What is the wormhole paradigm saying?

Samir D. Mathur[1]

Department of Physics
The Ohio State University
Columbus, OH 43210, USA

Abstract

[1] email: mathur.16@osu.edu

Contents

Lecture 1

1.1 Introduction

The information paradox is a deep problem: it gives a precise conflict between general relativity and quantum theory. Resolving this problem is therefore an essential step in building any unified theory of physics. It is also plausible that the resolution of this puzzle will give new insights about spacetime and matter, and thereby help resolve deep questions in theoretical physics and cosmology.

String theory has provided a theory of quantum gravity which does not suffer from the perturbative loop divergences that had long plagued attempts to quantize gravity. Remarkably, this theory reproduces exactly the Bekenstein entropy of black holes as a count of microstates. In the cases where these microstates have been construted explicitly they have been found to be *fuzzballs* – horizon sized quantum balls with no horizon of their own. Fuzzballs radiate from their surface like normal bodies, so if all microstates are of this type (the 'fuzzball conjecture') then we resolve the information puzzle in string theory.

These lectures are however not about the fuzzball paradigm. They describe my attempts to understand an alternative set of ideas that have been proposed to resolve the puzzle. These ideas involve proposals like wormholes, ER=EPR and spacetime being built by entanglement. It has been argued that using these ideas one can show the Page curve describing black hole information behaves like the Page curve of a normal body, even though these methods may not tell us how information escaped the black hole. For convenience I will refer to this circle of ideas as the 'wormhole paradigm'.

I was initially very excited to see these suggestions, and have spent quite some time trying to understand them. But I have not succeeded in this venture. In particular, I have not been able to extract a derivation of the Page curve from the ideas presented in various papers. In these lectures I will try to explain what I *think* is being claimed, and why I am worried that these claims may be in conflict with general principles of physics. The hope is that people will tell me where I am wrong, and that this will lead to clarity for everyone in the community.

1.2 The question to be addressed: Hawking's information paradox

We will review the information puzzle in Lecture 2, but for now let us recall the essential question in words. The black hole radiates particles, and these radiated particles are entangled with the remaining hole. The entanglement of the radiation at any time with the remaining hole is described by the von Neumann entropy $S_{ent}(N)$ of the radiation, where N is the number of quanta that have been emitted upto the given time. For a normal body, the entanglement $S_{ent}(N)$ at first rises with N, but after around the halfway point of evaporation it starts falling, and reaches zero when the body has completely radiated away. This graph of $S_{ent}(N)$ is called the Page curve. The fact the Page curve

comes down to zero at the endpoint of radiation is essential for consistency: if nothing is left of the radiating body, then there is nothing for the radiation to entangle with; thus $S_{ent}(N_f)$ must become zero at the final step $N = N_f$.

Black holes radiate because their gravitational field makes the vacuum around the horizon unstable. The radiation is pulled out of the *vacuum* around the horizon. In this process of Hawking radiation, the entanglement $S_{ent}(N)$ keeps rising monotonically all the way till the endpoint of evaporation. We then face a problem: the radiation is in a highly entangled state with entanglement $S_{ent}(N_f)$ at this endpoint, but what is it entangled with? This is the black hole information puzzle.

A crucial aspect of this puzzle is that the problematic radiation process happens around the horizon, which is a region of low curvature. For example the spacetime curvature scale at the horizon of the black holes found at galactic centers is of the same order as the curvature at the surface of the earth. There is of course an intense gravitational field at the center of the hole, but this region is causally separated from the horizon region by a large distance – the radius of the hole – which is $\sim 10^8$ Km for the galactic holes. Thus it would seem that the Hawking process appears to be well captured by semiclassical gravity, and does not need a knowledge of what the full quantum gravity theory is. Thus aspect is what makes the information paradox so robust: *any* theory of quantum gravity that yields the usual low energy approximation of semiclassical physics leads to the above described problem of monotonically rising entanglement.

1.3 The claim from the wormhole paradigm (one version)

With this background, let us try to describe the recent claims concerning the Page curve. One statement of the idea, as I understand it, goes as follows:

"Hawking did a semiclassical approximation, and found a monotonically rising Page curve. This semiclassical approximation can be regarded, in a Euclidean path integral description, as corresponding to a 'black hole saddle point' in the semiclassical action of gravity. But there is another saddle point of the semiclassical gravity theory that Hawking missed, and this latter saddle becomes dominant after we cross the halfway point on the Page curve. If we set up the computation of $S_{ent}(N)$ in a Euclidean formalism, this new saddle appears as a 'replica wormhole'. Once we take this new saddle into account, the Page curve comes down like the way it should for a normal body. This method does not tell us how the Hawking radiation process is getting altered to lead to this change in $S_{ent}(N)$, but at least we have a way of seeing that S_{ent} does come down to zero at the endpoint of evaporation."

Here, as in the rest of the wormhole paradigm, we do not make use of any detailed properties of string theory. Many of the arguments are based on tools as simple as JT-gravity, a well-studied 2-dimensional gravity theory with no propagating degrees of freedom. It would therefore appear that the arguments should be easily understood by the whole quantum gravity community, not just those doing string theory. But these

quantum gravity folks appeared as puzzled by the new saddle point as I was. Many relativists believe that the Page curve does keep rising monotonically all the way till the endpoint of evaporation, and we must end up with a violation if quantum theory or having this information locked up into tiny remnants. These relativists did not find any reason to change their view after being presented with this possibility of a new saddle point.

The situation is actually a little more concerning. There can certainly be subleading saddle points in any physical process. But a subleading saddle produces small corrections to the leading order process. We will see that the 'small corrections theorem' says that Hawking's argument for a monotonically rising Page curve is *robust against any small corrections*. The only assumption in this theorem is that once a radiated quantum recedes sufficiently far from the hole (say 100 times the horizon radius) then there is no significant change to its state. (This assumption is just the same one that we make when studying radiation from a normal body: the polarization of a photon radiated from a piece of burning coal does not keep changing after the photon leaves the region of the coal.)

Given this theorem, what could this subleading saddle point be doing? Is it implying that that our quantum gravity theory has a fundamental nonlocality across large distances? I asked this question to several proponents of the wormhole paradigm. Interestingly, there seemed to be a wide spread in the answers. I will attempt to list these answers in schematic form below. These schematic descriptions are only meant to illustrate the large number of diverse threads of thought that seem to have become intermingled under the umbrella of the wormhole paradigm. In later lectures we will investigate each of these threads carefully, to see if we have learnt something new about quantum gravity or black holes.

1.4 Different proposals under the umbrella of the the wormhole paradigm

In this section I will make a first pass at describing my understanding of the various proposals that I have encountered when reading papers on the wormhole paradigm or in talking to people who have contributed to the area. For convenience I group these proposals into four kinds, called A,B, C, D. I emphasize again that the schematic descriptions below just reflect my understanding of what is being suggested by different people; the actual claims might be different since I may just not have understood the claims correctly. (I will update these notes as and when people communicate to me any errors I might have in representing what they said.)

1.4.1 A: Genuine nonlocality in the fundamental quantum gravity theory

Some people have argued that the fundamental gravity theory has a nonlocality across arbitrary distances. Such nonlocalities had appeared in the old picture of Coleman et. al [] where wormholes can connect any point of spacetime to any other point of spacetime.

It was found that such wormholes can be integrated out, but the resulting theory is an 'ensemble averaged theory' where fundamental constants do not have fixed values; rather one has to do a statistical average over the values of these constants. Some recent papers have drawn on similar long distance nonlocalities in addressing the black hole puzzle. But most string theorists say that string theory does not have this kind on nonlocality.

1.4.2 B: Nonlocal wormholes from entanglement

A second point of view I have encountered is the following. We do not have wormholes connecting arbitrary points of spacetime to start with. But take two objects very far apart from each other, and entangle their states. Then gravity generates a geometric connection between these two objects, and it is this effect that is responsible for the departure from Hawking's original computation. Such ideas are encompassed by phrases like 'wormholes from entanglement', 'spacetime is built by entanglement' and 'ER=EPR'.

The difficulty with such ideas, as we will note below, is that they violate the linearity of quantum theory. This linearity implies that entanglement *cannot* generate any dynamical effects that one may try to invoke to avoid Hawking's puzzle.

The responses I have received to this difficulty have been in two different directions:

(a) Gravity is a novel theory, and new rules apply

(b) The spacetime connection emerging from entanglement is an approximate picture that is useful for computing some coarse-grained quantities.

The problem with (a) is that we have a dual description of gravity in terms of a CFT, and it is hard to postulate any novel rules here. With (b), the problem is that the entanglement entropy is a delicate quantity involving all the bits in the system, while the coarse grained approximations typically give averaged quantities like 2-point functions. I have not been able to reconstruct a computation where one actually looks at the entanglement of the actual radiation state and then uses such an approximation to extract an answer for $S_{ent}(N)$.

1.4.3 C: Exact gravity theory vs its semiclassical approximation

Another point of view that I have encountered involves the low energy approximation of the quantum gravity theory. In this view, the exact quantum gravity theory (which could be string theory) has no long distance nonlocalities. But when we take the 'semiclassical approximation' nonlocal terms appear that connect systems that are separated by arbitrarily large distances. These 'wormholes' of the low energy theory give additional contributions to the dynamics of a black hole and lead to a Page curve that comes down at the endpoint of evaporation.

One immediately notes that there is something strange about this suggestion. Consider the situation where the exact theory is a GUTS field theory, and the low energy approximation is QCD. Consider a rock near earth, and another rock near Mars. Each rock is described

exactly by the GUTS theory, and to an excellent approximation by the low energy limit which yields QCD. There were no relevant interactions between the two well-separated rocks in the exact GUTS theory. But equally, we do not expect to find an interaction between the rocks in the low energy approximation which yields QCD. So how did such an interaction emerge in the passage from exact quantum gravity to low energy semiclassical gravity?

To this question I received the response: gravity is special, and novel things can happen. But there is a problem with postulating such novelty. An object in the gravity theory can be described equally well in the dual CFT, and it does not seem possible to have such nonlocal effects appear when taking the low energy limit of two well separated and noninteracting field theories.

We will return to this issue several times, as a general theme in the wormhole paradigm has been the introduction of new rules that are postulated for the semiclassical limit of the exact gravity theory. But we will see that these postulated rules seem to conflict with the usual way we define low energy effective field theory limits from an exact theory. Further, as mentioned above, they appear to conflict with the original idea of AdS/CFT that gravity (described by closed strings) is just Yang Mills theory (described by open strings between branes) written in new variables. To this difficulty, I have been given the response that this original picture of AdS/CFT itself must be given up if we are to understand the newly postulated wormhole terms. Since this issue goes to the heart of how we understand string theory, we will spend some time in later lectures discussing this situation in detail.

One other claim in this context is the following. The Gibbons-Hawking computation of black hole entropy invokes a Euclidean solution that has the shape of a cigar (times a sphere), which is a geometry that we did not originally have in our Lorentzian theory. The proponents of the wormhole paradigm use this example to argue that novel geometries and topologies can appear extensively in gravity, especially in situations involving black holes, and the wormholes that are being assumed between distant objects are just more instances of this phenomenon.

As we will see, the Gibbons-hawking cigar can be shown to have a natural interpretation in terms of standard field theory ideas. But this interpretation does not yield the wormholes being used in the wormhole paradigm.

1.4.4 D: Argument used for normal statistical systems

In statistical mechanics there is a standard computation showing the following. Consider two systems A and B with Hilbert spaces of dimensions N_A, N_B respectively. Then the entanglement between them is bounded by

$$S_{ent} \leq min\{\log N_A, \log N_B\} \tag{1.1}$$

Suppose the system A is a piece of burning coal and B is its radiation. At the endpoint of radiation when the coal is gone (i.e., $N_A = 0$) we will find $S_{ent} = 0$; thus the Page curve curve come down to zero as expected.

For a generic state of the combined system $A + B$ one can write this computation in path integral language. This yields a diagrammatic representation containing a sum of terms. These diagrams resembles wormholes connecting multiple copies of systems A, B in different ways. One can then say that in the gravity theory these connections are 'wormholes in spacetime', It would seem that one has shown that the Page curve comes down to zero at the end of evaporation (since it does for any radiating body), and these wormholes have somehow played a role.

But a more careful look indicates that this argument for the Page curve is circular. It *assumes* that the black hole is a normal body like a piece of coal, with a Hilbert space of dimension N_A given by the Bekenstein entropy

$$S_{bek} = \frac{A}{4G} \tag{1.2}$$

of the hole. If we make this assumption, there is nothing to show: as the hole evaporates down to zero, its Bekenstein entropy goes to zero, and $N_A \to 0$. The entanglement S_{ent} must then also go to zero. The whole point of the information puzzle was that the classical picture of the black hole allows an *arbitrary* number if internal states N_A for the hole even though its mass (and therefore S_{bek}) go to zero at the endpoint of evaporation. (We will see this more explicitly when we discuss the information paradox in the next lecture.)

The 'wormholes' appearing in the above mentioned diagrammatic representation also do not seem to have anything to do with gravity. The action of these wormholes, is given by the weights of states used in the *averaging prescription* used to define the generic state of the system $A + B$, and is therefore not necessarily connected to the gravitational action of the theory.

1.5 What is ER=EPR?

Lecture 2

2.1 The black hole information paradox

Consider a mass m. Einstein taught us that this mass has an intrinsic energy $E = mc^2$. Now place m near another mass M. The gravitational potential energy is *negative*, a fact emerging from the attractive nature of gravity. Schematically, we write the total energy of m as

$$E_T = mc^2 - \frac{GMm}{r} \tag{2.1}$$

We note that at

$$r = \frac{GM}{c^2} \tag{2.2}$$

the total energy E_T of m becomes zero, and for smaller r is is *negative*. This means that our theory has an instability: we can create the mass m for 'free' out of the vacuum. The Hawking puzzle arises from this instability of the vacuum.

Since we are invoking both relativity (to get $E = mc^2$) and gravity (for the potential), we should really be doing all this with general relativity. It turns out that the essential instability noted above is still valid, with an additional factor of 2 in the critical radius (**??**). This radius, which will be the horizon radius for the mass M is therefore

$$R = \frac{2GM}{c^2} \tag{2.3}$$

How does the above instability manifest itself? Even in flat spacetime, the vacuum is filled with virtual particles. For example an electron-positron pair can emerge from the vacuum, last for a short time, and then annihilate away. The energy of this fluctuation $\Delta E \approx 2mc^2$ determines the typical time Δt for which the pair lasts through the uncertainty principle

$$\Delta E \Delta t \lesssim \hbar \tag{2.4}$$

Now suppose the fluctuation happens near the horizon radius. One particle (say the electron) can be inside this radius with net negative energy, while the other (the positron) can be outside with net positive energy, such that the total energy is $\Delta E = 0$. Then (2.4) gives $\Delta t \to \infty$; i.e., the particle pair is on-shell, and need not annihilate. The outer member of the pair floats off to infinity as 'Hawking radiation' while the inner member (carrying net negative energy) falls into the hole and reduces its mass. The process repeats with another pair, and so on, so that the hole slowly evaporates away.

The problem comes because the two members of the created pair are in an entangled state. The entanglement can come from many sources, but for simplicity we look at the spin. Since the particle pair emerged from the vacuum, it will typically be in a singlet state

$$|\psi\rangle_{pair} = \frac{1}{\sqrt{2}}\Big(|\uparrow\rangle_{e^-}|\downarrow\rangle_{e^+} - |\downarrow\rangle_{e^-}|\uparrow\rangle_{e^+}\Big) \tag{2.5}$$

If two systems A and B are in a state

$$|\Psi = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} |\psi_k\rangle_A \, |\chi_k\rangle_B \tag{2.6}$$

then the entanglement entropy of system A with B is $\log N$. Thus the emission (2.5) leads to an entanglement between the radiation and the remaining hole by $\log 2$. After N emissions, the entanglement becomes $N \log 2$ (we will see this more explicitly below). Thus the Page curve describing entanglement as a function of time keeps rising monotonically with each emission, as pictured in fig.**??**(a).

In this process the mass M of the remaining hole keeps going down, as does the radius of the horizon given through (2.3). We have to ask what happens at the endpoint of evaporation. There are two possibilities:

(a) The hole evaporates away completely, restoring the spacetime at its location back to the vacuum state $|0\rangle$. The radiation is still in an entangled state, but there is nothing that it is entangled *with*. This radiation cannot be described by *any* quantum state; it can only be described by its density matrix. The black hole was made from a star in some pure state $|\psi\rangle_M$, but after the evaporation process we are left with something that cannot be described by a usual quantum state. Thus the process of black hole formation and evaporation violates the unitarity of quantum evolution

$$|\psi\rangle_f = e^{-iHt}|\psi\rangle_i \tag{2.7}$$

which takes pure states to pure states. This option was the one assumed by Hawking in his original paper in 1975.

(b) Many others were unhappy with this strong conclusion. They argued that once the black hole reaches planck size ($M \sim m_p$, $R \sim l_p$) some (hitherto unknown) quantum gravity effects stop further evaporation, leaving this planck sized hole as a 'remnant'. Remnants are awkward however. To have an entanglement $N \log 2$ with the radiation, they must have 2^N internal states. Here N is unbounded, since we could have started with a black hole of arbitrarily large size and allowed it to evaporate down to planck size. Thus we need an infinite number of possible internal states for remnants, while having an energy budget $\sim m_p$ and a radius $\sim l_p$. There are difficulties with having an infinite number of objects of this type in a field theory; for example they tend to make all loop diagrams in QFT diverge due to the infinite degeneracy of remnants running in loops.

Given the radical nature of possibility (a), many gravitational physicists have reconciled themselves to remnants, and tried to construct models for them. (One model, which we will encounter later, consists of a baby universe attached to the rest of spacetime by a planck sized neck, carrying the initial matter that made the hole and the infalling members of the Hawking pairs.) But remnants are not possible in string theory, if we accept the conjecture of AdS/CFT duality. Conider string theory on global $AdS_5 \times S^5$,

whose dual is $\mathcal{N}=4$ supersymmetric Yang-Mills on S^3. A remnant at the center of AdS has energy $\sim m_p$ which corresponds to a finite energy in the boundary CFT. An infinite degeneracy for such remnants would imply an infinite number of states for SYM with finite N on a finite volume S^3, something that is known to be not true.

The above problem created by black hole evaporation is known as the black hole information paradox.

2.2 Second iteration

Let us now take a more careful look at how the particle creation process happens. The essential feature of a black hole horizon is that light cones 'point inwards' inside the horizon, so that nothing can escape to the outside. Consider the following three different null geodesics, each directed radially outward:

(i) The geodesic starts a small distance ϵ outside the horizon, and heads radially outwards. This geodesic eventually escapes to $r \to \infty$.
(ii) The geodesic is at the horizon radius R, and directed radially outwards. This geodesic stays at the radius R.
(iii) The geodesic starts a small distance ϵ inside the horizon, and is directed radially outwards. This geodesic gets pulled into the hole and heads to $r=0$.

Thus we see that geodesics 'separate' at the horizon. As a consequence, the spacelike slice around the horizon *stretches* as it evolves. It is a general fact that a change in the spacetime metric creates particles out of the vacuum. Let us see what the nature of the created particles will be. Consider a massless scalar field on our spacetime

$$\Box\hat{\phi}=0 \tag{2.8}$$

This quantum field can be described in terms of a set of masses m with separation a, with springs coupling neighbouring masses with spring constant k. Taking $a\to 0$ and scaling m,k appropriately, we recover the dynamics (2.8) as a set of coupled harmonic oscillators. The vacuum is the ground state of these coupled oscillators. Exciting an oscillator at position x_1 corresponds to a particle at x_1.

Consider the lower slice $\mathcal{S}_1$ in fig.**??**. For simplicity consider just two oscillators: one inside the horizon at $r=r_1$ and one outside at $r=r_2$. These oscillators are coupled to yield the gradient term in (2.8). Consider the lower slice S_1 of fig.**??**. We write the total Lagrangian for thee two oscillators on this slice schematically as

$$\mathcal{L}_{S_1}=\frac{(\dot{\phi}(x_1))^2}{2}+\frac{(\dot{\phi}(x_2))^2}{2}-\frac{K}{2}(\phi(x_1))^2-\frac{K}{2}(\phi(x_1))^2-\frac{1}{2}\frac{(\phi(x_2)-\phi(x_1))^2}{(\Delta s_{12})^2} \tag{2.9}$$

where Δs_{12} is the proper distance between x_1 and x_2. Let the quantum state on this slice $\Psi(\phi(x_1),\phi(x_2))$ be lowest allowed energy state, which we call the vacuum $|0\rangle_{S_1}$ on the slice S_1.

Now consider the time evolution to the upper slice S_2 in fig.**??**. The stretching of slices leads to a larger separation $\tilde{\Delta}s_{12} > \Delta s_{12}$ between x_1, x_2, so that the Lagrangian becomes

$$\mathcal{L}_{S_2} = \frac{(\dot{\phi}(x_1))^2}{2} + \frac{(\dot{\phi}(x_2))^2}{2} - \frac{K}{2}(\phi(x_1))^2 - \frac{K}{2}(\phi(x_1))^2 - \frac{1}{2}\frac{(\phi(x_2)-\phi(x_1))^2}{(\tilde{\Delta}s_{12})^2} \tag{2.10}$$

Thus we have a smaller coupling between the oscillators at x_1, x_2 on S_2.

If the time Δt taken to evolve from S_1 to S_2 is long, then the vacuum state $\Psi(\phi(x_1), \phi(x_2))$ will evolve to the lowest energy state $\tilde{\Psi}(\phi(x_1), \phi(x_2))$ on S_2. This follows from the adiabatic theorem, and is the case when $\Delta t \gg \omega^{-1}$ where $\omega \sim \sqrt{K}$ is the frequency of the oscillators. But if $\Delta t \ll \omega^{-1}$, then the initial state $\Psi(\phi(x_1), \phi(x_2))$ cannot evolve fast enough ton adjust to the new Lagrangian, and remains essentially unchanged. We are then in an excited state of the QFT on slice S_2, and this is the phenomenon of particle creation that we are seeking to understand.

To make things simple, let us assume that we are in the ground state $|0\rangle$ of the Lagrangian (2.9) for $t < 0$, and that the coupling in (2.9) goes to zero suddenly at $t = 0$; i.e., the Lagrangian for $t > 0$ is

$$\mathcal{L}_{S_2} = \frac{(\dot{\phi}(x_1))^2}{2} + \frac{(\dot{\phi}(x_2))^2}{2} - \frac{K}{2}(\phi(x_1))^2 \tag{2.11}$$

The vacuum for (2.11) is just the product of the ground states for the two oscillators

$$|0\rangle_{S_2} = |0\rangle_1 |0\rangle_2 \tag{2.12}$$

But the state $|0\rangle_{S_1}$ cannot suddenly change to $|0\rangle_{S_2}$. The state at $t = 0^+$ remains $|0\rangle_{S_1}$, which has the form

$$|0\rangle_{t^+} = |0\rangle_{S_1} = c_0|0\rangle_1|0\rangle_2 + c_1|1\rangle_1|1\rangle_2 + c_2|2\rangle_1|2\rangle_2 + \dots \tag{2.13}$$

Thus there is some amplitude c_0 to indeed get the vacuum on slice S_2, but there is also a amplitude c_1 to get a pair of excitations – one in the oscillator x_1 inside the horizon and one in the oscillator at x_2 outside the horizon. This is the Hawking pair creation process.

The crucial point is that the state (2.13) is entangled between the oscillator mode inside the horizon at $x-1$ and the oscillator mode outside at x_2. We have taken a scalar field that has no spin, but we still get an entanglement of the *occupation number* of the field modes around x_1, x_2. The particle excitation amplitudes like c_1 in (2.13) become order unity at wavelengths $\lambda \sim R$ where R is the radius of the hole. (Shorter wavelengths have higher frequencies, and their modes evolve adiabatically from the vacuum state on S_1 to the vacuum state on S_2; this makes $c_i \to 0$ for $i > 0$.)

To match with notation used elsewhere, we call the oscillator outside the horizon as b and the one inside as c. For simplicity we keep just the first two terms in the above state and write the entangled state as

$$|\psi\rangle_{pair} = \frac{1}{\sqrt{2}}\Big(|0\rangle_b|0\rangle_c + |1\rangle_b|1\rangle_c\Big) \tag{2.14}$$

The above explanation of Hawking pair creation may appear like a toy model, but it actually captures all the essential features of Hawking's actual computation. A full derivation involves expanding $\hat{\phi}$ in field modes

$$\hat{\phi} = \sum_k \left(\hat{a}_k f_k(x) + \hat{a}_k^\dagger f^*(x) \right) \tag{2.15}$$

and then studying the evolution of the vacuum state $|0\rangle$ on an early-time slice in the black hole geometry. But field modes can be made into wavepackets, and the essential pair creation is captured by toy Lagrangians like (2.9) where the two oscillators describe wavepacket of wavelength $\sim R$ on the two sides of the horizon.

2.3 Third iteration

So far we have not had to write down the black hole metric; all we used was the fact that outward directed geodesics separate at the horizon, leading to a local stretching of spacelike slices. But this description leaves the following question. The pair creation process creates one member of the pair inside the hole. Is it possible that after some time these infalling particles 'fill up' the black hole interior upto the horizon? If so, the quantum state around the horizon would no longer be the vacuum that we assumed on S_1 in the above discussion, and we would not be able to argue for the creation of entangled pairs (2.15).

But as we will now see, there is an *infinite* amount of space inside the black hole. This is the consequence of the existence of a horizon: inside the horizon space and time interchange roles, and spacelike slices inside the hole can be arbitrarily long. To see this, start with the Schwarzschild metric

$$ds^2 = -(1 - \frac{2M}{r})dt^2 + \frac{dr^2}{1 - \frac{2M}{r}} + r^2 d\Omega^2 \tag{2.16}$$

This metric looks time-independent, but that is an illusion. The t coordinates fails at $r = 2M$, and we must use another set of coordinates like Kruskal or Eddington-Finkelstein to cover both the outside and inside of the hole. We will then find that the metric is time *dependent*, and spacelike slices 'stretch' in this metric around the horizon region. In a time dependent geometry the vacuum is generically not stable – one gets particle creation, and this is the Hawking effect.

Let us construct a set of spacelike slices that cover both the inside and outside of the hole. In the region $r > 2M$ a spacelike slice is given by $t = constant$, so we take $t = t_1$ for, say, $r > 3M$. For $r < 2M$ we see that a $r = constant$ is a spacelike slice. We wish to not be close to the singularity at $r = 0$ or the horizon at $r = 2m$, so we take this part of the slice as $r = M$. These two parts can be connected by a smooth spacelike 'connector' segment; (an explicit form for this segment is given for example in []).

We imagine that the black hole was made by the collapse of a shell of massless particles falling in radially. We continue the segment $r = M$ of our slice downwards till it meets

this shell. Inside the shell is just flat spacetime, so we can then bring this spacelike slice smoothly to the origin $r = 0$ inside the shell. This makes one complete spacelike slice S_1, which does not approach the singularity anywhere.

To study evolution we need a 'later slice'. we advance the outer segment to a later time $t = t_2$. We let the 'connector segment' have the same intrinsic geometry as before. We do not advance the $r = M$ segment. Advancing this segment would take us closer to the singularity, and we dont wish to go there. Similarly, we do not change the part of the slice inside the shell. But we see that we must 'stretch' the $r = M$ segment to a longer length to join with the connector segment. With this extra part added to the $r = M$ segment, we have made a 'later' slice S_2 in our geometry.

While these slices look strange in the way they have been sketched in fig.**??**, we can understand their intrinsic nature and the Hawking pair creation by sketching the slices as in fig.**??**

(i) On the first spacelike slice S_1 we have the shell of mass M on the left, in some state $|\psi_M\rangle$. The quantum fields on the ret of the slice are in the vacuum state.

(ii) For the next slice S_2 we advance the $t = constant$ part in time by an amount $t_2 - t_1 \sim M$. The part of the slice carrying the shell $|\psi_M\rangle$ does no advance, so the state of the shell does not change. The $t = constant$ part of the slice also does not change its internal geometry. But a part of the slice does 'stretch', by a length $\sim M$. This stretching creates an entangled pair of quanta b_1, c_1, whose state we may write schematically as (2.14).

(iii) We now advance the time at infinity by a further amount $t_3 - t_2 \sim M$, to get the slice S_3. The shell and the quantum c_1 remain where they werer, since this part of the slice is not advanced. The quantum b_1 moves to larger r under the natural evolution which takes outgoing wavepackets away from the hole. The region in between stretches again by an amount $\sim M$, and creates a new entangled pair b_2, c_2.

After N emissions the state on the spacelike slice is

$$\begin{aligned} |\Psi\rangle &= |\psi_M\rangle \\ &\otimes \frac{1}{\sqrt{2}}\Big(|0\rangle_{b_1}|0\rangle_{c_1} + |1\rangle_{b_1}|1\rangle_{c_1}\Big) \\ &\otimes \frac{1}{\sqrt{2}}\Big(|0\rangle_{b_2}|0\rangle_{c_2} + |1\rangle_{b_2}|1\rangle_{c_2}\Big) \\ &\cdots \\ &\otimes \frac{1}{\sqrt{2}}\Big(|0\rangle_{b_N}|0\rangle_{c_N} + |1\rangle_{b_N}|1\rangle_{c_N}\Big) \end{aligned} \tag{2.17}$$

The radiated quanta $\{b_i\}$ have 2^N possible states, where each state is described by some sequence like $01101\ldots$. The state of the $\{c_i\}$ is the same state. Thus the overall state

has the form (2.6) with $n = 2^N$. The entanglement entropy of the radiation $\{b_i\}$ with the region of the hole is therefore

$$S_{ent} = \log 2^N = N \log 2 \tag{2.18}$$

and we get the rising Page curve of fig.**??**.

2.4 The Page curve of a normal body

For completeness, we recall the nature of the Page curve for a normal body that burns away to nothing. Consider a toy model of a piece of coal onsisting of two atoms, each in an excited state. The radiation process can be as follows:

(i) One excited atom (atom 1) de-excites, emitting a photon. The photon γ and the de-excited atom are in an entangled state:

$$|\psi\rangle = \frac{1}{\sqrt{2}}\Big(|\uparrow\rangle_\gamma |\downarrow_{atom} - |\downarrow\rangle_\gamma |\uparrow\rangle_{atom}\Big) \tag{2.19}$$

Upon this emission, the entanglement S_{ent} of the radiation region with the region of the coal goes from 0 to $\log 2$.

(ii) The second atom (atom 2) de-excites in a similar way. The entanglement S_{ent} becomes $2 \log 2$.

(iii) We are looking for a model where the coal disappears completely. So we imagine that atom 1 now floats out as ash. Now this atom and its entangled photon are both in the radiation region. They are entangled with each other, but not with anything in the region of the coal. The entanglement S_{ent} now is back down to $\log 2$, arising from the entangled of atom 2 with its radiated photon.

(iv) Atom 2 also floats out as ash. There is nothing left at the location of the original coal; the coal has completely radiated away. The entanglement of the radiation with the region where the coal was is now $S_{ent} = 0$.

These steps mimic the Page curve of fig.**??**.

More generally the quanta in the coal can interact and exchange spins before each emission (some examples were studied in [**?**]). But the key point is that as the radiating body reduces in size, it has fewer and fewer internal states, and thus S_{ent} has to become smaller and smaller. By contrast, in the Hawing state (2.17), the mass of the remaining hole goes to zero while the number of internal states keeps growing . The reason this is possible is because of the negative sign of the gravitational potential in (2.1). The hole contains the initial shell of mass M as well as all the negative energy $\{c_i\}$, but the negative gravitational energy between all these excitations makes the overall mass go to zero as the black hole reaches its endpoint. This negative energy if gravity (arising from its attractive nature) is what is responsible for the information puzzle.

2.5 Summary of the information paradox

Lecture 3

3.1 What is ER=EPR?

A large fraction of the ideas in the wormhole paradigm are related to the idea of ER=EPR. Here ER stands for the Einstein-Rosen bridge in the classical black hole geometry, and EPR stands for the entanglement in the Einstein-Rosen-Podolsky setup. The phrase ER=EPR arose in email discussions between H. Verlinde and E. Verlinde. It appeared as a proposal to resolve the black hole information puzzle in the 2013 paper of Maldacena and Susskind. The notion of entanglement building bridges in spacetime had appeared earlier in a 2010 paper of van Raamsdonk. Even before this, Maldacena in 2001 had conjectured that the dual to two entangled CFTs was a geometry connecting the two boundaries in the bulk. The overall idea behind all these studies is that entanglement between two well separated entities builds a gravitational bridge between them.

These ideas sound fascinating, but they have also caused a large amount of confusion. Is there a new physics principle being postulated here? Are we saying something new about how gravity behaves, something different from what the traditional quantum gravity community would believe? Are these ideas compatible with string theory as a complete theory of quantum gravity, or do we need something different from string theory for these ideas to be true?

I was initially very excited about these ideas, and tried to explore them in the context of string theory black holes. But after a while I found that I could not make these ideas work. The problem is very basic: these ideas are not compatible with the linearity of quantum theory. I will try to explain the problem in simple terms in this lecture.

It should be noted that some people working with ER=EPR have agreed that if the idea was taken as an exact statement, then it would violate the linearity of quantum theory. They argue that in a certain *approximation*, one can express correlators between entangled states by using a geometric bridge connecting the entangled systems. While this can be true, the problem is that the quantity we want to compute – S_{ent}, the entanglement entropy – is a delicate quantity that is very sensitive to the details of the state. I have not found any approximation which incorporates the idea of ER=EPR and also computes the required entanglement.

I will present the difficulties with this circle of ideas in small steps below. People working in this area can then tell me which step is wring, and the result will hopefully lead to clarity for everyone.

3.2 The basic idea of ER=EPR

I will assume that the theory of quantum gravity is string theory. The reader can adapt the argument below to any other theory of quantum gravity and see where that leads.

But with string theory we have a precise and rigorous framework that is accepted by most string theorists, so we can arrive at sharp conflicts when an idea may not be feasible.

Consider IIA string theory for concreteness. Place a D0 brane at position x_1, and another at position x_2. We will take the separation L of these two particles to be larger than any other length scale of interest

$$L \equiv |x_2 - x_1| \to \infty \tag{3.1}$$

The D0 brane forms a 256 dimensional spin multiplet. We take any two of these spin states which can form a singlet, and call them $|\uparrow\rangle$ and $|\downarrow\rangle$. We entangle the two D0 branes as

$$|\Psi\rangle = \frac{1}{\sqrt{2}}\Big(|\uparrow\rangle_1|\downarrow\rangle_2 - |\downarrow\rangle_1|\uparrow\rangle_2\Big) \tag{3.2}$$

where the subscripts $1, 2$ stand for the D0 branes at x_1, x_2 respectively.

The idea of ER=EPR is that this entanglement between the regions x_1, x_2 leads to a geometric connection between x_1, x_2, depicted by the thin (planck diameter) wormhole drawn to link the region near x_1 to the region near x_2. But what does this link mean?

Before asking this question, let us recall how the above notion was used in [] to address the black hole puzzle. We have seen that the process of Hawking evaporation creates a large number of entangled pairs with the quanta $\{c_i\}$ inside the hole being entangled with their partners $\{b_i\}$ near infinity. We imagine the above mentioned planck width wormhole connecting the pair (b_i, c_i) for each i. In the interior of the hole, we imagine that these thin wormholes fuse together to crate a wormhole with a large diameter. The resulting wormhole structure is given by the squid diagram in fig.**??**, taken from the paper [] by Maldacena and Susskind. This wormhole is argued to imply new physics not included in the Hawking argument for the monotonically rising Page curve, and it is hoped that this new physics will bring the Page curve down.

Let us return to the entangled state (**??**) and ask what the wormhole connecting points x_1, x_2 could mean. I have received two different answers to this question:

(a) *There is no new physical effect that is being claimed here: the wormhole between x_1, x_2 is merely a pictorial depiction of the fact that the two D0 branes are entangled. Entanglement in string theory, which is a theory with gravity, is not different in any way from the entanglement between two electron spins that we learn about in our undergraduate quantum textbooks.*

If the wormhole is merely a depiction of the entanglement, then we are free to *not* use this depiction; we can just write the entanglement explicitly as in (3.2). In that case the squid diagram fig.**??** has nothing new to say about the information paradox.

(b) *The wormhole arising from the entanglement between the D0 branes does have an implication for the dynamics, something new that is not captured by the usual physics of entanglement described in undergraduate textbooks. This new effect, amplified by the large number of entanglements in the squid diagram fig.**??** tells us that Hawking missed something in his analysis of entanglement.*

Though this picture looks exciting, we see that there is an immediate conflict with the linearity of quantum mechanics:

(i) Take the D0 at x_1 spin up, the D0 at x_2 to be spin down. This gives the state $|\Psi_1\rangle = |\uparrow\rangle_1 |\downarrow\rangle_2$. There is no entanglement, so this time we expect no novel effects from the idea of ER=EPR. Let the future evolution of this state be $|\Psi_1(t)\rangle$.

(ii) Take the D0 at x_1 spin down, the D0 at x_2 to be spin up. This gives the state $|\Psi_1\rangle = |\uparrow\rangle_1 |\downarrow\rangle_2$. There is no entanglement, so this time we expect no novel effects from the idea of ER=EPR. Let the future evolution of this state be $|\Psi_1(t)\rangle$.

(iii) We can now superpose these states to make the entangled state

$$|\Psi\rangle = \frac{1}{\sqrt{2}}\Big(|\Psi_1\rangle - |\Psi_2\rangle\Big) \tag{3.3}$$

By linearity of quantum mechanics, this state will have to evolve to

$$|\Psi(t)\rangle = \frac{1}{\sqrt{2}}\Big(|\Psi_1(t)\rangle - |\Psi_2(t)\rangle\Big) \tag{3.4}$$

So how can entanglement ever give any new effects?

(Note that all the above statements are in our exact theory of gravity which we have assumed for concreteness to be string theory; we have not done any averaging or coarse-graining. We will address such averaging attempts later.)

3.3 Moving towards black holes

The above difficulty with trying to implement ER=EPR looks clear enough, but one might say the following. The argument of the above section pertained to the entanglement between two D-branes. Could it be that black holes are somewhat different, and new rules incorporating ER=EPR can be postulated when we entangle black holes instead of D-branes?

In this section we will see that it is not possible to find such a way out of the difficulty. The reason is that in string theory black holes are made from D-branes. We can describe D-brane dynamics in a simple way using the language of open strings between the branes – this gives Yang-Mills theory. But the gravitational description of the hole is obtained from this open string dynamics by just a change of variables – we write open string diagrams in terms of closed string diagrams. So there is no room for new postulates in the gravitational description of black holes. In what follows, we will this argument in steps, so that if one wants to question the argument then he must point to which step is erroneous.

3.3.1 Many D-branes

In the discussion above we had placed one D0 brane at x_1 and one D0 brane at x_2. We now wish to move towards black holes in string theory. Thus take IIB string theory, and compactify it as

$$M_{9,1} \to M_{4,1} \times^2 \times T^4 \tag{3.5}$$

At the location x_1, we wrap n_1 D1 branes on S^1, and n_5 D-branes on $S^1 \times T^4$. For concreteness we take $n_1 = 10^6$, $n_5 = 10^6$, the string coupling to be $g = 0.1$ and the compactification radii of the compact directions to be $\sim \sqrt{\alpha'}$. This brane bound state has a degeneracy $N_{states} = Exp[S_{micro}]$ with

$$S_{micro} = 2\sqrt{2}\sqrt{n_1 n_5} \tag{3.6}$$

We make a similar brane bound state at x_2. Then we consider the maximally entangled state between these brane bound states

$$|\Psi\rangle = \frac{1}{\sqrt{N_{states}}} \sum_i |\psi_i\rangle_1 |\psi_i\rangle_2 \tag{3.7}$$

where $|\psi_i\rangle_1$ are the N_{states} states of the branes at x_1 and $|\psi_i\rangle_2$ are the N_{states} states of the branes at x_2.

Since we have $gn_1n_5 \gg 1$, we are in the domain of coupling where these branes describe the '2-charge extremal hole'. The radius R of this hole is much larger than planck size. But we take the distance L between the two brane bound states to be much much larger than R, and in fact think of the limit $L \to \infty$. In this situation all the arguments of section 3.2 go through as before; i.e., there cannot be any nontrivial effect of the entanglement in (**??**).

3.3.2 Gravity description of the 2-charge hole

For the above 2-charge extremal states, we happen to know the gravitational description as well. These are the 2-charge fuzzballs, and give a gravitational description of the states $\psi_i\rangle_1, |\psi_i\rangle_2$ in (**??**). These gravitational solutions are fuzzballs – explicit solutions of gravity with no horizon – thus we can just think of then as 'string stars'. Thus in the gravitational description we again get a state just like (**??**); i.e., two stars which are separated by a large distance L and have their states entangled.

The important point is that just as the description (**??**) of the brane states did not have any nontrivial effect like ERR=EPR, the gravitational description of these states cannot have a nontrivial effect of entanglement either. The reason is that by the conjecture of AdS/CFT duality, the gravitational description is the same as the gauge theory description of the branes, but now written in different variables. The gauge theory description used in (**??**) is given in terms of open strings stretching between the branes; the interactions between these open strings is given by yang-Mills theory. But open string loops can be written as closed strings in the cross channel – this is called open-closed

duality,, and is a basic feature of string theory. The open strings in the brane bound state at x_1 had no couplings to the open strings near x_2. Thus when we change variables to closed strings, there will be no couplings between the closed strings near $x-1$ and the closed strings near x_2. In other words, there is no dynamical interaction between the gravitational degrees of freedom at $x-1$ and x_2. We just have textbook entanglement, and no novel effects from ER=EPR.

3.3.3 3-charge extremal hole

The 2-charge extremal hole is sometime called the 'small black hole', which is an object on the threshold of being a black hole. But with the same compactification (**??**) we can add a third charge P obtain the 3-charge extremal hole, which is thought of as a perfectly good black hole in string theory. For concreteness let us take n_p units of momentum charge P, which yields the entropy of the brane bound state

$$S_{micro} = 2\sqrt{n_1 n_5 n_p} \tag{3.8}$$

The entropy (**??**) is computed from the open string description of the brane bound state; i.e., from the CFT description. This CFT computation has been extended to an *exact* count of microstates; (**??**) is the leading order approximation for large charges. Recent work has shown how to understand this exact degeneracy in terms of gravitational degrees of freedom – one extends the Gibbons Hawking computation to higher and higher orders. Given this fact, it is generally agreed that the gravitational description of the 3-charge extremal hole is also given by the usual AdS/CFT map: the open string degrees of freedom of the CFT description are to be rewritten as the closed string degrees of freedom of the gravity description. Following the reasoning of section **??** above, we cannot have any nontrivial effect of ER=EPR if we entangle the states of two well separated 3-charge extremal holes.

Interestingly, given the above arguments, some people have told me that indeed there should be no Er=EPR king of connection between two entangled *extremal* holes, but there *should* be a nontrivial ER=EPR effect for *nonextremal* holes. I find this strange, because I have always believed that near-extremal holes are described by D-brane bound states just the way that extremal black holes are. For example, one can add both P and $\bar{P}$ excitations to the $D1D5$ bound state described above, and this should give the near extremal hole in terms of open strings between branes.

Thus it is useful to summarize the arguments of this section in the following way. Suppose we assume that

(A1) Black holes in string theory (both extremal and nonextremal) are made by taking bound states of branes

(A2) The gravitational description of the hole is obtained just by a change of variables from the CFT description – one rewrites the open string degrees of freedom describing the brane state in terms of closed strings.

Then there cannot be any novel effect like ER=EPR.

3.4 Where did the idea of ER=EPR come from?

The above difficulty with implementing any idea like ER=EPR looks clear enough. So let us ask: why did people suggest this as an idea that might be true?

In my understanding the idea has its origins in the attempt to find a CFT dual to the eternal black hole. In fig.**??** we depict the eternal hole in, which has two asymptotic boundaries. If the black hole is in asymptotically flat spacetime then the two external regions are Minkowski spacetimes extending to $r \to infty$, while if it is a black hole in AdS then each of the two asymptotic boundaries is a boundary of AdS.

Consider the latter case, so that we have an eternal black hole spacetime, described by some black hole mass M, in an AdS spacetime with cosmological constant $\Lambda < 0$. The regions near the boundaries are locally AdS spacetime regions. We can then ask: what is the CFT dual to this spacetime?

The general belief is that an asymptotically AdS spacetime has a dual that is a CFT living on the AdS boundary. Thus time we have *two* boundaries, so the dual to the eternal black hole spacetime is the union of two CFT, which we will call the Left (L) and Right (R) CFT. These two CFT's are not interacting, so the total Hamiltonian is a sum

$$\hat{H} = \hat{H}_L + \hat{H}_R \tag{3.9}$$

Thus the only relation between the L and R CFTs can be the fact that we entangle their states. It is then natural to consider the 'thermofield double' state

$$|\Psi\rangle = \frac{1}{Z^{\frac{1}{2}}} \sum_i e^{-\frac{\beta}{2}E_i} |\psi_i\rangle_L |\psi_i\rangle_R \tag{3.10}$$

where

$$Z = \sum_i e^{-\beta E_i} \tag{3.11}$$

is a normalization factor, β is the inverse temperature of the eternal hole, and $\psi_i\rangle_L, |\psi_i\rangle_R$ are the energy eigenstates of the L and R CFTs respectively. Thus one conjectures that the state (**??**) is the CFT dual of the eternal black hole geometry.

The eternal black hole geometry is clearly connected between its left and right. More precisely, the gravitational degrees of freedom on the left and right sides can interact with each other in the following way. A graviton in the Left wedge L_g can propagate to the future wedge F_g, and a graviton in the Right wedge R_g can also propagate to the future wedge F_g. In this future wedge these gravitons can collide. Thus in the gravitational picture there is an interaction between the degrees of freedom on the Left and Right sides:

$$\hat{H}_g \neq \hat{H}_{g,L} + \hat{H}_{g,R} \tag{3.12}$$

We can now trace the origin of the idea of ER=EPR. The CFT degrees of freedom are *entangled* between the L and R sides (eq.(**??**)) but they are not *interacting* (eq.(**??**).

The gravitational dual has a physical connection between the L and R sides, given by the Einstein-Rosen bridge (ER) connecting the two sides. Thus somehow entanglement between the L and R CFTs has led to a physical connection between the gravity duals.

In the above description, entanglement in the *boundary CFTs* led to a physical connection in the *gravitational bulk*. Van Raamsdonk took the idea a little further, by arguing that entanglement between two regions in the *gravitational* theory leads to a physical connection between the two regions of the gravitational theory. It is this kind of connection that is invoked in the squid diagram for an evaporating black hole in [] (fig.**??**). This circle of ideas came to be called 'spacetime is built from entanglement'.

But we have seen in section **??** that such ideas of ER=EPR are in conflict with the assumptions (A1), (A2) that seem to be generally accepted in string theory. Thus we will now take a more careful look at the chain of ideas leading to the notion of ER=EPR.

3.5 The problem with the eternal hole

One way or another, the idea of ER=EPR rests on the structure of the eternal hole, which has the Einstein-Rosen bridge (ER) connecting its two sides. So let us start by taking a closer look at the eternal hole in AdS and its conjectured dual CFT description.

We see something right away that should be a cause for concern. In the CFT description, there is no interaction between the L and R CFTs (eq.(**??**)). These CFTs describe the dynamics of open strings between D-branes. If we accept assumption (A2) of section **??**, then the gravitational description is just a change of variables from open strings to closed strings. This would imply that

$$\hat{H}_g = \hat{H}_{g,L} + \hat{H}_{g,R} \tag{3.13}$$

in contradiction to (**??**). How do the proponents of ER=EPR address this?

The standard answer to this seems to run as follows. The interactions between gravitons from the L_g and R_g regions happens in the forward wedge F_g. But the outgoing particles from this interaction cannot escape to either the left or right asymptotic region; the horizons prevent any information from the region F_g from coming out. So this interaction somehow an artifact of the variables being used in the description; i.e., the gravitational variables are somehow a mixed up combination of the variables of the L and R CFTs.

But this suggests that there are alternative variables in gravity where we do *not* have the ER bridge joining the L_g, R_g regions. In any case this is an issue we must dig into deeper since it this connection between L_g, R_g is the essence of the ER=EPR argument. So let us take a deeper look at the nature of the interaction between the two sides in the gravitational theory. In doing this we will encounter a problem, which was described in [].

Lecture IV

4.1 Wormholes suggested by ensemble averaged models

We have seen that one circle of ideas in the wormhole paradigm revolved around the idea of ER=EPR. A second set of ideas emerged from the SYK model and its relation to JT gravity. People have sought to extend these JT gravity computations to address the black hole information puzzle. But as we will see in this lecture, this extension seems to run into some fundamental difficulties.

4.1.1 JT gravity

JT gravity has been often used in motivating wormhole ideas. So let us begin by summarizing JT gravity.

Gravity in D spacetime dimensions has $(D-1)(D-2)/2-1$ propagating degrees of freedom. To see this, consider a graviton $h_{ij}(t-x)$ propagating in some chosen direction x. The graviton is traceless and transverse, so the indices i,j take $D-2$ values each, and the above count gives the number of components of a $D-1$ dimensional symmtric traceless matrix.

In $D=2$ this count gives -1 degrees of freedom; i.e., we have an overconstrained system. So we add one degree of freedom – a scalar ϕ, the dilaton – to get dilaton gravity described by $\{g_{ab},\phi\}$. This theory has zero propagating degrees of freedom, but it can have interesting solutions when we consider different topologies or add external sources.

Let us restrict to terms with at most two derivatives in the action (we can also consider higher derivative theories, but these will not be of direct interest to us). Then the most general action has the form

$$S=\int\sqrt{-g}\Big[A(\phi)R+B(\phi)\partial_a\phi\partial^a\phi+C(\phi)\Big] \tag{4.1}$$

we can make a field redefinition

$$\tilde{g}_{ab}=f(\phi)g_{ab} \tag{4.2}$$

to remove the kinetic term for ϕ, getting an action

$$S=\int\sqrt{-\tilde{g}}\Big[\tilde{A}(\phi)\tilde{R}+\tilde{C}(\phi)\Big] \tag{4.3}$$

We can redefine the dilaton field $A(\phi)\to\tilde{\phi}$ to remove another function

$$S=\int\sqrt{-\tilde{g}}\Big[\tilde{\phi}\tilde{R}+\tilde{\tilde{C}}(\tilde{\phi})\Big] \tag{4.4}$$

Thus the general dilaton gravity theory is specified by one arbitrary function of the dilaton. If we dimensionally reduce 3+1 gravity on the angular sphere S^2, we get 1+1 dimensional dilaton gravity in the r,t directions with $\bar{C}(\tilde{\phi})=\phi^{-\frac{1}{2}}$.

If we choose $\tilde{\tilde{C}}(\tilde{\phi}) = \Lambda\tilde{\phi}$ the we get JT gravity

$$S = \int \sqrt{-g}\phi(R + \Lambda) \tag{4.5}$$

where we have now dropped the tilde symbol from all variables for simplicity. The ϕ equation of motion gives

$$R = -\Lambda \tag{4.6}$$

so that we have a constant curvature metric in the absence of sources. The metric equation gives

$$\phi_{;ab} - g_{ab}\phi_{;c}{}^{c} + \frac{\Lambda}{2}g_{ab}\phi = 0 \tag{4.7}$$

It may appear that JT gravity corresponds a very particular choice of potential $\tilde{\tilde{C}}(\phi)$, but such a potential arises naturally if we consider small fluctuations around an equilibrium configuration. Suppose that the equilibrium solution corresponds to $\phi = \phi_0$, and we are interested in small fluctuations of ϕ around this value. Then we write

$$\tilde{\tilde{C}}(\phi) \approx \tilde{\tilde{C}}(\phi_0) + \alpha(\phi - \phi_0) \tag{4.8}$$

Dropping the constant term $\tilde{\tilde{C}}(\phi_0)$ in the action and redefining $\phi - \phi_0 \to \phi$ we get the action for JT gravity.

One case of the above type that will be of interest to us is the description of small fluctuations near the horizon of extremal black holes. In extremal hole in 3+1 dimensions haas a metric

$$ds^2 = -(1 - \frac{Q}{r})^2 dt^2 + \frac{dr^2}{(1 - \frac{Q}{r})^2} + r^2 d\Omega_2^2 \tag{4.9}$$

The metric is flat space at $r \to \infty$, then we have a 'neck' region around $r \sim Q$, and then we have an infinite length throat upto the horizon. In this throat region $r = Q + \epsilon$ the angular sphere approaches a constant radius $r_0 = Q$. The r, t part of the geometry has the form

$$ds^2 \to Q^2[-\tilde{r}^2 d\tilde{t}^2 + \frac{d\tilde{r}^2}{\tilde{r}^2}] \tag{4.10}$$

where we have written $\tilde{r} = r - Q$ and rescaled $\tilde{t}t$. This metric has constant curvature, so it is locally AdS_2. Thus the throat geometry approaches $AdS_2 \times S^2$ as we approach the horizon.

Now suppose we look at small fluctuations around this metric. For simplicity we can restrict to fluctuations where we only allow a change in the radius of the S^2 as a function of $\tilde{r}, \tilde{t}$. This perturbation of the radius acts as a scalar on the $\tilde{r}, \tilde{t}$ space, and since we are looking at small fluctuations around the equilibrium radius Q, the dynamics of these fluctuations is given by the JT gravity action (**??**).

4.2 The conflict

Instead of following the history of JT gravity in the wormhole paradigm, we will proceed in the following way. We will set up a situation with black holes (where JT gravity appears in the near horizon region). We will then ask a precise question about the possible contribution of wormholes in this setting. I have found disparate answers to this question. These differences will allow me to explain the he conflicts in the wormhole paradigm that I have not been able to resolve.

Consider two extremal black holes, with radius $R = Q$ each. The two holes are placed at a very large distance L apart. As before, we will let L be larger than any other length scale in the problem, so we should think of the limit $L \to \infty$. In particular, $L \gg Q$ and the two black holes are not in the 'same AdS envelope'; i..e, their near horizon AdS regions do not overlap.

As we saw above, the near-horizon geometry of each hole has an AdS_2 region, described by JT gravity. We can cut a hole in the AdS region of one black hole, and a similar hole in the AdS region of the other black hole. We can join these two holes by a short neck, making a wormhole that connects the near horizon region of one hole to the near horizon region of the other hole. Note that the size of the wormhole will be order $\sim Q$, so it is much smaller than the very large distance L. (Recall that we should really think of the limit $L \to \infty$, so we are asking if there is a nonlocal interaction in our theory which does not fall off with distance.)

The question now is: *In the quantum gravity path integral, are manifolds with such wormholes to be included?*

Note that the action of the wormhole is finite, and so including such wormholes in the path integral it will generally yield different answers for quantities of interest as compared to the case where do not include such wormholes. Thus it is crucial to know the answer to the above posed question. But when I asked this question to several people working on the wormhole paradigm, I received different answers. I summarize these different answers (and their problems) in the subsections below.

Before proceeding, let us note that we can have wormholes of two types:

(i) In the Lorentzian theory

(ii) In the Euclidean theory

Thus where necessary we will be explicit about which kind of theory we are dealing with.

4.2.1 (a) Such wormholes do contribute in the full gravity path integral

Some people have told me that such wormhole paths *will* contribute to the exact gravity path integral, both in the Lorentzian theory and an the Euclidean theory. They believe that gravity follows a general rule: any manifold that can be imagined will contribute to the full path integral. With such a rule, we would indeed get the wormhole described above connecting the near horizon regions of two holes.

But there is an immediate problem with this view, which we explain as follows:

(i) Consider the first extremal black hole. In string theory, we can let this be a black hole of the Strominger-Vafa type: made of, say 10^6 D1 branes, 10^6 D5 branes and 10^6 momentum excitations P. The states of the hole can therefore be described through open strings carrying the momentum P, running between the $D1$ and $D5$ branes making the bound state. The dynamics of these open strings is given by a Higgs state of the Yang-Mills theory living on the $D1-D5$ bound state. The other hole is described by a similar set of branes in this open string language. Let us call the open string degrees of freedom for the first set of branes as $\{a_i^{(1)}\}$ and those for the second of branes as $\{a_i^{(2)}\}$. The energy eigenstates of the first set of branes will be called $|\psi_k\rangle_{CFT}^{(1)}$ and energy eigenstates of the second set of branes will be called $|\psi_k\rangle_{CFT}^{(2)}$.

(ii) Now consider the idea of AdS/CFT duality. This duality is supposed to be just a change of variables: using open-closed duality we can write the above open string states in terms of closed strings. These closed string variables give the gravitational description of the black hole states. This gravitational description will in general involve not just the gravitons and other massless fields, but strings, branes, and any other nonperturbative mess that string theory possesses. Let us call the closed string degrees of freedom for the first hole as $\{b_i^{(1)}\}$ and those for the second hole as $\{b_i^{(2)}\}$. AdS/CFT duality then says that there is a one-to one and onto map of states in the open string description (CFT) and the closed string description (the 'AdS' theory; we will label these states with a subscript 'grav')

$$\begin{aligned} |\psi_k\rangle_{CFT}^{(1)} &\leftrightarrow |\psi_k\rangle_{grav}^{(1)} \\ |\psi_k\rangle_{CFT}^{(2)} &\leftrightarrow |\psi_k\rangle_{grav}^{(2)} \end{aligned} \tag{4.11}$$

(iii) Now we can address the issue of wormholes. In the CFT picture, we just have two widely separated, noninteracting field theories. Thus the total Hamiltonian acting on the degrees of freedom $\{a_i^{(1)}\}$, $\{a_i^{(2)}\}$ is a sum of two terms

$$\hat{H}_T^{CFT} = \hat{H}^{(1)}(\{a^{(1)}\}) + \hat{H}^{(2)}(\{a^{(2)}\}) \tag{4.12}$$

where each term involves only the degrees of freedom of *one* of the two brane bound states; i.e., there is no interaction term between $\{a_i^{(1)}\}$, $\{a_i^{(2)}\}$. Now change variables to the closed string (i.e. gravity) description. If we just rewrite open strings in terms of closed strings, then the change of variables proceeds separately in each system

$$\{a_i^{(1)}\} \leftrightarrow \{b_i^{(1)}\}, \qquad \{a_i^{(2)}\} \leftrightarrow \{b_i^{(2)}\} \tag{4.13}$$

Thus the Hamiltonian in the gravity description must have the form

$$\hat{H}_T^{grav} = \hat{H}^{(1)}(\{b^{(1)}\}) + \hat{H}^{(2)}(\{b^{(2)}\}) \tag{4.14}$$

In other words, there cannot be any interaction term between the two holes, in the limit of widely separated holes ($L \to \infty$). How then can we have wormhole paths connecting these two systems in the path integral? There can be all kinds of nontrivial topologies in each black hole region separately, but any wormhole between the two holes will lead to *some* interaction between the two gravitating systems, and this is not possible given what we know about the dual CFT description.

The above argument (i)-(iii) is very elementary, and the reader may wonder why we have taken pains to explain it in such detail. The reason is that the responses to this argument from the wormhole community are quite diverse and striking. I will try to list below what (I believe) different people have told me:

(1) Some people have told me that there is no exact theory of gravity – i.e., since we do not know what string theory is in all its details, we should really say that there is no exact theory of gravity. In this view the string theory that we know is some perturbative theory which cannot be completed to a full theory if quantum gravity. Thus they claim that we cannot think of gravitational degrees of freedom $\{b_i^{(1)}\}$, $\{b_i^{(2)}\}$ as giving an exact description of the theory. This argue that this situation leaves them room room to postulate wormhole terms in the gravity theory while there are none in the CFT description.

I find this belief rather unsatisfactory; I have always thought that string theory is in principle a complete theory of quantum gravity, though we may not have uncovered all its phenomena yet. Further, I find that most string theorists seem to share my belief on this view of string theory.

(2) A second, related, viewpoint that I have heard is that AdS/CFT duality is not just a change of variables from open strings to closed strings. That is, the gravity theory in AdS is not obtained by just a change of variables from the CFT description. Thus they argue that we do not have the relation (**??**), and the above arguments against wormholes need not hold.

I find this view unsatisfactory as well, and again I find that most string theorists seem to share my belief on this. Over the years we have learnt how supergravity fields map to susy protected operators in the CFT (e.g. scalars in $AdS_5 \times S^5$ map to operators $tr[X^i X^j]$ in $\mathcal{N} = 4$ SYM) , string states in the AdS bulk show up as operators of the type $tr[ZZZXZZZXZZZ\dots]$ in the pp-wave approximation, and D-branes in the AdS bulk emerge as determinants of Yang-Mills traces. All this hard work has been directed to showing that the quantum gravity theory is obtained by a change of variables from the CFT. Even black hole states, which are not well understood, are supposed to just be described by a deconfined plasma state of the SYM. Thus there are strong reasons to accept that (**??**) is true; at any rate, people denying (**??**) should make an effort to state clearly that they are postulating something that is generally accepted by the string theory community.

(3) A third view I have been given is that there are no wormhole paths in the path integral when the two holes are in a product state

$$|\psi_i\rangle^{(1)}_{grav}|\psi_j\rangle^{(2)}_{grav} \tag{4.15}$$

but there *are* wormhole paths in the gravity theory if we have an entangled state

$$\sum_{i,j} C_{ij}|\psi_i\rangle^{(1)}_{grav}|\psi_j\rangle^{(2)}_{grav} \tag{4.16}$$

But this view does not make sense to me either. First consider the CFT description. We can certainly entangle the states of two CFTs, but if the Hamiltonian had the form (**??**) with no interaction term between the two systems, the there is no interaction term when we consider an entangled state either. The proponents of wormholes with this view agree with this fact about CFTs, but they argue that gravity is a novel theory and an interaction *can* develop when we consider entangled states. However the problem with their argument is the AdS/CFT map (**??**), which implies (**??**), indicating that there cannot be any novel effects of entanglement in gravity if there are no novel effects in the CFT.

Recall we had already seen in Lecture III that if we require entanglement to generate any dynamical effects then we would violate the linearity of quantum mechanics. In essence we are again encountering the same problem with such a proposal for wormholes.

4.2.2 (b) Wormholes in the low energy EFT

A second set of people working with wormholes have given me the following view. They agree that string theory an exact theory of gravity, and that (**??**)-(**??**) are all true. They agree that this means that there are no wormholes connecting the two well separated holes in the exact gravity theory. But, they argue, one can consider a low energy effective field theory that arises from the full quantum gravity theory. This low energy theory *will* have wormholes connecting the two well separated systems. These wormholes contributions will 'correct' the EFT calculation to yield the full quantum gravity calculation of the quantity of interest (which in the present case is the entanglement entropy between the two systems).

But this view also makes no sense to me. We take low energy limits of exact to obtain EFTs in many settings, but this limit never generates nonlocal interaction terms if the starting exact theory had no such terms. As a concrete example, consider a GUTS theory as the exact theory of particle physics, and the low energy EFT to be QCD. Consider two lumps of matter, separated by a distance $L \to \infty$. In the exact GUTS theory the total Hamiltonian is a sum of terms

$$\hat{H}^{exact}_T = \hat{H}^{exact}_1 + \hat{H}^{exact}_2 \tag{4.17}$$

where the subscripts $1, 2$ denote the two lumps, and we have no interaction terms H^{exact}_{12}. The low energy limit just integrates out the heavy GUTS particles to yield QCD. But

the loops of heavy particles appear independently in the two lumps, and the effective theory is described by

$$\hat{H}_T^{EFT} = \hat{H}_1^{EFT} + \hat{H}_2^{EFT} \tag{4.18}$$

with again no interaction term H_{12}^{EFT}. So how can wormholes connect the two systems in the low energy theory when they were not present in the exact theory?

The answer I have obtained to this obvious difficulty from some wormhole people is: ' Gravity is a strange theory, and unique things can happen'. But I see no room for such a possibility. Using the map (**??**), we can rewrite the exact degrees of freedom of the gravitational theories ($\{b_i^{(1)}\}, \{b_i^{(2)}\}$) in terms of the exact degrees of freedom of the corresponding CFTs ($\{a_i^{(1)}\}, \{a_i^{(2)}\}$). Thus any low energy limit of the gravitational degrees of freedom can be mapped, under this change of variables, to a low energy limit of the CFTs. Thus if there are no novel nonlocal effects possible for the EFTs of two disconnected CFTs, then there are no novel nonlocal effects possible for the EFTs of two gravitating systems.

4.2.3 (c) Averaged theories

A third set of people have considered systems that are *averaged* in some way. They argue that it is this averaging that produces the wormhole connections between systems separated by a distance $L \to \infty$. The problem we will find with these assertions is that this averaging alters the very quantity that we are interested in – the entanglement entropy. I have not been able to extract any path through these averaging computations which shows that the Page curve will come down. The arguments typically become circular: one starts by assuming a system which is like a piece of coal, and then shows that the Page curve for this piece of coal comes down as expected. But the whole point of the information puzzle was that a black hole with horizon does *not* radiate like a piece of coal.

This issue of averaging seems to have caused great confusion in the field. One source of this confusion is that several different kinds of averaging have been considered:

(i) *Statistical averaging:* This is the averaging that we do for normal statistical systems, like a box of gas. Consider a gas in a box with $N \gg 1$ atoms. We replace a given state of the gas $|\psi\rangle$ with an average over all $|\psi_i\rangle$ which have similar macroscopic properties. This replacement is called 'coarse-graining', and does not significantly change the value of simple observables like the 2-point function $\langle\hat{\phi}(x_1)\hat{\phi}(x_2)\rangle$ in the gas. But this replacement does lose all the information about the entanglement of the state $|\psi\rangle$. For an example, consider the air in this room. It could be in a state that is pure; i.e., not entangled with any other system, or it could be in a maximally entangled state with the air in the next room. The 2-point function in both cases is the same, (to the order of accuracy that we preserve in the process of coarse-graining), but the entanglement S_{ent} is completely different.

(ii) *Ensemble averaging:* In the statistical averaging mentioned above in (i), we average over different states for a given system with some Hamiltonian $\hat{H}$. In ensemble averaging, we replace the actual Hamiltonian $\hat{H}$ with an average over similar-looking Hamiltonians $\hat{H}_i$. With this kind of averaging it is even harder to preserve the entanglement S_{ent} that we wished to compute for the original system. We will spend some time looking at ensemble averaging below, since some of the wormhole ideas trace their origins to the analysis of the SYK model, which is an ensemble averaged theory.

(iii) *Hydrodynamic approximation:* In this approximation we replace a gas of atoms by a continuous fluid, described by some hydrodynamical equations. In such an approximation we lose any trace of the original atoms in the state $|\psi\rangle$ of the gas, so we also lose the entanglement S_{ent} hat these atoms may have had with an external system.

(iv) *Nonlocality in the fundamental theory of quantum gravity:* As we have noted before, one can imagine a fundamental theory of gravity where wormholes can connect any point to any other point. Such a theory has an inherent nonlocality across arbitrarty distances. We can integrate out these wormholes to obtain an effective theory which is an ensemble averaged theory if gravity. In such an ensemble averaged theory fundamental constants like the fine structure constant α will not have definite values; instead we will have to compute amplitudes with a given value of α then then average the answer over all values of α. String theory is not believed to be an ensemble averaged theory.

Each of the above approximations (i)-(iv) involve some kind of averaging. I believe that a leading source of confusion in the wormhole paradigm is that such averaged procedures have been assumed to be the same as the procedure of taking the low energy effective field theory (EFT) limit that we encountered in section **??** above. It is crucial to understand the difference between averaged theories and the EFT limit, so we describe the EFT limit more explicitly now.

4.2.4 Defining the effective field theory (EFT)

In any quantum theory, we can look for a low energy limit which has a simpler dynamics than the full theory. This low energy limit is sometimes called an effective field theory (EFT), and serves as a good approximation to answer questions that can be posed in the low energy theory. As noted above, in the wormhole paradigm we have to avoid confusing the notion of an EFT with the notion of averaging. Thus we first describe the notion of an EFT more precisely, as follows:

(i) The exact stheory is described by a Hilbert space $\mathcal{H}_{exact}$, with states ψ_i^{exact}. These states evolve by a Hamiltonian $\hat{H}_{exact}$, on which we have observables $\hat{O}_{exact}$. Equivalently, we have a path integral given through an action S_{exact}; the space of paths can involve nontrivial topologies where appropriate.

(ii) The EFT approximation is a map to a low energy Hilbert space $\mathcal{H}_{EFT}$, with states ψ_i^{EFT}. These states evolve by a Hamiltonian $\hat{H}_{EFT}$, on which we have observables $\hat{O}_{EFT}$. Equivalently, we have a path integral given through an action S_{EFT}; the space of paths can involve nontrivial topologies where appropriate. The operators of interest $\hat{O}_{EFT}$ are, typically, a small (low energy) subset of the full set of operators $\hat{O}_{exact}$.

(iii) There is a map $\mathcal{H}_{EFT} \rightarrow \mathcal{H}_{exact}$ with

$$|\psi_i\rangle_{EFT} \rightarrow |\psi_a\rangle_{exact}, \quad \hat{O}_{EFT} \rightarrow \hat{O}_{exact} \tag{4.19}$$

such that the evolution between the exact states is well approximated by the semiclassical evolution between the approximate states

$${}_{exact}\langle\psi_j|e^{-i\hat{H}_{exact}t}|\psi_i\rangle_{exact} \approx {}_{EFT}\langle\psi_b|e^{-i\hat{H}_{EFT}t}|\psi_a\rangle_{EFT} \tag{4.20}$$

and a similar approximation holds for expectation values

$${}_{exact}\langle\psi_j|\hat{O}_{exact}|\psi_i\rangle_{exact} \approx {}_{EFT}\langle\psi_b|\hat{O}_{EFT}|\psi_a\rangle_{EFT} \tag{4.21}$$

4.2.5 Ensemble averaged theories

Let us now recall the idea of an ensemble averaged theory. Suppose we have a field theory with a coupling constant α. In a normal field theory of this type, we have a factorization of correlators when the operators in the correlator are separated by a large distance

$$\langle\hat{O}_1(x_1)\hat{O}_2(x_2)\rangle_\alpha \rightarrow \langle\hat{O}_1(x_1)\rangle_\alpha\langle\hat{O}_2(x_2)\rangle_\alpha \tag{4.22}$$

in the limit $|x_1 - x_2| \equiv L \rightarrow \infty$. We have added a subscript α to the correlators to denote the fact that the coupling is set to a value α.

Now consider an ensemble averaged version of this theory. Now we are supposed to compute the correlators as above, but then follow this by an averaging over the values of α. Thus we get the ensemble averaged correlators

$$\langle\langle\hat{O}_1(x_1)\hat{O}_2(x_2)\rangle\rangle \equiv \int d\alpha W(\alpha)\langle\hat{O}_1(x_1)\hat{O}_2(x_2)\rangle_\alpha \tag{4.23}$$

where $W(\alpha)$ is some weighting function. Such ensemble averaged correlators do not have the factorization (**??**). As a simple example, suppose we just average over two different values of α with equal weights.

$$\begin{aligned}\langle\langle\hat{O}_1(x_1)\hat{O}_2(x_2)\rangle\rangle &= \frac{1}{2}\langle\hat{O}_1(x_1)\hat{O}_2(x_2)\rangle_{\alpha_1} + \frac{1}{2}\langle\hat{O}_1(x_1)\hat{O}_2(x_2)\rangle_{\alpha_2} \\ &\rightarrow \frac{1}{2}\langle\hat{O}_1(x_1)\rangle_{\alpha_1}\langle\hat{O}_2(x_2)\rangle_{\alpha_1} + \frac{1}{2}\langle\hat{O}_1(x_1)\rangle_{\alpha_2}\langle\hat{O}_2(x_2)\rangle_{\alpha_2}\end{aligned} \tag{4.24}$$

where the second line corresponds to the limit $|x_1 - x_2| \equiv L \rightarrow \infty$. We see that, unlike (**??**), we do not get a factorized form $\langle\hat{O}_1(x_1)\rangle\langle\hat{O}_2(x_2)\rangle$ for the ensemble averaged

correlator. This lack of factorization is due to ensemble averaging: here is an extra correlation between the points x_1, x_2 because the correlator is computed with the same value of α for the region near x_1 and the region near x_2.

An example of a system with such ensemble averaging is the SYK model. We have a large number N of fermions $\psi_i(t)$ in 0+1 dimensions, with the Hamiltonian

$$\hat{H} = \sum_{i,j,k,l} J_{ijkl}\psi_1 i\psi_j\psi_k\psi_l \tag{4.25}$$

The couplings J_{ijkl} are random variables with vanishing mean, and we do an ensemble average over these couplings.

Since there is no spatial variable x in this 0+1 dimensional theory, we cannot consider a large separation limit like (**??**). But we can see the lack of factorization by taking multiple *copies* of the SYK model, and ensemble averaging the overall system. Consider the SYK on a Euclidean time circle of length β. The ensemble averaged partition function of one copy (called copy 1) is given by

$$\langle\langle Z_1\rangle\rangle = \int D[J]W[J]\Big(tr[e^{-\beta H_1[J]}\Big) \tag{4.26}$$

Now consider *two* copies of the system, denoted by subscripts $1, 2$. The two copies have no interaction between them. The partition function of this system is

$$\langle\langle Z_{12}\rangle\rangle = \int D[J]W[J]\Big(tr[e^{-\beta H_1[J]}tr[e^{-\beta H_2[J]}\Big) \tag{4.27}$$

Analogous to (**??**), we find a lack of factorization

$$\langle\langle Z_{12}\rangle\rangle \;\neq\; \langle\langle Z_1\rangle\rangle\langle\langle Z_2\rangle\rangle \tag{4.28}$$

On the other hand, if we did not have an ensemble average over couplings $[J]$, then we *would* find a factorization for the partition function of two disconnected systems $1, 2$

$$Z_{12} = Z_1 Z_2 \tag{4.29}$$

Saad, Shenker and Stanford [] found that the lack of factorization (**??**) could be elegantly expressed in a graphical way. Take the SYK model at large N, and consider its dynamics at low energies. The partition function of a single copy of the SYK model $\langle\langle Z_1\rangle\rangle$ can be written at leading order using a 2-dimensional surface filling the Euclidean time circle.: the partition function is given by the by action of JT gravity computed on the disc. In this language of surfaces the lack of factorization (**??**) was given by the action of JT gravity on surfaces that *connected* the two copies of the SYK model. This joining surface looks like a wormhole. We therefore see that the effect of ensemble averaging leads to a relation between two disconnected copies of a field theory, and that for the SYK model the effect of this connection is given by 'wormholes' in a low energy effective description that involves JT gravity on surfaces.

4.3 Summary

We can now pull together the various threads that we have discussed in this lecture; this will allow me to explain some of the confusions I have encountered with wormholes.

There are two distinct places where JT gravity has appeared in our discussion:

(i) Consider string theory, which is not an ensemble averaged theory. Consider the classical solution of an extremal black hole in this theory. The small fluctuations of the metric in the region just outside the horizon are described by JT gravity,

(ii) Consider the SYK model, which is an ensemble averaged theory field theory. The ensemble averaging leads to a nonfactorization of amplitudes even for systems which are not connected by any term in the Hamiltonian, and this lack of fatorization can be encoded through the JT gravity action of a wormhole surface connecting the noninteracting systems.

I think some part of the confusion in the wormhole paradigm has resulted from thinking of (i) and (ii) as being the same thing, while they are clearly not. The wormholes in (ii) arise from ensemble averaging, while the string theory used in (i) is not an ensemble averaged theory. With this in mind, let us review the conflicts we noted in section **??**.

(a) We saw in section **??** that if we consider wormholes between the near horizon regions of two well separated black holes in string theory, then we get a sharp conflict with the idea of AS/CFT duality. We now see that even though the small fluctuation dynamics of near horizon regions can be described by JT gravity, and that JT gravity wormholes have appeared in the ensemble averaged SYL model, we should *not* consider such wormholes as part of the fundamental description of the string theory path integral. This fact is equally true wether we are considering the Lorentzian theory or the Euclidean theory. We will encounter this issue more explictly when considering the motivations for drawing a 'replica wormhole.

(b) As we noted above in section , JT gravity emerges as an effective description in large N, low energy limit of the SYK model. This led some people to argue as follows. The exact string theory does not have wormholes connecting distant holes. But we can take a low energy limit obtaining an effective field theory (EFT) which is described by the usual gravity action

$$S_{low\,energy} = \int \sqrt{-g}R + \dots \tag{4.30}$$

In taking this low energy limit some details have been lost, and they argue that these details can be accounted for by introducing wormholes that can connect distant systems.

But in string theory the low energy limit giving (**??**) is just like the way QCD appears as the low energy limit of GUTS. This EFT limit was described in section **??**. In such a limit we do *not* generate a wormhole type connection between disconnected systems.

(Note that the wormhole term in the low energy SYK theory did not emerge from taking a low energy limit, but from the existence of ensemble averaging.)

(c) We noted in section **??** that some approaches study wormhole type terms that arise from an averaging procedure. The SYK model is an ensemble averaged theory, while string theory is not. In order to see the wormholes similar to those of the low energy SYK theory, these approaches have sought to first replace the black hole state or Hamiltonian by an ensemble of states or Hamiltonians, and then to argue for wormholes in this averaged theory.

But we noted that such averaging generally destroys the entanglement entropy S_{ent} that we are trying to find. We will look at explicit examples of such averaging in later lectures, and find that they lead to circular arguments, of the following kind. One starts by assuming that the black hole is like a piece of coal with a normal Page curve that comes down at the end of the evaporation process. One then takes an ensemble average of such pieces pf coal, and tries to argue some aspect of this average is described by wormhole type surfaces. The problem with such approaches is twofold. First because we have assumed that the black hole is like a piece of coal with a normal Page curve, there is nothing to show; we have already assumed a situation with no information paradox. Second, if there was indeed some aspect of the average that could be described by gravitational wormholes, then then aspect would apply to all statistical systems including pieces of coal; thus it would no bearing on the issue of black holes that we are studying.

Acknowledgments

This work is supported in part by DOE grant DE-SC0011726.