Black Hole Thermodynamics: Then and Now

Edward Witten, IAS

Majorana Lecture 1, University of Naples, June, 2024

I will start today by reviewing some of the high points of black hole thermodynamics as it developed originally in the 1970's. Then in the second half of the lecture, I will turn to a more modern perspective.

Broadly the two parts of the lecture correspond to two points of view about entropy. "Entropy" can mean thermodynamic entropy and then it obeys a Second Law of Thermodynamics:

$$\frac{\mathrm{d}S_{\mathrm{thermal}}}{\mathrm{d}t} \geq 0.$$

Alternatively, one can adopt a microscopic point of view about entropy. Though we could start with classical physics, for brevity I will consider the quantum case. A quantum mechanical system in a general, possibly "mixed" state has a density matrix ρ and its von Neumann entropy is defined by

$$S_{\mathrm{vN}} = -\mathrm{Tr}\,\rho \log \rho.$$

Under unitary quantum mechanical evolution, $\rho \to e^{-\mathrm{i}Ht} \rho e^{\mathrm{i}Ht}$, so

$$\frac{\mathrm{d}S_{\mathrm{vN}}}{\mathrm{d}t} = 0.$$

The relation between the two is that S_{thermal} is the largest that S_{vN} can be given the macroscopic state of a system (its energy, for instance):

$$S_{\text{vN}} \leq S_{\text{thermal}}.$$

If one looks like a system that appears thermal, it is very hard to know if it is truly in a thermal state with $S_{\text{vN}} = S_{\text{thermal}}$, or if it is actually in (or almost in) a typical microstate drawn from a thermal ensemble, in which case $S_{\text{vN}} \ll S_{\text{thermal}}$.

To explain in more detail the inequality $S_{\mathrm{vN}} \leq S_{\mathrm{thermal}}$: Consider a quantum system that has N states with observed values of the macroscopic observables (such as energy, charge, ...). A general density matrix ρ of such a system can be diagonalized

$$\rho = \sum_i p_i |i\rangle\langle i|, \quad \sum_i p_i = 1,$$

which one can interpret to mean that the system is in state $|i\rangle$ with probability p_i. Then

$$S_{\mathrm{vN}} = -\sum_i p_i \log p_i.$$

This vanishes for a pure state (one of the $p_i = 1$ and the others are 0). A simple exercise with Lagrange multipliers shows that under the constraint $\sum_i p_i = 1$, $p_i \geq 0$, the quantity S_{vN} is maximal if all $p_i = 1/N$ – in other words, if all states consistent with macroscopic observations are equally probable. This is a (microcanonical version of) an equilibrium state with the given macroscopic observables. Its von Neumann entropy – namely $\log N$ – is the thermodynamic entropy of such a system.

“Coarse-graining” and forgetting all microscopic details of a system that are not captured by macroscopic observables such as its energy and density will replace the actual density matrix of a system by a thermal one and reduce to the case

$$S_{\mathrm{vN}} = S_{\mathrm{thermal}}.$$

Black hole thermodynamics, however, began with thermodynamic entropy, so we will begin there.

Jacob Bekenstein (1972), inspired by questions from his advisor John Wheeler, asked what the Second Law of Thermodynamics means in the presence of a black hole.

The Second Law says that, for an ordinary system, the “entropy” can only increase. However, if we toss a cup of tea into a black hole, the entropy seems to disappear. Bekenstein wanted to “generalize” the concept of entropy so that the Second Law would hold even in the presence of a black hole. For this, he wanted to assign an entropy to the black hole.

What property of a black hole can only increase? It is *not* true that the black hole mass always increases. A rotating black hole, for instance, can lose mass as its rotation slows down. But there is a quantity that always increases: Stephen Hawking had just proved the "area theorem," which says that the area of the horizon of a black hole can only increase. So it was fairly natural for Bekenstein to propose that the entropy of a black hole should be a multiple of the horizon area. For example, for a Schwarzschild black hole of mass M

$$\mathrm{d}s^2 = -\left(1 - \frac{2GM}{r}\right)\mathrm{d}t^2 + \frac{\mathrm{d}r^2}{1 - \frac{2GM}{r}} + r^2\mathrm{d}\Omega^2,$$

the horizon is at

$$R = 2GM$$

and the horizon area is

$$A = 4\pi R^2 = 16\pi G^2 M^2.$$

Since entropy is dimensionless, to relate the entropy of a black hole to its area, one requires a constant of proportionality with dimensions of area. From fundamental constants $\hbar, c$ and $G = \mathrm{Newton's\ constant}$, one can make the Planck length $\ell_P = (\hbar G/c^3)^{1/2} \cong 10^{-33}$ cm, and the Planck area ℓ_P^2. In units with $c = 1$, Bekenstein's formula for the entropy was

$$S = \frac{A}{4G\hbar},$$

where the constant $1/4$ was not clear in Bekenstein's work and was provided by Stephen Hawking a few years later, in a way that I will explain. For a Schwarzschild black hole

$$S = \frac{4\pi G M^2}{\hbar}.$$

(I sometimes will set $\hbar = 1$, but it is useful to include the $1/\hbar$ in these formulas to explain why black hole entropy is so large.)

Bekenstein's idea was that the entropy of a black hole was supposed to capture the information lost when the black hole was formed – he interpreted it as the logarithm of the number of possible ways the black hole might have formed. Bekenstein proposed a "Generalized Second Law" saying that the "generalized entropy"

$$S_{\rm gen} = \frac{A}{4G\hbar} + S_{\rm out}$$

always increases. Here $S_{\rm out}$ is the ordinary entropy of matter and radiation outside the black hole. The claim is that when something falls into the black hole, $S_{\rm out}$ may go down but $A/4G\hbar$ increases by more. Since $S_{\rm gen}$ increases when a cup of tea falls into a black hole, clearly $S_{\rm gen}$ is a thermodynamic entropy, not a microscopic von Neumann entropy.

Bekenstein made a few tests of the Generalized Second Law. Here is one. Shine photons with a wavelength λ and (therefore) energy $E = 1/\lambda$ on the black hole. The entropy of a single photon is of order 1, for example because the photon has two polarization states. When the black hole absorbs one photon, its mass shifts by

$$\Delta M = \frac{1}{\lambda}$$

so its entropy $S_{\rm bh} = 4\pi G M^2$ shifts by

$$\Delta S_{\rm bh} = 4\pi G((M + 1/\lambda)^2 - M^2) \cong 8\pi G \frac{M}{\lambda}.$$

Bekenstein wanted $\Delta S_{\rm bh} > \Delta S_{\rm out} \cong 1$. He observed that if the black hole is capturing a photon of size smaller than the Schwarzschild diameter $2R = 4GM$ of the black hole, say

$$\lambda << 4GM$$

then

$$\Delta S_{\rm bh} >> 2\pi$$

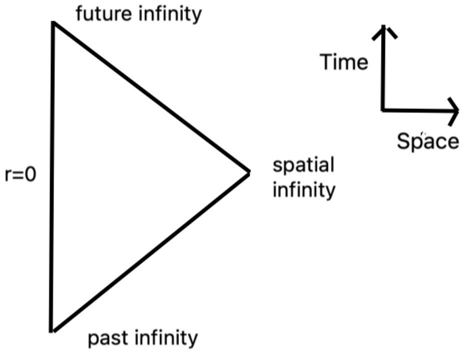
which is satisfactory.

However, Bekenstein did not really get a satisfactory answer if the black hole is absorbing photons of wavelength *larger* than the black hole size – which can happen, though not very efficiently. This question really does not have a satisfactory answer in the framework that Bekenstein was assuming, which was that whatever falls behind the black hole horizon stays there forever. In thermodynamic terms, since Bekenstein assumed that the black hole does not radiate, one would have to assign it a temperature of 0. Thermodynamics says that at equilibrium the changes in energy E and entropy S of a system are governed by

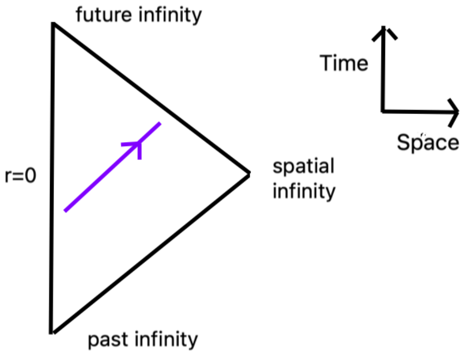
$$\mathrm{d}E = T\mathrm{d}S$$

or $\mathrm{d}S = \mathrm{d}E/T$, so a system with $T = 0$ should have $\mathrm{d}S = \infty$ if $\mathrm{d}E \neq 0$. But Bekenstein wanted to attribute a finite, not infinite, entropy to the black hole. One cannot analyze the absorption of very long wavelength photons by the black hole while ignoring the fact that the black hole is strongly emitting such photons.

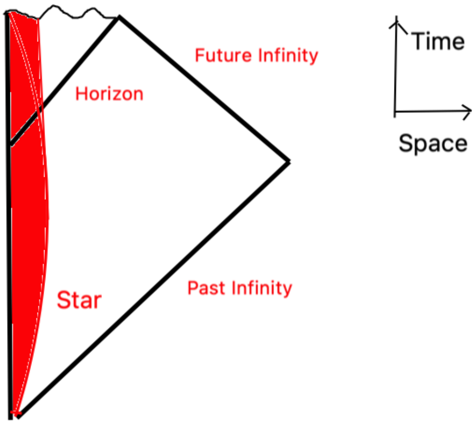
Famously, Stephen Hawking discovered in 1974 that at the quantum level, a black hole is not really black – it has a temperature proportional to $\hbar$. As preparation for explaining what Hawking did, I want to recall the idea of a Penrose diagram as a convenient way to depict and visualize spacetime:



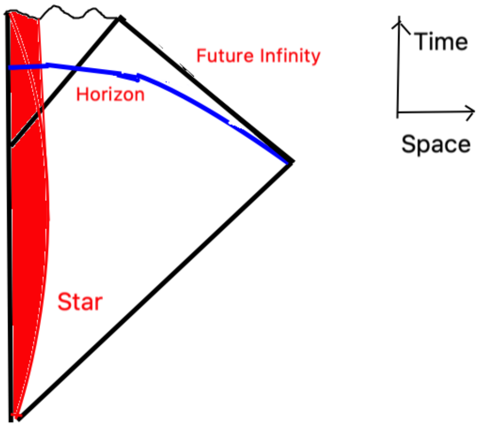
A Penrose diagram (for a spherically symmetric spacetime) is always drawn so that radially ingoing or outgoing light rays are at a $\pi/4$ angle to the vertical:



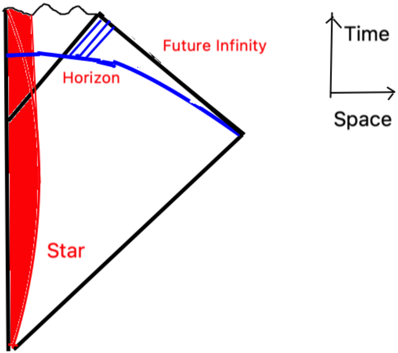
Hawking made his discovery by analyzing the behavior of quantum fields in a black hole geometry:



Measurements that an observer at future null infinity will make can be traced back to initial conditions of the quantum field on a Cauchy hypersurface. It is convenient to pick a hypersurface that crosses the horizon to the future of the collapsing star:

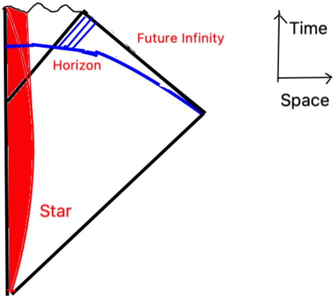


This picture shows signals propagating out at the speed of light from the initial value surface to the observer at infinity:



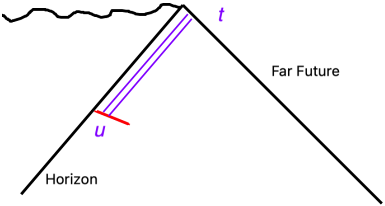
What will the observer see? Part of Hawking's insight was that although the full details of exactly what the observer will see depend on the details of the collapsing star, if we ask what the observer will see *in the far future* after transients die down, we will get a universal answer.

The most important point about this picture is that a signal that is received very late



originated from very close to the horizon. This means that observations made at late times depend on measurements of the state of the quantum fields at short distances. But every state is equivalent to the vacuum at short distances. So the late time observations of the observer probe the vacuum state near the horizon at short distances. That is why Hawking got a universal answer for the late time behavior, regardless of exactly how the black hole formed.

Let u be a coordinate function that vanishes on the horizon on some particular Cauchy slice - it doesn't matter precisely how u is defined.



And let t be the time at which a signal is detected by an observer at infinity. The relation between u and t is

$$t = 4GM \log \frac{1}{u} + C_0 + \mathcal{O}(u),$$

where C_0 is an integration constant. One finds this formula by solving the geodesic equation for an outgoing null geodesic. Rescaling u will only shift the unimportant constant C; nonlinear redefinitions of u will affect the unimportant $\mathcal{O}(u)$ terms.

We can solve the equation $t = 4GM \log \frac{1}{u} + C_0 + \mathcal{O}(u)$ for u:

$$u = C_1 \exp(-t/4GM).$$

At late times, that is if t is large, u is exponentially small. Moreover, $\mathrm{d}u/\mathrm{d}t$ is also exponentially small, which means that a mode observed at infinity will have undergone an exponentially large redshift on its way. A mode of any given energy E that is observed at a sufficiently late time will have originated from a very high energy mode near the horizon. That is why there is a simple answer. A mode of very high energy propagates freely, along the geodesics that I've been drawing.

The observer at infinity probes the radiation by measuring a quantum field $\psi(t)$. A typical observable is a two-point function

$$\langle \psi(t)\psi(t')\rangle.$$

Near the horizon, if the field ψ is for simplicity a free fermion with dimension $1/2$ in the $1+1$-dimensional sense, one would have had

$$\langle \psi(u)\psi(u')\rangle = \frac{(\mathrm{d}u\,\mathrm{d}u')^{1/2}}{(u-u')}.$$

Setting $u = C_1 \exp(-t/4GM)$, we see that for the observer at infinity, this translates to

$$\langle \psi(t)\psi(t')\rangle = \frac{(\mathrm{d}t\mathrm{d}t')^{1/2}}{\exp((t-t')/8GM) - \exp(-(t-t')/8GM)}.$$

This is antiperiodic in imaginary time, that is it is odd under $t \to t + 8\pi GM\mathrm{i}$. That antiperiodicity corresponds to a thermal correlation function at a temperature $T_H = 1/8\pi GM$, which is the Hawking temperature.

In other words, a black hole, after transients that depend on how it was created die down, radiates thermally at a temperature $T_H = 1/8\pi GM$. This explains why Bekenstein had had trouble making sense of the interaction of the black hole with low energy photons. It also lets us confirm the value of the entropy: using

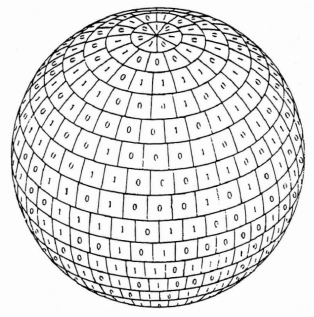
$$\mathrm{d}E = T\mathrm{d}S$$

where $E = M$ and $T = 1/8\pi GM$ gives $\mathrm{d}S = 8\pi GM\mathrm{d}M$ so $S = 4\pi GM^2$. The area of a Schwarzschild black hole is $A = 16\pi G^2 M^2$ so the entropy is

$$S = \frac{A}{4G}.$$

This is how Hawking confirmed Bekenstein's ansatz and determined the constant that was unclear in Bekenstein's work.

Many researchers have thought that, somehow, the entropy $S = A/4G$ means that the black hole can be described by some sort of degrees of freedom that live at its surface – one bit or qubit for every $4G$ of area. For example, in a famous article in 1992, John Wheeler illustrated that idea with this picture:



Hawking's approximations assume that the radius R of the black hole is much bigger than the Planck length $\left(\hbar G/c^3\right)^{1/2} \approx 10^{-33}\,\mathrm{cm}$. For example, a solar mass black hole, with initial mass of order 10^{33} grams, will shrink down to a mass of order 10^{-5} grams before Hawking's approximations break down. We don't really know what happens at that point, but we presume that eventually the evaporation ends and we are left with stable elementary particles.

A fundamental point about Hawking radiation is that the radiation appears to be thermal even though the black hole could have formed from a pure state. This has presented a puzzle that drives much of the research in this field and that to this day is only partly resolved. Hawking's approximations are valid for almost the whole evaporation process and seem to show that the outgoing state is thermal, ultimately with a very large entropy of order $M/T \sim GM^2$ (or about 10^{76} in the case of a solar mass black hole). But if the formation and evaporation of the black hole are described by the ordinary laws of quantum mechanics, then if the initial state is pure, the final state should also be pure.

Concretely the reason that the Hawking radiation seems to be thermal even if the black hole is in a pure state is that the observations of the distant observer amount to observing the quantum fields only outside the horizon. Even if a black hole formed from a pure state – so that we can assume that the state of the whole universe is pure – the quantum fields restricted to only part of spacetime are in a mixed state. That is the essence of the Hawking effect.

Let us remember how this works in ordinary quantum mechanics. We start with a pure state $\psi_{\rm AB}$ in a tensor product Hilbert space $\mathcal{H}_{\rm A} \otimes \mathcal{H}_{\rm B}$. We first make the "pure state" density matrix

$$\rho_{\rm AB} = |\psi_{\rm AB}\rangle\langle\psi_{\rm AB}|$$

The expectation value of any operator $\mathcal{O}_{\rm AB}$ is

$$\langle\psi_{\rm AB}|\mathcal{O}_{\rm AB}|\psi_{\rm AB}\rangle = {\rm Tr}_{\rm AB}\,\mathcal{O}_{\rm AB}\rho_{\rm AB}.$$

Note that $\rho_{\rm AB}$ is the orthogonal projection operator on the state $\psi_{\rm AB}$; in particular, it is hermitian, non-negative, satisfies

$${\rm Tr}_{\rm AB}\,\rho_{\rm AB} = 1$$

and has rank 1.

Now suppose we are only going to observe the subsystem A. That means that we consider only operators of the form $\mathcal{O}_{\mathrm{AB}} = \mathcal{O}_{\mathrm{A}} \otimes 1_{\mathrm{B}}$. The expectation of this operator in the state ψ_{AB} is

$$\mathrm{Tr}_{\mathrm{AB}}\,(\mathcal{O}_{\mathrm{A}} \otimes 1_{\mathrm{B}})\rho_{\mathrm{AB}} = \mathrm{Tr}_{\boldsymbol{A}}\,\mathcal{O}_{\mathrm{A}}\rho_{\mathrm{A}}$$

where

$$\rho_{\mathrm{A}} = \mathrm{Tr}_{\mathrm{B}}\,\rho_{\mathrm{AB}}.$$

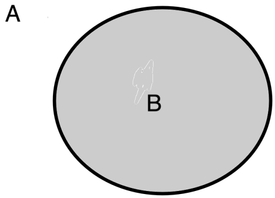
In other words, for measurements on system A only, we can use the density matrix ρ_{A} which is obtained from ρ_{AB} by taking a "partial trace" on $\mathcal{H}_{\mathrm{B}}$.

System A then has a von Neumann entropy

$$S_A = -\mathrm{Tr}\, \rho_A \log \rho_A$$

that in this case is called an "entanglement entropy" because it arose entirely from the entanglement of system A with another system B – with the overall state being pure.

The idea that black hole entropy should be understood in such terms was apparently first put forward by R. Sorkin in 1983 (in a paper that attracted only modest attention at the time). The idea was just the following. In a quantum field theory, divide space into two regions A and B

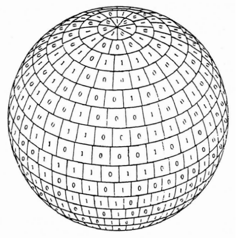


Let Ψ be the vacuum state, and ρ_{A} the "reduced density matrix" of the vacuum for the state Ψ. One can try to calculate the entanglement entropy S_{A}. One finds that it is ultraviolet divergent but the coefficient of the divergence is proportional to the area A of the boundary between regions A and B.

Sorkin's idea, in modern language, was that somehow gravity cuts off the ultraviolet divergence, replacing $A\infty$, which is the answer without gravity, by the Bekenstein-Hawking answer $A/4G$. This makes a lot of intuitive sense, as it matches two ideas:

(1) $A/4G$ is the irreducible entropy of the system for someone who has access only to the region outside the horizon

(2) the divergence in the entanglement entropy is proportional to A because it comes from short wavelength modes near the "horizon," as if (after cutting off the divergence) the density of quantum degrees of freedom on the horizon per unit area is $1/4G$ as in Wheeler's picture:



Susskind and Uglum (1993) made a simple observation that strongly supports the idea of interpreting $S_{\rm out}$ as entanglement entropy. If we interpret $S_{\rm out}$ this way, the generalized entropy defined by Bekenstein is better defined than either term on the right hand side is separately:

$$S_{\rm gen} = \frac{A}{4G\hbar} + S_{\rm out}.$$

The second term has an ultraviolet divergence that Sorkin noted. The first term has a similar problem, because there is an ultraviolet divergence in the relation between the bare Newton constant G_0 and the physical, observed Newton constant G:

$$\frac{1}{G\hbar} = \frac{1}{G_0\hbar} + c\Lambda^2 + \cdots .$$

Here Λ is an ultraviolet cutoff and c is a constant (at 1-loop level, c is independent of $\hbar$). Susskind and Uglum argued that the ultraviolet divergences in $S_{\rm out}$ cancel those in $1/G$. Later authors improved this derivation and confirmed the claim.

Twenty-first century developments have supported these ideas, though leaving us with plenty of mysteries.

I will use the remaining time to try to explain something of the modern perspective.

As I have explained, black hole thermodynamics originated with the mysterious claim that the purely geometric quantity $A/4G$ can be interpreted as an entropy. As I stressed already, this is a thermodynamic entropy which is believed to obey a Generalized Second Law. I also emphasized that the other side of entropy is the purely microscopic von Neumann entropy $S_{\rm vN} = -{\rm Tr}\,\rho\log\rho$. In general the relation between the microscopic entropy and the thermodynamic entropy is an inequality

$$S_{\rm vN} \leq S_{\rm thermal}.$$

Is there a geometric or gravitational formula for the von Neumann entropy?

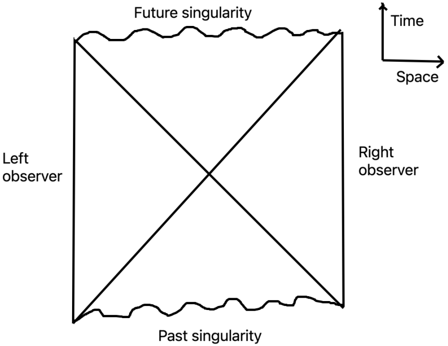
In fact, there is a geometric formula for von Neumann entropy. The first version was discovered by Ryu and Takayanagi (2006) with later refinements by several groups (notably Hubeny-Rangamani-Takayanagi; Lewkowycz-Maldacena; and Engelhardt-Wall).

I will try to motivate the Ryu-Takayanagi formula. The simplest solution of Einstein's equations that describes a black hole is the Schwarzschild solution

$$\mathrm{d}s^2 = -\left(1 - \frac{2GM}{r}\right)\mathrm{d}t^2 + \frac{1}{1 - \frac{2GM}{r}}\mathrm{d}r^2 + r^2\mathrm{d}\Omega^2.$$

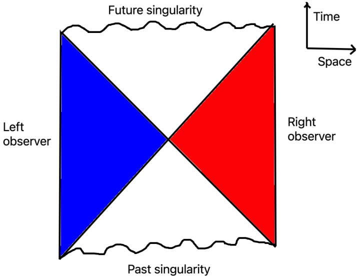
This is a good solution for $r > 2GM$. It describes the gravitational field exterior to any body that is spherically symmetric, such as the Sun (to high accuracy). However, if a spherically symmetric body undergoes gravitational collapse to $r < 2GM$, the solution is incomplete because it has a singularity at $r = 2GM$. Various researchers (Synge, Eddington, Finkelstein, Kruskal, Szekeres, ...) discovered in the 1950's and 1960's that this is only a coordinate singularity and that the solution can be continued to $r < 2GM$.

The full "maximally extended" Schwarzschild solution is quite remarkable – it describes two asymptotically flat worlds connected by a "wormhole" (the wormhole is sometimes called an Einstein-Rosen bridge as Einstein and Rosen discovered part of this picture in the 1930's). Here is a cartoon version (depicted via a "Penrose diagram":



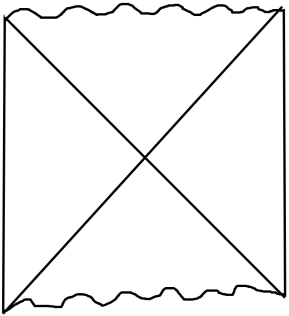
In drawing this cartoon version, I have assumed a small negative cosmological constant, which provides a sort of infrared cutoff.

The part of the spacetime that a "right observer" can see and influence is shown in red



and the part that a "left observer" can see and influence is shown in blue. These regions are bounded by past and future horizons, which are diagonal lines in the picture. As understood by Hartle-Hawking and Israel in the 1970's, and reformulated in the 2000's by Maldacena, in this universe there is a natural pure state (the "thermofield double"), that to an observer on either the left or the right looks exactly thermal at the Hawking temperature.

According to Bekenstein and Hawking, the entropy of this density matrix is the horizon area, and for this ideal solution it does not matter which horizon one picks or where one measures it



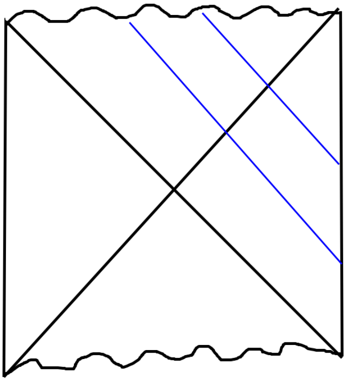
However, suppose two observers, one living on the left, and one on the right, decide to disturb this system. They do this by applying unitary operators, since that is all that one can do to a quantum system.

Here we have a two-sided system AB, where A and B represent the left and right exterior regions of the spacetime, and a pure state, the Hartle-Hawking-Israel state Ψ_{AB}. The combined system has a pure state density matrix $\rho_{AB} = |\Psi_{AB}\rangle\langle\Psi_{AB}|$ but the left or right observer separately sees a thermal density matrix

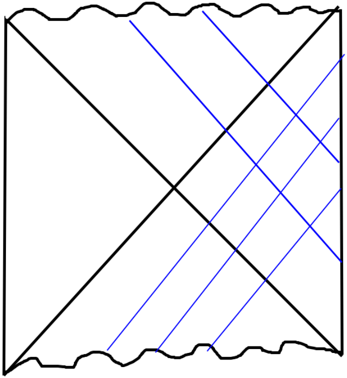
$$\rho_A = \mathrm{Tr}_B\, \rho_{AB}, \quad \rho_B = \mathrm{Tr}_A \rho_{AB}.$$

The left observer can manipulate the system by applying a unitary operator U_A to the left side A of the spacetime, and the right observer can manipulate the system by applying a unitary operator U_B to the right side B.

What happens to the geometry when the left or right observer applies a unitary transformation? For example, the right observer can create particles that head into the interior



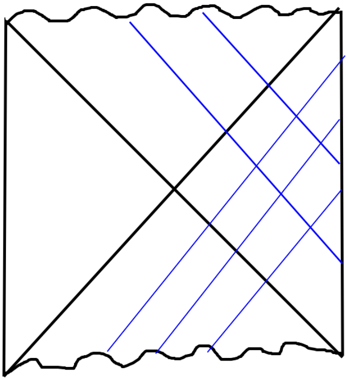
But the problem is time symmetric so the right observer can also create a state that had additional particles in the past



When I say the right observer can throw in "particles," these particles can just as well be macroscopic objects like tables and chairs. Throwing in tables and chairs increases the black hole mass and therefore the horizon area, so it increases the Bekenstein-Hawking entropy

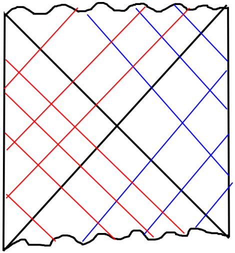
$$S_{\rm thermal} = \frac{A_{\rm horizon}}{4G}.$$

In the new spacetime that the right observer creates, the black hole mass is bigger than before and the horizon moves outward; the past and future horizons are no longer given by the diagonal lines that represented the horizons in the original Schwarzschild solution:



(I have not tried to draw the figure realistically to show the new horizons.)

The left observer can also manipulate the state



All these manipulations change the state:

$$\Psi_{AB} \to U_A \otimes U_B \Psi_{AB}.$$

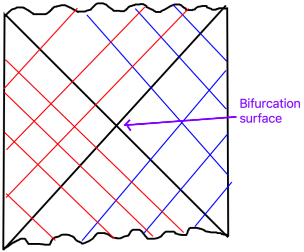
The two observers can do a lot to the system, including increasing the thermodynamic entropy, but there is a limit to what they can do. Unitary evolution will only change the left or right density matrices by conjugation

$$\rho_A \to U_A \rho_A U_A^{-1}, \qquad \rho_B \to U_B \rho_B U_B^{-1}$$

and this will not change the von Neumann or entanglement entropy

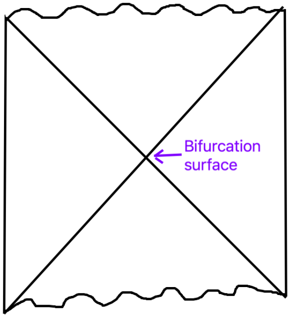
$$S_{\mathrm{vN}} = -\mathrm{Tr}\, \rho_A \log \rho_A = -\mathrm{Tr}\, \rho_B \log \rho_B .$$

If von Neumann entropy can be represented by something in geometry, it will be something that observers on the two sides cannot change. There is only one place in the spacetime that the observers cannot manipulate. It is the “bifurcation surface”



which in the original Schwarzschild solution was the surface where the two horizons cross. The observers on left and right can do nothing to change the area (or even the geometry) of this surface, though they can make all sorts of manipulations on its left or right.

Since the picture has become rather busy, let us go back to where we started:



The only place in this picture that the left and right hand observers cannot change the geometry is the bifurcation surface where the two horizons cross. So it is natural that the von Neumann entropy should be the area of this surface, if it is any kind of area.

In a formula for the von Neumann entropy, we shouldn't refer to the bifurcation surface as "the horizon" because once the spacetime is perturbed on left and right, this surface is not a horizon any longer. But it is an *extremal surface*, that is a surface whose area is stationary among all nearby surfaces. And it remains extremal in the more general spacetime that left and right observers can create by their manipulations. This follows from the Einstein equations.

Thus the idea of Ryu and Takayanagi (in the type of example that I have been describing) is that the von Neumann entropy of the left or right observers is the area of an extremal surface that separates the two observers:

$$S_{\mathrm{vN}} \approx \frac{A_{\mathrm{ext}}}{4G}$$

compared to the thermodynamic entropy which is given by

$$S_{\mathrm{thermal}} \approx \frac{A_{\mathrm{horizon}}}{4G}.$$

(Both formulas have quantum corrections, which is why I've written $\approx$ rather than $=$.) One can use Einstein's equations to prove that the extremal surface is always behind the horizon, consistent with the expectation

$$S_{\mathrm{vN}} \leq S_{\mathrm{thermal}}.$$

There is a subtlety, however: there might be multiple extremal surfaces between the left and right boundaries of the figure. According to Ryu and Takayanagi, in that case the von Neumann entropy is the area of an extremal surface *of minimal area* that separates the left and right:

$$S_{\mathrm{vN}} \approx \frac{A_{\mathrm{min}}}{4G}.$$

Once one realizes that one has to talk about *minimal area* (among extremal surfaces), it becomes clear that there might be a *phase transition* in the location of the Ryu-Takayanagi surface. As a black hole evaporates, there might be a phase transition in the location of the minimal surface that represents the von Neumann entropy.

Just such a phase transition in the field of an evaporating black hole was found in 2019 (Almheiri, Engelhardt, Marolf, and Maxfield; Penington) and resolved some of the paradoxes that arise in trying to reconcile the thermal nature of the Hawking process with quantum mechanical unitarity. I want to give at least some idea of why a phase transition was needed.

Remember that in the real world, a black hole forms in a state of very low entropy. Let us idealize the situation by imagining that the black hole forms in a state of zero entropy, that is, a quantum mechanical pure state $\Psi_B(t=0)$. Then let the black hole evaporate for a long time t. It emits radiation that is very nearly thermal according to Hawking. Hence the entropy of the radiation grows roughly in proportion to t. At time t, the black hole is no longer in a pure state, but if quantum mechanics is correct, then there is a pure state $\Psi_{BR}(t)$ that describes the joint state of the radiation R and the black hole B. At this point, either the black hole or the radiation can be described by a "mixed state" density matrix

$$\rho_B = \mathrm{Tr}_R \, |\Psi_{BR}\rangle\langle\Psi_{BR}|, \qquad \rho_R = \mathrm{Tr}_B \, |\Psi_{BR}\rangle\langle\Psi_{BR}|.$$

Now, it is a basic fact of quantum mechanics that in such a situation the von Neumann entropy of ρ_B and ρ_R are equal:

$$-\mathrm{Tr}_B \, \rho_B \log \rho_B = -\mathrm{Tr}_R \, \rho_R \log \rho_R.$$

In fact, more generally it is true that ρ_B and ρ_R have the same eigenvalues. That is because the canonical form of a joint state $\Psi_{BR} \in \mathcal{H}_B \otimes \mathcal{H}_R$, up to unitary transformations of the two Hilbert spaces $\mathcal{H}_B$ and $\mathcal{H}_R$, is

$$\Psi_{BR} = \sum_i \lambda_i \chi_i(B) \otimes \widetilde{\chi}_i(R),$$

where $\chi_i(B)$ and $\widetilde{\chi}_i(R)$ are states in $\mathcal{H}_B$ and $\mathcal{H}_R$, respectively, that we can assume to be orthonormal. Then the density matrices are

$$\rho_B = \sum_i |\lambda_i|^2 |\chi_i(B)\rangle\langle\chi_i(B)|, \qquad \rho_R = \sum_i |\lambda_i|^2 |\chi_i(R)\rangle\langle\chi_i(R)|,$$

so both density matrices have the same eigenvalues $|\lambda_i|^2$ and the same von Neumann entropy

$$S(\rho_B) = S(\rho_R) = -\sum_i |\lambda_i|^2 \log |\lambda_i|^2.$$

Now in the case of an astrophysical black hole, the evaporation is a very slow process. The temperature that we found was of order $1/R_S$ where $R_S = 2GM$ is the Schwarzschild radius of the black hole. That means that in each time interval R_S the black hole emits an energy of order $1/R_S$ so the luminosity is of order

$$\frac{1}{R_S^2} \sim \frac{1}{G^2 M^2}.$$

Thus the black hole evolves according to (roughly)

$$\frac{\mathrm{d}M}{\mathrm{d}t} = -\frac{1}{G^2 M^2}$$

and the time for its mass to change appreciably is of order

$$\Delta t = \frac{M^3}{G^2},$$

which for a solar mass black hole is very roughly 10^{71} seconds.

Anyway, for this very long time, the von Neumann entropy $S_{\rm vN}$ of the black hole keeps slowly growing, and the von Neumann entropy of the radiation likewise keeps growing Meanwhile the black hole mass is slowly shrinking, and therefore its thermodynamic entropy, which is the Bekenstein-Hawking entropy $S_{\rm thermal} = \frac{A}{4G}$, is slowly decreasing. As Don Page pointed out just over 30 years ago, there is no problem with this until a time called the Page time at which a naive extrapolation would violate the fundamental inequality

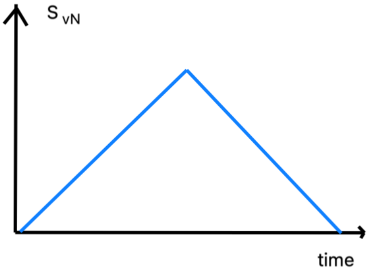
$$S_{\rm vN} \leq S_{\rm thermodynamic}.$$

When we reach the Page time, the claim that the radiation is purely thermal and that $S_{\rm B} = S_{\rm vN}(R)$ is steadily increasing has to break down. Rather, upon further evaporation of the black hole, with its mass continuing to decrease, its thermodynamic entropy decreases and its von Neumann entropy must therefore decrease in order to satisfy the inequality $S_{\rm vN} \leq S_{\rm thermodynamic}$, Page argued that what should happen is that once the Page time is reached and the inequality

$$S_{\mathrm{vN}} \leq S_{\mathrm{thermodynamic}}$$

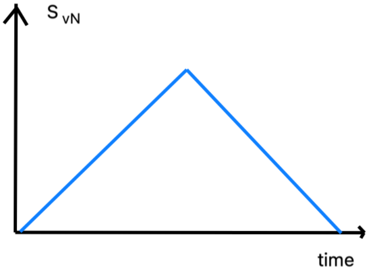
is saturated, it will remain saturated for all times. The idea is that we started with a black hole in a pure state – thus if one is capable of measuring its microstate, one would say that the black hole is far thermal equilibrium – that is, it is far from having a thermal density matrix. But when we reach the Page time, and saturate the inequality $S_{\mathrm{vN}} \leq S_{\mathrm{thermodynamic}}$, the black hole is actually in thermal equilibrium, with (very nearly) a thermal density matrix. Once the black hole reaches true thermal equilibrium, one would expect that this would be maintained as the black hole evaporates adiabatically.

Thus the evolution of $S_{\mathrm{vN}}(B)$ is proposed to follow a "Page curve":



This isn't really an exotic claim, in the following sense. It would be very hard to initialize a burning lump of coal in a quantum mechanical pure state, but if one could do that, then theoretically one would expect its von Neumann entropy to follow a Page-like curve, for basically the reasons that I've explained. The black hole is different in that nature does initialize black holes in states of very low S_{vN}.

In the limit that the initial black hole hole mass is very large, the Page curve



is believed to converge to a true phase transition. And this phase transition has been interpreted as a phase transition in the quantum extremal surface that computes S_{vN}.

According to Penington and also Almheiri, Engelhardt, Marolf, and Maxfield (2019) the quantum extremal surface is the empty surface prior to the Page time, and at the Page time it jumps to a location very near the black hole horizon. In coming lectures, I will speak a little more of the basis for this claim, and also I will describe a little more what has *not* been explained.