation, but this should be compensated by the growing entropy of the radiated quantum fields. Indeed, this is how we came to the realization that black holes must have entropy. Bekenstein proposed that one should apply the *generalized second law* to the whole system[11] which now has *generalized entropy*,

$$S_{\text{gen}} = \frac{A_H}{4G\hbar} + S_{\text{rad}}, \tag{5.1}$$

where S_{rad} is the entropy of quantum fields that are being radiated away. The generalized second law then simply states that

$$\Delta S_{\text{gen}} \geq 0. \tag{5.2}$$

One could still be worried that the change in the area will be greater than the change in the entropy of the matter outside. However, a calculation done by Don Page reassured the world that the change in the matter entropy will always be greater than the area decrease [3]. One can try to argue for massless radiation that

$$\Delta S_{\text{rad}} \simeq \frac{4}{3}|\Delta S_{\text{BH}}|. \tag{5.3}$$

[11] This is really just the usual second law where a bath has been included.

We can see that the entropy produced is of the same order as the entropy of the black hole. The reason is that the evaporation produces a number of photons $N \sim S_{\rm BH}$. The factor $4/3$ comes from the relation $s_{\rm rad} = \frac{4}{3}\epsilon_{\rm rad} T$ between the entropy and energy density of massless radiation ($\frac{P}{V} = p = 1/3\epsilon_{\rm rad}$ and $\epsilon_{\rm rad} = T s_{\rm rad} + p$) at temperature T (in D spacetime dimensions $s_{\rm rad} = \frac{D}{D-1}\epsilon_{\rm rad} T$).

Life(time) of a black hole

Radiating away quantum fields also shrinks the mass of the black hole, so we expect that the black hole will fully evaporate away at some point. Let us compute how long would it take a Schwarzschild black hole to evaporate.

The area of a Schwarzschild black hole is given by $A = 4\pi r^2 = 16\pi G^2 M^2$ as we saw in the previous section. Now, the radiating power P is related to the black hole area and its temperature – for this, we use the Stefan-Boltzman law since we can treat the black hole radiation as black body radiation, and so

$$P = e\sigma A T^4, \tag{5.4}$$

where e is the emissivity of an object and we set it to 1 for an ideal radiator, and σ is the Stefan-Boltzman constant, which depends on the specific (massless) fields that are being radiated. Replacing all the known variables, we get

$$P = \frac{\sigma \hbar^4 c^8}{256\pi^3 G^2 k_B^4 M^2} \equiv \frac{K}{M^2}, \tag{5.5}$$

where K consists of all the constants in the equation. Now, given that the power of the Hawking radiation is the rate of energy loss of the black hole, we can write

$$P = \frac{dE}{dt} = -\frac{dM}{dt}, \tag{5.6}$$

since $dE = -dM$. Equating the above expressions, we get a differential equation for the mass loss of the black hole

$$-\frac{dM}{dt} = \frac{K}{M^2} \quad \rightarrow \quad M^2 dM = -K dt. \tag{5.7}$$

Integrating over M' from M to zero, and t from zero to τ_{ev}, we get

$$\tau_{ev} \propto M^3, \tag{5.8}$$

where τ_{ev} is the evaporation time. We see that the black hole evaporates in a finite time. This calculation is of course very rough since it neglects the backreaction effect that the emission of radiation and loss of mass have on the black hole geometry and on the radiation process itself. These effects should be small when the energy of emitted quanta is much smaller than the mass of the black hole,

$$T_H << M, \tag{5.9}$$

that is, for $M >> M_p$, where M_p is the Planck mass. Therefore, for most of the black hole lifetime the approximation is good, and its mass will reach Planck size in a time $\propto M^3$.

During this evaporation process, one can ask what kind of particles will we get out? The wavelength of a single Hawking quanta is given by

$$\lambda_H = \frac{\hbar}{T_H} \sim GM, \tag{5.10}$$

which is comparable to the Schwarzschild radius of the black hole (it had to be, since this is the only scale in the system). Including numerical factors, one in fact finds $\lambda_H \gtrsim r_h$. Given that the size of the black hole is comparable to or smaller than the wavelength of the radiation, Hawking radiation cannot be traced to any point on the horizon – the image one forms of a black hole from its Hawking quanta is a blurred one. This is unlike for the Sun whose size $\sim 10^9$ m is much larger than the wavelength of the radiation it emits $\sim 10^2 - 10^3$ nm, and so, we can use it to get a detailed image of the Sun.

As for what kind of particles we can get, the black hole will produce any particle that is permitted by local conservation laws, with mass possibly all the way up to the Planck energy. Initially, the black hole will radiate mostly photons, gravitons, and then neutrinos, and as it reaches different mass thresholds, all other particles will be produced: electrons and positrons around $T_H \sim 1$ MeV, mesons at $T_H \sim \mathcal{O}(100)$ MeV, nucleons at $T_H \sim 1$ GeV, Higgs bosons at $T_H \sim 126$ GeV, then X? at ???eV etc. However, given the long evaporation time, the black hole spends most of its lifetime, and releases most of its energy, emitting low-energy quanta in copious quantities. We can see that if we restore the units for the Hawking temperature

$$T_H = \frac{\hbar c^3}{8\pi k_B GM} = 6 \times 10^{-8} \frac{M_\odot}{M} K, \tag{5.11}$$

so for a solar-mass black hole, $T_H \sim 10^{-7}$ K. This is much colder than the temperature of the CMB for instance, $T_{\rm CMB} \sim 3$ K. The smallness of the effect should not be surprising: it is a quantum effect and therefore one expects it to be small for macroscopic objects (and for a black hole, macroscopic means larger than the Planck scale). So, by the time it reaches the threshold to produce more interesting massive stuff, little energy is left and few of these particles are produced.

On the other hand, the entropy of astrophysical black holes is huge:

$$S_{BH} = \frac{c^3}{\hbar G} \frac{A_H}{4} \sim 10^{76} \left(\frac{M}{M_\odot} \right)^2 . \tag{5.12}$$

A single galactic black hole, with $M \sim 10^6 - 10^9 M_\odot$, has more entropy than all the matter and radiation in the universe ($S_{CMB} \sim N_{photons} \sim$ volume of universe in mm$^3 \sim 10^{87}$).

Can we say something about black holes that are finishing their evaporation today? The initial mass of a black hole that started evaporating in the early universe and ends its evaporation today, so $\tau_{ev} \sim 10^{10}$ yr, is $M \sim 10^{15}$ g. Black holes with these masses are necessarily primordial, that is, formed by density fluctuations in the early universe, since astrophysical collapse cannot yield black holes with masses much below a solar mass.

Information in black holes

What can we know about a black hole from the outside? We already saw that some parameters play a role in the laws of thermodynamics, like the mass, charge and spin. In fact, it turns out these are the *only* parameters that we can measure (classically) as asymptotic observers. This is not only true for spherically symmetric black holes, where a few parameters might do the job for standard systems; these parameters are the only ones even for highly non-symmetric black holes. This statement is best known as the "no-hair conjecture", posed by John Wheeler[12] in the 1950s as an indicator that black holes are determined by only a handful of parameters. These are usually said to be conserved charges at infinity, although one can find examples, for instance in higher dimensions, where non-trivial field configurations still exist. Nevertheless, the spirit of the no-hair statement is preserved – only a few parameters determining a black hole are available to asymptotic observers. In effect, the no-hair statement tells us the black hole tends to quickly settle down to a state where the region around the horizon is vacuum.

Therefore, when a system collapses to form a black hole, it appears that almost all of the initial information of the configuration of collapsing matter is lost in the process. The final state only knows about a few aggregate, macroscopic quantities. Classically, there is no way that this information can be retrieved. What about retrieving the information through quantum effects? After all, we saw that Hawking radiation exists – shouldn't all of the information be encoded in such quanta? It has to, otherwise our black hole will evaporate and all of that information will be lost forever. Right? Although, the spectrum of the radiation is thermal, so all of the information we can receive is similarly macroscopic in nature. Thermal spectrum basically says our radiation is so scrambled, we can only deduce the total mass of the black hole, and maybe some conserved charges conjugate to 'chemical potentials' such as angular velocities, electric potentials etc. In other words, we get a mixed state at asymptotic infinity. What's going on then?

Entanglement entropy

The discussion above was realized by Hawking immediately after his paper on particle production in black hole spacetimes. To understand the paradox more precisely, we need to introduce some terms, specifically types of quantum states and how we can distinguish them. There are two types of states in quantum mechanics: pure and mixed. Pure states can be represented by a state vector, while mixed states cannot and they arise for two different reasons. First, when the preparation of the system is not fully known, and thus one must deal with a statistical ensemble of possible preparations, and second when one deals with partitions of the system, which result in entangled states (but one can have pure entangled states as well). If you think about a Bell pair,

$$|\Psi\rangle = \frac{1}{\sqrt{2}} \left(|0\rangle_A |0\rangle_B + |1\rangle_A |1_B\rangle \right), \tag{5.13}$$

[12] He actually coined many cool terms we still use today – like a black hole!

this is a pure, yet an entangled state of the Hilbert space $\mathcal{H} = \mathcal{H}_A \otimes \mathcal{H}_B$. Mixed states consist of an ensemble of pure states ψ_n, each with a probability of occurrence p_n,

$$\rho = \sum_n p_n \left|\psi_n\right\rangle \left\langle\psi_n\right| . \tag{5.14}$$

If $p_n = 1$, we obtain back a pure state, that is, $p_n = 1$ means that we know with a 100% probability that we are in the state ψ_n. Otherwise, we can only say that we have $p_1 = 30\%$ probability of being in ψ_1 and $p_2 = 16\%$ probability of being in ψ_2, and so on (until $\sum_n p_n = 1$). In other words, mixed states carry a certain degree of uncertainty, ignorance about our state.

Another way to see that mixed states carry a degree of ignorance is by explicitly "tracing out" some part of our pure state – the resulting state will be mixed. Consider a quantum system composed of subsystems A and B, fully described by the density matrix ρ_{AB}. The reduced density matrix of subsystem A is then given by a partial trace over the system B,

$$tr_B\ \rho_{AB} = \rho_A. \tag{5.15}$$

For instance, let us take the Bell state in (5.13), and trace over the system B

$$\rho_A = tr_B\ \left|\psi\right\rangle \left\langle\psi\right| = \sum_{i=0}^{1} {}_B\left\langle i|\psi\right\rangle \left\langle\psi|i\right\rangle_B = \frac{1}{2}\left(\left|0\right\rangle \left\langle 0\right|_A + \left|1\right\rangle \left\langle 1\right|_A\right). \tag{5.16}$$

We see that we obtained a mixed state after tracing out. Note that tracing out does not constitute a physical evolution of a system – it is a mathematical way of expressing our ignorance about a subsystem. Now, in quantum mechanics, it is forbidden to *evolve* a pure state to a mixed state. This would be tantamount to losing information, which cannot happen if our evolution is unitary – in other words, probabilities must add up to 1. So, if I start with a pure state, I must end up in a pure state. But how can we tell?

The inherent ignorance in our mixed states inspired a notion of entropy with which we can tell the difference between pure and mixed states. The entropy in question is *entanglement entropy*, also known as von Neumann entropy. Note that this entropy is not the same as the coarse-grained entropy talked about in the introduction. This von Neumann entropy is *fine-grained* since it encompasses knowledge about our microscopic system. That is, it quantifies our ignorance about the precise quantum state of the system. It vanishes for a pure state, indicating complete knowledge of the quantum state. However, for mixed states, it gives a non-zero value since we do not know everything about our system. Take, for instance, the density matrix (5.16); it is easy to show that the von Neumann entropy will be $\ln 2$ in that case.

On the other hand, when calculating coarse-grained entropy, we fix only some macroscopic data, like the energy or the temperature, and we stay ignorant about the underlying microscopic details. To be more precise, we start with some density matrix ρ describing the system, but we do not measure all observables, we only measure a subset of simple, or coarse-grained observables A_i. Then we obtain the coarse-grained entropy in a few steps:

first, we consider all possible density matrices $\bar{\rho}$ which give the same result as our system for the observables that we are keeping fixed,

$$tr\rho A_i = tr\bar{\rho}A_i. \tag{5.17}$$

We then compute the von Neumann entropy $S[\bar{\rho}]$, and maximize it over all possible choices of $\bar{\rho}$. The A_i are often chosen to be a few observables, say the approximate energy and the volume. The thermodynamic entropy is obtained by maximizing the von Neumann entropy among all states with that approximate energy and volume.

Let us show that the (fine-grained) von Neumann entropy does not increase with time[13]. The only thing we will assume is that our evolution is unitary, that is

$$\rho(t) = U(t)\rho_0 U^\dagger(t), \tag{5.18}$$

where $U(t) = e^{-iHt}$ with H the Hamiltonian of our system, and $UU^\dagger = 1$. The von Neumann entropy is given by

$$S_{vN} = -tr\rho\log\rho, \tag{5.19}$$

where ρ is the density matrix associated with our system. In the case we have a pure state $|\psi\rangle$, the density matrix can be written as $\rho = |\psi\rangle\langle\psi|$ and one can show that the von Neumann entropy is zero. Note also that the density matrix is simply a list of probabilities, and so $tr\rho = 1$, which is simply to say that probabilities are preserved.

Let us evolve our density matrix,

$$\frac{d\rho}{dt} = iH\rho(t) - i\rho(t)H = -i[\rho(t), H]. \tag{5.20}$$

This equation is simply the Ehrenfest theorem, where the initial density matrix does not depend on time. Now we can evolve our entropy,

$$\frac{dS}{dt} = -tr\left(-i[\rho, H]\ln\rho + \rho\rho^{-1}\frac{d\rho}{dt}\right), \tag{5.21}$$

and notice that the trace can go beyond the derivative (it is just a sum after all), so $tr\frac{d\rho}{dt} = \frac{d}{dt}(tr\rho) = 0$, and

$$\frac{dS}{dt} = itr\left([\rho, H]\ln\rho\right) = itr\left(\rho H\ln\rho - \rho\ln\rho H\right), \tag{5.22}$$

where in the last line we used the cyclicity of the trace to move the Hamiltonian to the right. Now we see

$$\frac{dS}{dt} = itr\left(\rho[H, \ln\rho]\right) = -tr\left(\rho\rho^{-1}\frac{d\rho}{dt}\right) = 0, \tag{5.23}$$

since we can use the same argument as before. Therefore, von Neumann entropy is preserved under time evolution. Note that this is not in contradiction with the second law since the entropy still does not decrease – it also just does not increase. This is because we have not

[13] One can show that the second law holds for thermal states using a notion of relative entropy; a quantum-information oriented perspective on this is given here.

done any coarse-graining in our system and all the knowledge is preserved in the density matrix.

Exercise: Fill in the gaps in the calculations above and calculate the fine-grained von Neumann entropy of (5.16).

The beginning of a paradox

Recall how entanglement arises physically: the entangled particles must come from the same source or have some sort of an interaction in order for them to be entangled. And now we see why we went through the trouble of defining entanglement entropies and so forth – pair production is a process that involves a same source for the particles, that is, the vacuum. So, when we derived the Hawking radiation as pair production, we see that those pairs are actually entangled (and one can show this rigorously through a proper calculation).

Knowing that we now have entangled particles to deal with, we can use the methods described above to tackle the problem of black hole evaporation. We will start by identifying that we are dealing with subsystems – black hole A and the radiation B. If this process was unitary, then we could have a description as shown in Fig. 1. The global initial state was pure, and it evolved into another globally pure state. A way of representing that the evolution was unitary is through a plot of the von Neumann entropy for both subsystems, as shown in Fig. 2. Plotting the unitary entropy evolution against time is known as the Page curve, and it was first discussed by Don Page. The Page curve tells us how the entropy of subsystems must evolve if the evolution of the entire system is unitary.

Now, let us take a black hole that was formed by a shell of matter in some pure state. As time goes on, the quantum state of the radiation field outside becomes more and more mixed, which we can quantify by saying that its entanglement entropy is increasing. This may not seem so bad at first, since after all in looking at the late radiation we are looking just at the part of the state which is outside of the black hole horizon. Eventually, the entanglement entropy of the radiation is bigger than the area of the black hole – this is the moment when the paradox emerges. We saw that in the unitary model described in Fig. 1, the entanglement entropies always have to be the same during the evolution – it is basically telling us that the purification of modes always exists. If one entropy becomes larger than the other, this indicates that there are some modes which cannot be purified anymore; they do not have their partner in other words. The moment this happens is called the Page time, and it is roughly around the time when the black hole has evaporated $\sim$ half of its mass. As the black hole evaporates further, it decreases in size until at some point it becomes Planckian, and the entanglement entropy of the radiation field outside still continues to increase. This is how we get the evolution of a pure state to a mixed state.

Proposals for a resolution

Clearly, there is something wrong with the picture we were using (and the much more sophisticated picture that people actually calculated for; we simply presented a toy argument). However, the "beauty" of the paradox lies in apparent innocuous assumptions made throughout Hawking's calculation. Hawking assumed three things in total:

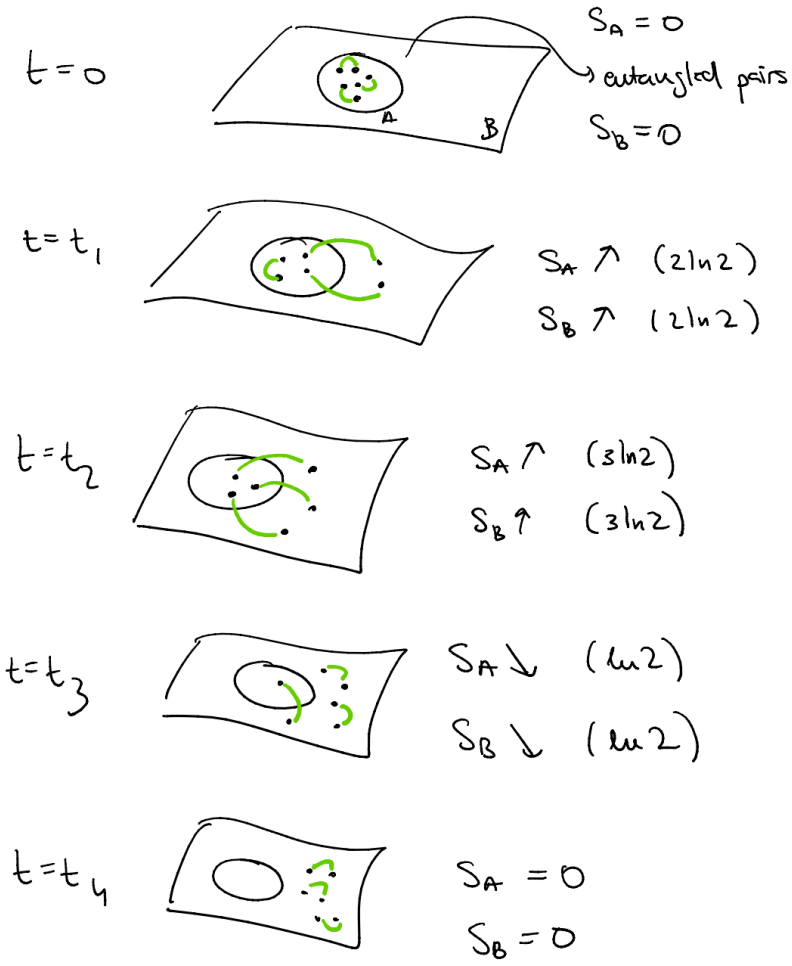

Figure 1: Unitary evolution of radiation transfer.

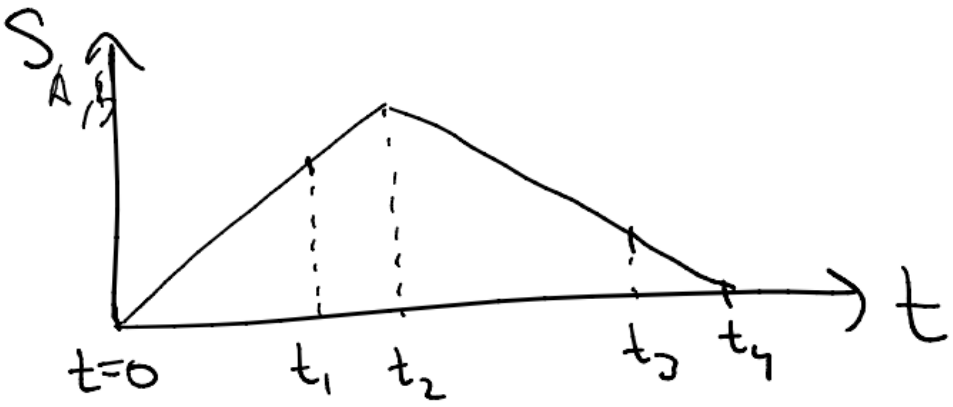

Figure 2: The Page curve.

1. The black hole entropy is given by its area measured in Planck units;
2. The equivalence principle should hold for large black holes, where large = not Planckian;

that is, freely falling observers (and fields) experience the horizon as a smooth place, locally indistinguishable from Minkowski space;

3. Quantum physics is valid, even if black holes are involved.

Listing the assumptions in this way, it is difficult to say which one of these should be dropped. This is the main reason why this calculation has been bothering (and exciting) physicists for over 50 years now. And throughout these 50 years, many possible resolutions have been proposed. We will not go through the details of proposed resolutions in these lecture notes; for this, I highly recommend reading [4] and consulting other references within. We will simply outline three possible directions[14].

Remnants. One popular idea at the time was that maybe the black hole never evaporates fully, but leaves behind a Planckian-sized remnant. Suffice to say that this possibility has been ruled out by now on many grounds, but the simplest one is to see that the problem already begins around the Page time – what happens after it does not matter if the ever-growing entropy is not controlled by that time. Another reason why they are ruled out is due to the fact that, as stable objects of arbitrary amount of information, they would have an enormous amount of entanglement entropy, but their size would be only Planckian. This clearly violates the entropy bound discussed in the introduction.

Corrections due to backreaction. Another popular idea was thinking that Hawking simply did not take into account the backreaction of the metric on the modes: as more radiation comes out, the more effect they have gravitationally, so maybe that can help change the outcome of the calculation. Corrections really do exist and they bring small deviations from the exactly thermal spectrum. However, the entanglement between the interior and the exterior of the black hole grows monotonically along the evaporation. If backreaction results in a small modification in the individual emission process, the entanglement will still keep growing, only now at a slightly lower rate. So it does not seem possible that backreaction, which must be small for large black holes, can effect the reversal in the growth of the entanglement at mid-evaporation that is required to maintain unitarity of the entire process. In fact, this was proven by Samir Mathur in 2009 in [5]. This way, the paradox has been made into a rigorous theorem, (using something called "strong subadditivity of entanglement entropy"), so that it is stable against all subleading corrections. In other words, to overcome the large deviation from the unitary Page curve, one would need large corrections to the calculation.

Fuzzballs, firewalls, brick walls and other -alls. A very popular idea is to simply replace the black hole with something would-be more fundamental, or in other words, something that does not allow for a smooth horizon to form. Although different in their origin and their purpose, all of these proposals suffer from one main issue: what prevents a black hole from forming? By all means of low-curvature, classically valid General Relativity, a black hole is the most natural end-state of a collapse of a massive star. Quantum effects coming into play at such large and arbitrary scales seems very unusual – but of course, the

[14] There are many.

whole paradox is quite unusual. Overall, the hope is that this paradox will shed light on new physics that we have been all waiting for. And in a way, it did.

Holographic theory of gravity

The black hole information problem stirred up a lot of research throughout its fruitful life. Given that all the proposed resolutions require some deviation from our standard assumptions about gravity and quantum physics, it was clear that, whatever the resolution turned out to be correct, it will lead us down the path to a theory of quantum gravity.

But the biggest hint that our standard understanding was incomplete, to say the least, came before all the proposed resolutions. Namely, the biggest hint came already from Bekenstein's insight on the nature of the black hole entropy. Specifically, its scaling with size.

Recall how entropy behaves in standard thermodynamic systems. The change in the entropy is given by the change in the energy, multiplied by a factor of temperature. Now, the temperature is an *intensive* quantity: it reflects the average speed of molecules in our system, but it does not depend on the total amount of molecules. On the other hand, the energy is an *extensive* quantity – adding more molecules means adding more energy to the system since each molecule carries energy with it. Entropy is therefore, an extensive quantity as well: it simply counts the number of degrees of freedom of the molecules, roughly. As we see, extensive quantities are those that depend on the "amount of stuff" we have. Adding more degrees of freedom means I have to rearrange them somehow, and for standard thermodynamic systems this means that each degree of freedom occupies a box of a certain volume in phase space. In other words, the more volume we have, the bigger extensive quantities are.

But the scaling of the black hole entropy is different: the entropy scales with the area of the black hole, not the volume. In fact, we do not know of any thermodynamic systems that behave this way – they all obey the volume law. If the entropy scales with the area, that means we have to seriously rethink the way these degrees of freedom are arranged. This also implies that we have far less degrees of freedom than we initially thought, and that their distribution must be somehow highly non-local.

This observation led Gerard 't Hooft to argue that the true space of quantum states in a finite region must be finite dimensional and associated with the lower-dimensional boundary of the region rather than the volume. He made an analogue with the way a hologram would work – hence the *holographic theory of gravity*. A hologram, recall, manages to reconstruct a three-dimensional object by using only a two-dimensional representation. That is, a lower-dimensional object has all the information necessary to reconstruct an object of one dimension more, but that information is distributed in a highly non-local way. This holographic point of view was then developed further and understood within the framework of string theory by Leonard Susskind.

There is one more observation we can make [6]: even though we do not know of any thermodynamic systems that have their entropies scaling with the area, we know that entanglement entropies *can* exhibit this behaviour. Take for instance, a snapshot in time of some quantum fields evolving on some background, for example, in Minkowski space.

Imagine that we draw an imaginary surface of area $\mathcal{A}$ in this background, and now we want to calculate the entanglement entropy associated with the fields within the surface, and those outside of it. Now we can ask a question: how will the entropies scale? First thing to note is that the information must be preserved – just like with Hawking radiation, whatever is inside the surface must purify the stuff outside; in other words, it must carry the same entanglement entropy. So, it is clear that the entropy cannot scale with the volume of the surface, since then how can the entropy for the fields outside scale in the same way; simply put, they have different volumes. The only thing these fields have in common is the area of the surface, and so, the entropy must be proportional to that area.

Therefore, it seems that the coarse-grained entropy of a black hole might have an interpretation of an entanglement entropy of some quantum fields across a surface of the same area as the black hole. The answer to what quantum fields and in what way are they holographically related to the black hole, was provided by Juan Maldacena when he discovered what is now known as the AdS/CFT correspondence. Here AdS stands for Anti-de Sitter spacetime, and CFT is a conformal field theory, which is a special type of a field theory with no intrinsic scale. In a nutshell, the correspondence tells us that gravity in d dimensions is equivalent to strongly-coupled quantum field theory in $d-1$ dimensions[15]. This correspondence encodes in a precise way the holographic ideas of 't Hooft and Susskind, and it gives us a completely new way of thinking about quantum gravity. Given that these lectures must end at some point, I highly recommend reading [7] as a good first introduction to the physics of AdS/CFT.

Here, let us briefly come back to the question of black hole information paradox – what can we say about it given this new principle? Since gravity is now *dual* to a theory of quantum fields, and such a theory is manifestly unitary, then that means that our gravity theory must evolve in a unitary way as well! For instance, we now know how to obtain the Page curve thanks to the insights that the boundary theory provided; a less-technical review of this progress can be found in [8]. However, we still do not know *how* to implement that unitarity, nor how to extract the physics of Hawking evaporation. In other words, there are many exciting questions left to be answered – possibly by the future Petnica participants.

[15] Technically, it says that Type IIB superstring theory on $\mathrm{AdS}_5 \times \mathrm{S}^5$ is dual to $\mathcal{N}=4$ supersymmetric Yang-Mills theory in 4 dimensions, but we do not need to get into these details here.

A Coarse-grained evolution

Coarse-grained entropy is the von Neumann entropy with fixed simple observables, like the energy. In other words, we can measure the one-point functions, but maybe not higher-point functions. When the energy is fixed, we are in the microcanonical ensemble.

Let us try and show the second law in the microcanonical ensemble – when coarse-graining has been performed. We start with the von Neumann entropy and we work in a basis where the density matrix diagonalizes,

$$\rho = \sum_n \lambda_n |n\rangle \langle n| , \qquad \ln \rho = \sum_n \ln \lambda_n |n\rangle \langle n| , \tag{A.1}$$

where λ_n are now probabilities in this diagonalized matrix: probability that the system could be found in the n-th state, and n are the eigenstates. An easy setup to show that the entropy grows is when we are in the microcanonical ensemble – this will put a restriction on n, that is, we will work in the energy eigenbasis, and it will also put a restriction on λ_n's. In this basis (A.1), we have

$$S = -tr\rho \ln \rho = -\sum_{n,m} \langle n| \rho |m\rangle \langle m| \ln \rho |n\rangle = -\sum_n \lambda_n \ln \lambda_n. \tag{A.2}$$

Let us now write the evolution of the entropy,

$$\frac{dS}{dt} = -\sum_n \frac{d\lambda_n}{dt} \ln \lambda_n, \tag{A.3}$$

and we see there is no second term since the probabilities are conserved, $\sum_n \lambda_n = 1$ (this is the same reasoning as in the quantum case). Now, to see the evolution of this probability λ_n, we will need to make some constraints. First, because we want this evolution to be compatible with equilibrium physics, we require that the evolution must halt if the probabilities are the same: the entropy grows only until reaches equilibrium. Second, this evolution should tend towards the equilibrium, so if $\lambda_\beta > \lambda_\alpha$, then $\frac{d\lambda_\alpha}{dt}$ must grow – probability to reach an equilibrium state should be the desired one. With these assumptions in mind, we can write

$$\frac{d\lambda_\alpha}{dt} = \sum_\beta k_{\alpha\beta}(\lambda_\beta - \lambda_\alpha), \tag{A.4}$$

where the function in brackets could in principle be any odd-powered[16] polynomial with the same difference $\Delta\lambda = \lambda_\beta - \lambda_\alpha$, but we will be satisfied with the linear derivation. One can also view this as being close to the equilibrium so that one can expand the function for small $\Delta\lambda$. We can also write the evolution with reversed indices; one can show that the coefficients $k_{\alpha\beta}$ must be the same due to unitarity,

$$\frac{d\lambda_\beta}{dt} = \sum_\alpha k_{\alpha\beta}(\lambda_\alpha - \lambda_\beta). \tag{A.5}$$

[16] It cannot be an even-powered function since one requires a change of the sign of the left hand side, that is, we require our probabilities to evolve towards equilibrium.

Now we can rewrite

$$\sum_n \frac{d\lambda_n}{dt} \ln \lambda_n = \frac{1}{2} \left(\sum_\alpha \frac{d\lambda_\alpha}{dt} \ln \lambda_\alpha + \sum_\beta \frac{d\lambda_\beta}{dt} \ln \lambda_\beta \right) \tag{A.6}$$

and get for the entropy

$$\frac{dS}{dt} = \sum_{\alpha,\beta} k_{\alpha\beta} (\lambda_\beta - \lambda_\alpha)(\ln \lambda_\beta - \ln \lambda_\alpha). \tag{A.7}$$

Since these two brackets always have the same sign (the log is a monotonic function), we get that the entropy will always increase, unless $\lambda_\alpha = \lambda_\beta$ in which case $dS = 0$.

References

[1] J. Schwinger, "On Gauge Invariance and Vacuum Polarization," *Phys. Rev.* **82** (Jun, 1951) 664–679. https://link.aps.org/doi/10.1103/PhysRev.82.664.

[2] S. M. Carroll, *Spacetime and Geometry: An Introduction to General Relativity*. Cambridge University Press, 2019.

[3] D. N. Page, "Time dependence of Hawking radiation entropy," *Journal of Cosmology and Astroparticle Physics* **2013** no. 09, (Sep, 2013) 028–028. https://doi.org/10.1088%2F1475-7516%2F2013%2F09%2F028.

[4] D. Harlow, "Jerusalem lectures on black holes and quantum information," *Reviews of Modern Physics* **88** no. 1, (Feb, 2016) . https://doi.org/10.1103%2Frevmodphys.88.015002.

[5] S. D. Mathur, "The information paradox: a pedagogical introduction," *Classical and Quantum Gravity* **26** no. 22, (Oct, 2009) 224001. https://doi.org/10.1088%2F0264-9381%2F26%2F22%2F224001.

[6] M. Srednicki, "Entropy and area," *Phys. Rev. Lett.* **71** (1993) 666–669, arXiv:hep-th/9303048.

[7] V. E. Hubeny, "The AdS/CFT correspondence," *Classical and Quantum Gravity* **32** no. 12, (Jun, 2015) 124010. https://doi.org/10.1088%2F0264-9381%2F32%2F12%2F124010.

[8] A. Almheiri, T. Hartman, J. Maldacena, E. Shaghoulian, and A. Tajdini, "The entropy of Hawking radiation," *Rev. Mod. Phys.* **93** no. 3, (2021) 035002, arXiv:2006.06872 [hep-th].