Position space analysis of the AdS (in)stability problem

Fotios V. Dimitrakopoulos, Ben Freivogel, Matthew Lippert and I-Sheng Yang

ITFA and GRAPPA, Universiteit van Amsterdam,
Science Park 904, 1090 GL Amsterdam, Netherlands

E-mail: f.dimitrakopoulos@uva.nl, benfreivogel@gmail.com,
matthewslippert@gmail.com, isheng.yang@gmail.com

ABSTRACT: We investigate whether arbitrarily small perturbations in global AdS space are generically unstable and collapse into black holes on the time scale set by gravitational interactions. We argue that current evidence, combined with our analysis, strongly suggests that a set of nonzero measure in the space of initial conditions does not collapse on this time scale. We perform an analysis in position space to study this puzzle, and our formalism allows us to directly study the vanishing-amplitude limit. We show that gravitational self-interaction leads to tidal deformations which are equally likely to focus or defocus energy, and we sketch the phase diagram accordingly. We also clarify the connection between gravitational evolution in global AdS and holographic thermalization.

KEYWORDS: Classical Theories of Gravity, Black Holes

Contents

1   Introduction

Recently, the gravitational (in)stability of classical gravity in asymptotically global AdS spacetime has drawn a lot of attention and also generated a lot of confusion. Given a spherically symmetric perturbation of arbitrarily small initial amplitude ϵ, two dramatically different behaviors have been observed at the timescale $\sim \epsilon^{-2}$, the earliest time on which interactions can have a significant effect [1–13]. Sometimes a black hole forms around this time; sometimes a long-lived quasi-periodic behavior emerges and gravity does not become strong. This is a great puzzle concerning both the gravitational dynamics in the bulk and the corresponding thermalization process in the holographic boundary theory.

In this paper we will focus on the bulk perspective and on the simple case of a free massless scalar field coupled to gravity. We treat the system classically and impose spherical symmetry. In the limit of small amplitude ϵ, the energy density is proportional to ϵ^2 and controls the strength of gravitational effects. Therefore, behavior at the time scale ϵ^{-2} is

sensitive to the leading-order effects of gravitational interactions. One framework to study this is to analyze the nonlinear couplings between the linearized modes induced by the gravitational interactions. Linearized modes in AdS space all have frequencies which are integer multiples of the AdS scale. A mode that is initially unexcited can be resonantly driven by the excited modes, which allows for the possibility of efficient transfer of energy. Such efficient energy transfer between modes generically leads to the breakdown of naïve perturbation theory, since the true solution does not remain close to the solution in the non-interacting theory. This resonance effect was argued to be the cause of an energy cascade — energy spreads out into more and higher modes — in order to explain black hole formation and the power-law spectrum observed during such processes [1, 2, 4, 14]. It was also argued that since the AdS spectrum is resonant, such an instability should be the generic outcome of small perturbations.

Counter-examples to the above claim in the form of the stable, quasi-periodic solutions were initially viewed as being special. It was conjectured in [4] that these stable solutions will shrink to a set of measure zero in the small ϵ limit, and the term "stability island" was used to describe their existence in the generically unstable sea of phase space. However, more recent evidence suggests that such a conclusion is too strong.[1] Numerical evidence suggests that, at finite ϵ, the stable and unstable solutions both have nonzero measure in the space of initial conditions [7, 9, 11].[2] One can then apply a simple scaling argument, described in more detail in section 2, to show that in the $\epsilon \to 0$ limit, the stable solutions persist. However, the same argument fails for unstable solutions. The open question now becomes whether there are "instability corners." Namely, in the $\epsilon \to 0$ limit, do the unstable solutions shrink to a set of measure zero, or do also continue to have finite measure?

There are some important misconceptions and misunderstandings in the current literature regarding the status of the AdS (in)stability problem, due in part to three points of confusion, which we would like to clarify here. First of all, an energy cascade is not identical to, nor does it guarantee black hole formation. This distinction has not been made clear enough. Both have been frequently used interchangeably and referred to as the "instability of AdS space." Black hole formation requires energy to be focused into a small spatial region. According to the uncertainty principle, energy flowing to high momentum is certainly a necessary condition for that, but it is not sufficient. It is entirely possible for even unboundedly high momentum modes to be populated, but for the energy distribution to stay roughly spatially homogeneous.

Therefore, in this paper, **AdS instability** strictly refers to **black hole formation** only. Because the AdS geometry changes dramatically in this case, such nomenclature aligns with a more gravity point of view.[3] This also allows us to study its implications on

---

[1]I.S. Yang thanks Jorge Santos for stressing this point during a discussion.

[2]Note that we are always discussing stability on the interaction time scale ϵ^{-2} in this paper. The question of the behavior on longer time scales is a fascinating one that touches on issues of ergodicity, Arnold diffusion, and the KAM theorem. We do not know how to attack questions on these longer time scales analytically or numerically.

[3]From the hydrodynamic point of view, the existence of an energy cascade might be a suitable definition of instability. Indeed, this is the perspective taken by some authors, and we wish the reader to see the distinction clearly.

the boundary CFT. When we refer to a solution or initial condition as stable or unstable, we will always be indicating whether it collapses to form black hole or not.

The second point of confusion is the use of term "generic." Numerical evidence suggests that, at finite ϵ, the stable and unstable solutions both form sets of nonzero measure in the space of initial conditions [7, 9, 11].[4] We are interested in the $\epsilon \to 0$ limit, and in this paper we use the following definition:

1. "Generic instability" means the set of stable initial conditions (not forming black holes) shrinks to measure zero.

2. "Generic stability" means the set of unstable initial conditions (forming black holes) shrinks to measure zero.

3. "Mixed" means that both sets have nonzero measure as $\epsilon \to 0$.

Until recently, references in the literature did not clearly distinguish between (1) and (3). For example, it was conjectured in [4] that "stability islands" shrink to a set of measure zero in the small-ϵ limit, which is certainly arguing for only (1). However, the numerical evidence in [1] showing that black holes continue to form as ϵ is reduced is consistent with both (1) and (3). Since these are three physically different cases, we think such a clear distinction is needed.

Finally, when addressing the question of instability, one needs to specify a time scale. In this paper, we will only discuss the time scale that goes to infinity as ϵ^{-2} in the $\epsilon \to 0$ limit.[5] Indeed a naïve perturbation analysis shows that something interesting can happen at this time scale. The physical question we will address is whether that "something interesting" is generically black hole formation? In the end, we will try to relate the answer to the boundary CFT: does the boundary system thermalize at this time scale?

After making all these definitions clear, in section 2 we briefly review the recent progress on this topic. We then present a very simple scaling argument which shows that possibility (1) defined above, "generic instability", is the most unlikely given by existing evidence.[6] This directly argues against the "stability island" conjecture [4]. The remaining question then is whether AdS space is generically stable (2) or mixed (3).

In section 3 we set up our perturbative method for studying gravitational self-interaction. This position-space approach is more directly relevant than the usual

---

[4]Strictly speaking, numerical results only cover discrete choices of initial conditions. So, it is therefore impossible on numerical grounds alone to prove that any such set has either zero or nonzero measure. This fact holds equally for both stable and unstable solutions. Nevertheless, if either set really were measure-zero, unless the numerical code secretly enforced extra symmetries, it would be extremely unlikely to find such a result even once in simulations. Thus, despite the numerical controversy over some of the stable solutions [15], we still interpret the current evidence that stable and unstable solutions both have nonzero measures.

[5]Behaviors at shorter time scales are somewhat trivial. For example, for a given fixed time, black hole forming solutions disappear as $\epsilon \to 0$, so the system is generically stable, case (2). The behavior at longer time scales is a very deep problem that touches on issues of ergodicity, Arnold diffusion, and the KAM theorem. We do not know how to attack those questions analytically or numerically.

[6]I.S. Yang thanks Jorge Santos for stressing this point during a discussion.
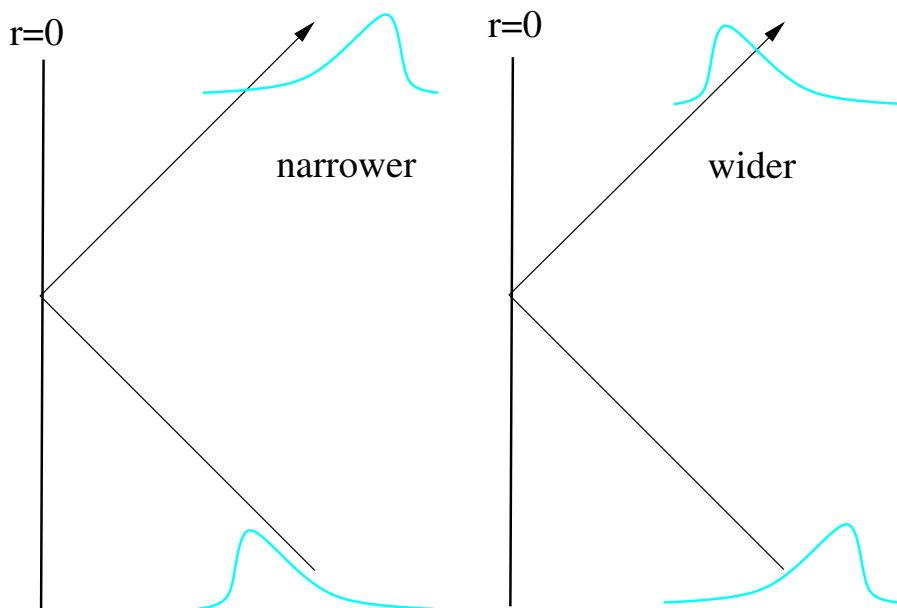
Figure 1. A thin shell that has higher energy density in its front will come out narrower after gravitational self-interactions as it bounces through $r = 0$. A shell with higher energy density in its tail will come out wider after the bounce.

momentum-space analysis to the question of whether or not black holes form.[7] If energy gets focused into a smaller region, then the solution is evolving toward a black hole. If energy is defocused into a larger region, then the solution is evolving away from a black hole. We explicitly demonstrate that in the $\epsilon \to 0$ limit, the focusing/defocusing dynamics depend only on the gravitational self-interaction near the origin of AdS, when the energy of the perturbation is maximally concentrated. The propagation through the rest of asymptotic AdS space plays no dynamical role.

In section 4 we prove a one-to-one correspondence between focusing and defocusing energy in the near-center dynamics. Heuristically, our result is shown in figure 1: a shell of massless scalar field will become narrower, its energy focused, if it is denser in the front. On the other hand, if it is denser in the tail, it will become wider and energy will defocus.[8] More generally, the leading-order dynamics of focusing and defocusing are related by time reversal, so a local maximum of energy density is also equally likely to grow or diminish within the time scale $\lesssim \epsilon^{-2}$.

In section 5 we present the new intuition our method provides and propose a conjecture on the structure of the phase space. Based on the symmetry between focusing and defocusing dynamics, the stable, quasi-periodic solutions can be understood as trajectories

---

[7]In principle, one can include all the information about relative phases in the spectral analysis to achieve the same result. Our position-space approach is simply more direct. In addition, it technically circumvents the subtlety that the gravitational interaction imposes significant additional constraints on possible resonances [16].

[8]The shell profile will change in other ways, but all changes are suppressed by ϵ^2. The focusing or defocusing behavior will last for a time scale $\lesssim \epsilon^{-2}$, so it is the dynamics we are interested in here.

that alternate between the two. As a result, they may form quasi-closed loops in phase space. In fact, some unstable solutions are also known to exhibit this alternating behavior while in the weak-gravity regime. Based on this understanding of the dynamics, we propose a conjecture on how to visualize the phase space of small perturbations in AdS space.

We also discuss how these gravitational calculations can shed light on the general concepts of thermalization in a closed system. In particular, contrary to conventional wisdom, black hole formation at the ϵ^{-2} time scale is not necessarily the holographic dual of thermalization in the boundary field theory. If the thermal gas phase is the equilibrium state, then black hole formation describes prethermalization that significantly delays true thermalization [17, 18].

In section 6, we provide a quick summary of six major points of this paper. In appendices A and B, we provide the computational details of our method and numerical examples to demonstrate how the shape of the profile determines whether its energy is focused or defocused.

2 Stability islands or instability corners?

We are interested in the perturbative stability of global AdS space. We will work in $(3+1)$ dimensions and employ the following metric for vacuum AdS_4:

$$ds^2_{AdS_4} = -\left(1 + \frac{r^2}{R^2_{\text{AdS}}}\right)dt^2 + \frac{dr^2}{1 + \frac{r^2}{R^2_{\text{AdS}}}} + r^2 d\Omega_2^2 \tag{2.1}$$

where R_{AdS} is the AdS radius and $d\Omega_2^2$ is the round metric on S^2.[9] Our perturbations will take the form of a real, massless scalar field ϕ minimally coupled to Einstein gravity with a negative cosmological constant:

$$S = \int d^4x \ \sqrt{g}\left(\frac{1}{16\pi}R + \frac{6}{R^2_{\text{AdS}}} - \frac{1}{2}\partial_\mu\phi\partial^\mu\phi\right) . \tag{2.2}$$

The Planck scale has been set to one. We will consider only spherically symmetric solutions, so both the scalar ϕ and the metric functions g_{tt} and g_{rr} will only depend on t and r.

Here we will review some existing evidence and argue that a careful interpretation strongly supports the following conclusion for spherically symmetric perturbations of a massless scalar field in AdS space:

In the $\epsilon \to 0$ limit, at the $T \sim \epsilon^{-2}$ time scale, AdS space is either generically stable, or that neither stable nor unstable perturbations are generic.

The first part of our argument is based on ample numerical evidence at small but finite ϵ. The initial conditions that lead to black hole formation (unstable) and those that lead to quasi-periodic solutions (stable) both form open sets in the phase space of nonzero measure. Note that the phase space is infinite dimensional, so no numerical evidence can prove that any set is really open. Nevertheless, whatever extrapolations are being made

---

[9]Our radial coordinate r is related to the radial coordinate x used in [1] by $r = \tan x$.

should be applied equally to both stable and unstable solutions, and the existing numerical evidence is quite sufficient to show that they are on equal footing. More specifically, numerical tests can scan a one-parameter family of initial conditions, corresponding to a line in phase space. It has been clearly demonstrated that for a few such lines, the initial conditions that lead to stable and unstable solutions both form finite segments [7, 9, 11]. We will pragmatically take this as evidence that both stable and unstable sets in phase space have nonzero measure at small but finite ϵ.

In particular, within the set of stable solutions, one can identify a subset for which "gravity never becomes strong" during the $\sim \epsilon^{-2}$ time scale; that is,

$$\exists\, \phi(\epsilon, r, t)\,,\ \text{such that}\ \left(\dot{\phi}^2 + \phi'^2\right) < \delta \ll 1 \ \text{for}\ 0 \le t \le T \sim \epsilon^{-2}. \tag{2.3}$$

Our next step is to show that in the $\epsilon \to 0$ limit, these stable solutions cannot disappear. We can use the scaling behavior observed in [1, 11], which was trustworthy to leading order in ϵ. We will demonstrate that in the $\epsilon \to 0$ limit, this scaling behavior is exact for stable, weak-gravity solutions.

The spectrum of a massless field in the AdS background is given by integer multiples of the AdS energy scale R_{AdS}^{-1}, meaning that the field profile is exactly periodic in time. Heuristically, a spherical wavefront shrinks toward the origin $r = 0$, passes through it, expands again to infinity, and finally bounces off the boundary back to the original position.[10] It is natural to describe the dynamics as a function of the "number of bounces" $N = \frac{t}{\pi R_{\text{AdS}}}$ instead of the microscopic time t:

$$\phi(r, N+1) \equiv \phi(r, t + \pi R_{\text{AdS}}) = \phi(r, N) \equiv \phi(r, t)\,. \tag{2.4}$$

Now, introducing gravitational self-interaction, as long as the field amplitude (and therefore the resulting back-reaction) is small, we have a small correction to the above exactly periodic solution,

$$\phi(r, N+1) - \phi(r, N) = A[\phi, \dot{\phi}] + \mathcal{O}(\phi^5)\,. \tag{2.5}$$

The functional A describes the small, leading-order changes to the profile, which we will analyze further in the following sections. Here we only need to know that it scales like ϕ^3. It is convenient to introduce the rescaled field, $\bar{\phi} \equiv \phi/\epsilon$, whose evolution is given by

$$\bar{\phi}(r, N+1) - \bar{\phi}(r, N) = A[\bar{\phi}, \dot{\bar{\phi}}]\epsilon^2 + \mathcal{O}(\epsilon^4)\,. \tag{2.6}$$

Although the value of N is discrete, in the $\epsilon \to 0$ limit, the change due to each bounce goes to zero. We can therefore take the continuum limit, in which eq. (2.6) becomes

$$\frac{d\bar{\phi}}{d(\epsilon^2 N)} = A\,. \tag{2.7}$$

Thus, the scaling behavior is exact:

$$\bar{\phi}_\epsilon(r, N) = \bar{\phi}_{\frac{\epsilon}{\alpha}}(r, \alpha^2 N)\,. \tag{2.8}$$

---

[10]The periodicity of geodesics in AdS is $2\pi R_{\text{AdS}}$, and in that time they pass through the origin twice. However, a shell of massless scalar field with Dirichlet boundary conditions at the boundary is actually periodic in half that time, πR_{AdS}, during which the wavefront passes through the origin only once.

Reducing the amplitude of the fluctuation simply slows down the dynamics by α^2: if ϵ is reduced by a factor of α, it takes α^2 more bounces to reach the same configuration. Therefore, if there is a stable solution at some finite ϵ and for a time $T \sim \epsilon^{-2}$ during which gravity never becomes strong, this must also be a stable solution at any smaller ϵ, all the way to the $\epsilon \to 0$ limit.[11]

Interestingly, this same argument is not applicable to unstable solutions. In order to form a black hole, the scalar field profile must first evolve to have large energy density somewhere, $\left(\dot{\phi}^2 + \phi'^2 \right) \sim 1$. In other words, gravity must become strong, at which point the higher order terms in eq. (2.6) cannot be ignored. In those cases the scaling behavior is lost. A collapsing solution at some small but finite ϵ might escape that fate if we reduce ϵ further [21].

At this point, we are left with two possibilities:

- Neither stable nor unstable perturbations are generic, since they both occupy sets of nonzero measure in the phase space.

- AdS space is perturbatively stable generically, but there are special "instability corners", which shrink to measure zero in the limit $\epsilon \to 0$.

Finally, recall that we have so far limited ourselves to spherical symmetry. Intuitively, spherical symmetry arranges for matter to converge at the origin, which is helpful for gravitational collapse. So, even if the first of the above possibilities holds within spherical symmetry, it may be that without spherical symmetry the second is instead the case.

3 Weak gravitational self-interaction in position space

3.1 The two-region approximation

We now present our approach to explicitly calculating the functional A in eq. (2.5). Our result, a precise expression for A, is given in eq. (3.21). Many of its properties will help us to better understand the dynamics and the possibility of instability corners. Our calculation will be in position space. The advantage for this approach is easily seen if we first picture the evolution of a thin shell of total energy $E \sim \epsilon^2$, thickness w and initial size r_0, such that $r_0 \gg w$. This corresponds to an initial field profile that is roughly given by

$$\phi_0(r,t)|_{t \sim t_i} \sim \frac{-\epsilon\sqrt{w}}{r} \, f\left(-\frac{r - r_0 + t - t_i}{w} \right) \ . \tag{3.1}$$

We will take the profile $f(x)$ to be some function that peaks at $x = 0$ and has compact support an order-one range around around this peak (i.e. $f(x) = 0$ for $|x| \gtrsim 1$).[12] Note that we have carefully chosen the dependence on w such that it does not affect the total mass, which is controlled solely by ϵ. The small-perturbation limit then corresponds to $\epsilon \to 0$.

---

[11] We should note that the expansion in powers of ϵ is most likely asymptotic [19], but its leading-order result has been accurate for many similar applications [20].

[12] We choose the profile to have compact support only to make the subsequent calculations somewhat cleaner. The shell only needs a narrow, well-defined width. Alternately, one could take f to have, for example, Gaussian tails without affecting the results.

Other papers studying similar scenarios choose various different initial conditions for the scalar field perturbation. Some authors take initial conditions that place the energy near $r = 0$. In other cases, the perturbation originates from a quench in the boundary CFT and appears as a wavefront coming in from $r = \infty$ [22]. Remember that in the small-ϵ limit, the leading-order behavior is the same as in empty AdS space; that is, the radiation shell simply bounces back and forth between $r = 0$ and $r = \infty$. Therefore, all of these initial positions of the shell are related by a shift in time on the order of R_{AdS}. Since we are interested in the outcome at longer time scales, they are all equivalent for our purposes.

One advantage of our position-space approach is that we can choose an r_0 which implements the following "two-region" approximation:

- For $r < r_0$, we will ignore that the background is AdS space and consider only the back-reaction of the scalar field on Minkowski space.

- For $r > r_0$, we will ignore the scalar field back-reaction and treat the geometry as empty AdS space.

In order justify this simplification, we first recall the general form of the Schwarzschild-AdS metric:

$$ds^2_{SAdS_4} = -\left(1 - \frac{2M(r)}{r} + \frac{r^2}{R^2_{\mathrm{AdS}}}\right)dt^2 + \frac{dr^2}{1 - \frac{2M(r)}{r} + \frac{r^2}{R^2_{\mathrm{AdS}}}} + r^2 d\Omega^2_2 \qquad (3.2)$$

where $M(r)$ is the total mass located inside the sphere of radius r.

For $r < r_0$ we will ignore the r^2/R^2_{AdS} terms in g_{tt} and g_{rr} responsible for the AdS asymptotics and calculate $M(r)$ due to the back-reaction of the radiation shell. This effect is strongest when the shell is near the origin and its energy is concentrated in a small region within $r < w$. We find that $M(r) \sim \epsilon^2$.

At $r = r_0$, we will start including the AdS terms and ignoring the back-reaction terms, such that for $r > r_0$ the metric is just that of empty AdS space. Naïvely, this is allowed if the metric at r_0 is approximately that of Minkowski space; that is, the corrections due to both AdS and back-reaction must be small:

$$\frac{r_0^2}{R^2_{\mathrm{AdS}}} \ll 1 \quad \text{and} \quad \frac{\epsilon^2}{r_0} \ll 1 \,. \qquad (3.3)$$

However, we should really ask for a stronger condition.

Our perturbative back-reaction calculation will be organized as an expansion in powers of ϵ^2/w, and we will work up to some power n using the Minkowski background. In order to be able to trust our results up to that order, we cannot allow the transition at r_0 to have a competing effect, meaning

$$\frac{r_0^2}{R^2_{\mathrm{AdS}}} \ll \left(\frac{\epsilon^2}{w}\right)^n \quad \text{and} \quad \frac{\epsilon^2}{r_0} \ll \left(\frac{\epsilon^2}{w}\right)^n \,. \qquad (3.4)$$

For any R_{AdS}, we can choose the shell small enough and thin enough to accommodate the hierarchy of scales

$$R_{\mathrm{AdS}} \gg r_0 \gg w \gg \epsilon^2 \,, \qquad (3.5)$$

which satisfies eq. (3.4) for any choice of n.

The two-region approximation provides a very simple picture. In the $\epsilon \to 0$ limit, the dynamical evolution is totally controlled by the central Minkowski region. For the AdS instability problem, the only meaningful calculation is the gravitational self-interaction of a thin-shell when it passes through $r < r_0$. The propagation in the $r > r_0$ region is just propagation in an empty AdS space; the shell simply travels out, reflects off the boundary, and repeats the gravitational evolution near the origin. Since the profile is modified by a small fraction $\sim (\epsilon^2/w)$ during each bounce, we expect on the time scale $\sim \epsilon^{-2}$ an order-one change to accumulate.

For example, the self-interaction might make the shell thinner after each bounce, meaning that the gravitational effect becomes stronger, since more energy is squeezed into a smaller region. If that behavior persists, then eventually the energy will be compressed during a bounce into a region near the origin smaller than its Schwarzschild radius. At this point, the weak-gravity approximation will break down, and it is very likely that in the ϵ^{-2} time scale, the shell will evolve into a black hole. On the other hand, it is also possible that the shell becomes wider after each bounce, and energy is dispersed into a larger region. In this case, there is no particular reason why gravitational effects would necessarily become strong and no indication that a black hole would form in the ϵ^{-2} time scale. The main goal here is to set up a calculation that can capture these two different behaviors.

Before moving on, we need to address the applicability of the thin-shell approximation. A full dynamical picture should accommodate energy distributions of all thicknesses. However, when $w \sim R_{\mathrm{AdS}}$, there is no clean way to separate the self-interaction from the effects of the AdS space. Nevertheless, our main interest is the instability in AdS toward black hole formation. In the small-ϵ limit, the energy must become concentrated into thin shells to even have a chance of eventually forming a black hole. Note that not all of the energy needs to be in one thin shell. But, the evolution toward a black hole is determined by the shell with the highest radial energy density, which is dominated by its self-interaction, so we can ignore the influence of other energy distribution outside the shell.

3.2 Near-Minkowski expansion

According to our approximation scheme, we can adopt the weak-gravity expansion in Minkowski space [20]:

$$\phi = \epsilon\phi_0 + \epsilon^3\phi_1 + \dots \tag{3.6}$$

$$g_{\mu\nu} = g_{\mu\nu}^0 + \epsilon^2 g_{\mu\nu}^1 + \dots \tag{3.7}$$

At zeroth order in ϵ, the background is empty Minkowski space,

$$\sim \mathcal{O}(\epsilon^0), \quad ds^2 = -dt^2 + dr^2 + r^2 d\Omega^2, \tag{3.8}$$

into which we put the initial shell profile. To first order, the equation of motion for ϕ is just that of a free field,

$$\sim \mathcal{O}(\epsilon), \quad \ddot{\phi}_0 - \phi_0'' - \frac{2}{r}\phi_0' = 0. \tag{3.9}$$

At the next order, gravity responds to the stress-energy tensor of the first-order profile. We therefore must solve the Einstein equation $G_{\mu\nu} = 8\pi T_{\mu\nu}$ to leading order in small

perturbations around empty Minkowski space. Spherical symmetry excludes dynamical degrees of freedom in the metric, so we only need to solve constraint equations. The tt and rr components suffice to provide the full answer, and the solution is parametrized by two intuitive quantities: enclosed mass M and gravitational potential V.

$$ds^2 = -[1 + 2\epsilon^2 V(r,t)]dt^2 + \left[1 + \frac{2\epsilon^2 M(r,t)}{r}\right] dr^2 + r^2 d\Omega^2 \ . \tag{3.10}$$

Note that we have also explicitly extracted the ϵ scaling from M and V, which are given in terms of the leading-order fields:

$$\sim \mathcal{O}(\epsilon^2)\,, \qquad \frac{2M'}{r^2} = 8\pi \frac{\dot{\phi}_0^2 + \phi_0'^2}{2}\,, \tag{3.11}$$

$$\frac{2}{r}\left(-\frac{M}{r^2} + V'\right) = 8\pi \frac{\dot{\phi}_0^2 + \phi_0'^2}{2}\,, \tag{3.12}$$

with boundary conditions $M(0,t) = 0$ and $V(\infty,t) = 0$. Finally, the leading nontrivial dynamics comes at the next order — the change in geometry back-reacts on the field profile.

$$\sim \mathcal{O}(\epsilon^3)\,, \quad \ddot{\phi}_1 - \phi_1'' - \frac{2}{r}\phi_1' = C\left(\ddot{\phi}_0 + \phi_0'' + \frac{2}{r}\phi_0'\right) + \dot{C}\dot{\phi}_0 + C'\phi_0' \ . \tag{3.13}$$

Here we have abbreviated $C = (V - M/r)$. We see that the field at this order obeys the same wave equation as in the previous order with the addition of a nontrivial source term.

The radial wave equation can be rewritten as a $(1+1)$-dimensional wave equation by introducing $u = r\phi$:

$$r\left(\ddot{\phi} - \phi'' - \frac{2}{r}\phi'\right) = \ddot{u} - u'' \ . \tag{3.14}$$

This implies that the initial shell profile given in eq. (3.1) is really just the left-moving part of an exact, leading-order solution,

$$r\phi_0(r,t) = u_0(r,t) = \sqrt{w}\left[f\left(\frac{r-t}{w}\right) - f\left(\frac{-r-t}{w}\right)\right] \ . \tag{3.15}$$

We remind the reader that in eq. (3.6), the ϵ dependence has been extracted explicitly for ϕ_0, therefore also for u_0. We have taken the liberty to choose the initial time $t_i = -r_0$ to simplify this expression. This allows us to start this calculation once the shell enters the $r < r_0$ region, and the center of the shell reflects off the origin at $t = 0$. Later, we will be interested in corrections to the profile at $t_f = r_0$, when the shell is leaving the central Minkowski region.

Rewriting the system in terms of the (1+1)-dimensional function u is essentially employing a method of images; we extend the range of r into the unphysical $r < 0$ region. To implement boundary conditions at $r = 0$ such that all physical quantities are finite and smooth, we require $u(r,t)$ to be antisymmetric. Similarly, we can extend the definition of M to negative r,

$$M(r,t) = \int_0^r d\tilde{r} \, \frac{\dot{\phi}_0^2 + \phi_0'^2}{2} 4\pi \tilde{r}^2 \ , \tag{3.16}$$

which is naturally an odd function of r. The same extrapolation shows that V is an even function of r.

In terms of these new variables, the problem of a shrinking shell has been mapped to the problem of two wavepackets colliding at $r = t = 0$. Note that this picture is more realistic than it seems; antipodal points of the shell do indeed collide with each other. When the shell is far from the origin, even the leading-order radial energy density is approximately equal to the naive definition of energy in this $(1 + 1)$-dimensional simplification:[13]

$$\rho_0 = 4\pi r^2 \frac{\dot{\phi}_0^2 + \phi_0'^2}{2} \approx 4\pi \frac{\dot{u}_0^2 + u_0'^2}{2} . \tag{3.17}$$

To leading order, the colliding shells simply pass through each other. Our goal is to solve the next-order nontrivial effect of such a collision by solving eq. (3.13), which in terms of u is simply

$$\ddot{u}_1 - u_1'' = C \left(\ddot{u}_0 + u_0'' \right) + \dot{C}\dot{u}_0 + C' \left(u_0' - \frac{u_0}{r} \right) \equiv S(r, t) . \tag{3.18}$$

This description has a striking resemblance to soliton collisions [23, 24]. The key to this type of problem is that, before solving the equations, we should already anticipate the physical meaning of the answer. At $t_f = r_0$, after the collision, the leading-order solution implies that an out-going shell of the opposite sign reaches exactly $r = r_0$. On top of that, we can organize the next-order correction into the following form:

$$u_0 + \epsilon^2 u_1 = u_0 - \epsilon^2 \left(\frac{\partial u_0}{\partial r} \Delta r + \frac{\partial u_0}{\partial w} \Delta w + \dots \right) . \tag{3.19}$$

We have again extracted the ϵ dependence explicitly. The shell is actually shifted by $\epsilon^2 \Delta r$ from its expected position, its width has changed by $\epsilon^2 \Delta w$, and there will be other changes orthogonal to these.

The function u_1 at $t_f = r_0$ can be solved from eq. (3.18) by integrating the retarded Green's function:

$$u_1(r, r_0) = \frac{1}{2} \int_{-r_0}^{r_0} dt \int_{r-r_0+t}^{r+r_0-t} dr' \ S(r', t) . \tag{3.20}$$

Note that the lower limit of r' can be negative, which is allowed due to our method of images. The result, however, is the same if we replace the lower bound of the integration range by its absolute value.

Note that this u_1 is only the difference between the incoming shell at $t = -r_0$ and the out-going shell at $t = r_0$, both at position $r = r_0$. Nevertheless, as we have argued that the propagation further to $r = \infty$, the reflection, and the propagation back to $r = r_0$, can all be taken as trivial. This allows us to directly relate u_1 to the functional A from eq. (2.5), which gives the leading-order change due to one bounce.

$$A \left[\frac{u_0}{r}, \frac{\dot{u}_0}{r} \right] = -\frac{u_1(\tilde{r}, t)}{\tilde{r}} , \tag{3.21}$$

---

[13]Note that the total energy $E \equiv M(\infty, t) = \int_0^\infty \rho(r) dr$ is in fact equal to the naive $(1 + 1)$-dimensional energy $\int_{-\infty}^\infty 2\pi \frac{\dot{u}_0^2 + u_0'^2}{2} \ dr$.

where $\tilde{r} = 2r_0 - r$ is the spatial reflection of r around r_0. The extra minus sign and changing to this "flipped" position are due to the trivial propagation to and from $r = \infty$.

The full procedure to calculate u_1 and extract physical information like Δr and Δw are tedious but straightforward. We will present the analytical and numerical details in appendices A and B. Here we highlight two relevant features of the results:

1. u_1 has a $\sim \log r_0$ contribution, which comes entirely from the position shift,

$$\Delta r = -\frac{\int u_1 \partial_r u_0 \ dr}{\int (\partial_r u_0)^2 \ dr} \,, \tag{3.22}$$

which has a clear physical meaning. The leading-order profile u_0 follows the $t = |r|$ trajectory, but the next-order correction to the metric modifies the null geodesics. The shell will therefore return to $r = r_0$ not exactly when $t = r_0$. However, this shift is irrelevant to the pertinent question of whether energy gets focused.[14]

2. The change in the shell's width is given by

$$\Delta w = \frac{\int u_1 \partial_w u_0 \ dr}{\int (\partial_w u_0)^2 \ dr} \,. \tag{3.23}$$

Since we have already scaled out the ϵ dependence, Δw only depends on the shape of the shell (i.e. the function f one chooses in eq. (3.15)), and it is independent of both ϵ and w.

In particular, our main result is that Δw is just as likely to be positive as negative. Specifically, when we flip the profile of the incoming shell, $f(x) \to f(-x)$, then $\Delta w \to -\Delta w$. As a special case, a symmetric profile with $f(x) = f(-x)$ will result in no first order Δw during one bounce. This demonstrates that the gravitational self-interaction in AdS is not biased toward focusing energy, and the collapse of small perturbations into black holes is probably not the generic behavior, at least not on time scales $\lesssim \epsilon^{-2}$.

As a complementary calculation, we also investigate how the maximum radial energy density ρ_{Max} of the shell behaves under the same $f(x) \to f(-x)$ transformation. Like the width w, we find that if for a given profile ρ_{Max} increases with each bounce, then for the flipped profile it decreases. This provides another indication that the weak-gravity dynamics are biased neither toward nor against focusing energy.

In the next section, we will give general proofs of these statements. We will also present numerical examples in appendix B.

4 Focusing and defocusing

4.1 Antisymmetry of the field correction

As a preliminary step in proving the statements of the previous section, we need to determine how the first-order correction u_1 responds to a spatial flip of the initial profile u_0

---

[14]This position shift is related to a shift in frequency in the momentum space analysis observed in other papers [11].

such that $f(x) \to \tilde{f}(x) = f(-x)$. We find that

$$u_1(r, r_0) \to \tilde{u}_1(r, r_0) \simeq -u_1(\tilde{r}, r_0), \qquad (4.1)$$

where $\tilde{r} = 2r_0 - r$ is again the spatial reflection of r around r_0. Note that this is an approximate statement; for a shell of width w, the error in eq. (4.1) is of order w^2/r_0. As we argued in section 3, in the $\epsilon \to 0$ limit, we can choose r_0 to make this error arbitrarily small.

The quantities that enter the expression (A.1) for u_1 are u_0 and its derivatives and C and its derivatives. So, let us first see how these quantities transform under the flip. From eq. (3.15), we can see that:

$$u_0(r, t) \to \tilde{u}_0(r, t) = -u_0(r, -t). \qquad (4.2)$$

Then, simply by differentiating the two sides of the equation (either with respect to r or t), we obtain the same transformation behavior for the derivatives of u_0. Now, to see how C transforms, all we need is to determine the transformation of M, defined in eq. (3.16):

$$\begin{aligned}
M(r, t) \to \tilde{M}(r, t) &= 2\pi \int_0^r dr' \left(\dot{\tilde{u}}_0^2(r', t) + \tilde{u}_0'^2(r', t) + \left(\frac{\tilde{u}_0(r', t)}{r'} \right)^2 - 2\frac{\tilde{u}_0(r', t)\tilde{u}_0'(r', t)}{r} \right) \\
&= 2\pi \int_0^r dr' \left(\dot{u}_0^2(r, -t) + u_0'^2(r, -t) + \left(\frac{u_0(r, -t)}{r'} \right)^2 - 2\frac{u_0(r', -t)u_0'(r, -t)}{r'} \right) \\
&= M(r, -t).
\end{aligned} \qquad (4.3)$$

Since V has the same behavior as M, then $C(r, t) = V - \frac{M}{r}$ transforms under the flip as:

$$C(r, t) \to \tilde{C}(r, t) = C(r, -t). \qquad (4.4)$$

Again, a similar relation holds for the derivatives of C. Combining the above results, we see that the source term $S(r, t)$, defined in eq. (3.18), behaves as

$$S(r, t) \to \tilde{S}(r, t) = -S(r, -t) \qquad (4.5)$$

under flipping of the initial profile. Also, by demanding regularity at the origin $r = 0$, the initial profile is antisymmetric in r, which in turn implies that $M(r, t)$ is also antisymmetric in r; hence $C(r, t)$ is symmetric. These properties imply the antisymmetry of $S(r, t)$ in its first argument, $S(r, t) = -S(-r, t)$.

Now we are ready to prove eq. (4.1), starting from the integral expression eq. (A.1) for u_1. The integration regions are illustrated in figure 2.

We first make an approximation to eq. (A.1). The upper limit of the r' integral is $r + r_0 - t$. Instead, we will extend the region of integration up to $r' = \infty$. Because the wavepacket has compact support only over a region of width w, the error introduced by this approximation comes just from the yellow shaded triangle in figure 2. The area of this added triangle is $\mathcal{O}(w^2)$ and, since $C(r, t) \sim \frac{1}{r}$, the integrand is of order $\frac{1}{r_0}$. Hence, the error is suppressed by a factor of $\frac{w^2}{r_0}$.
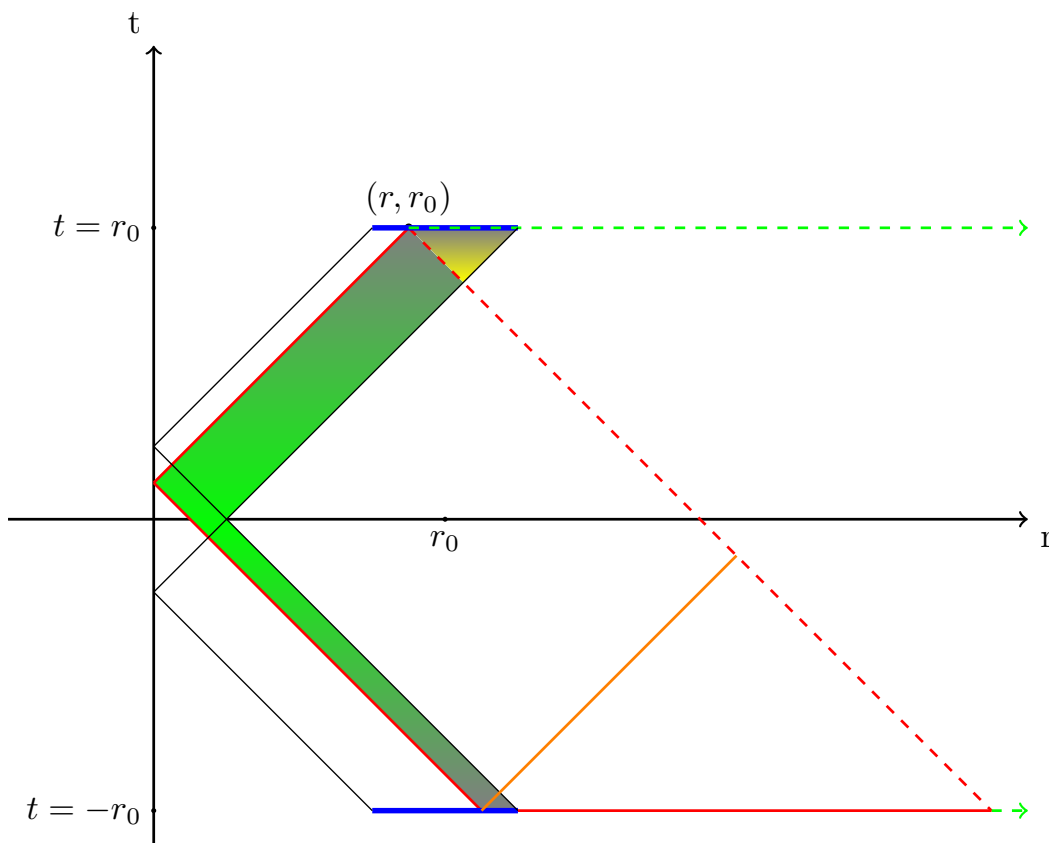
Figure 2. The areas of integration for the computation of $u_1(r, r_0)$. The blue solid line represents the source at times $t = \pm r_0$. The red lines (solid and dashed) indicate the actual area of integration in eq. (A.1) (i.e. the integral $\int_{-r_0}^{r_0} dt \int_{|r-r_0+t|}^{r+r_0-t} d\tilde{r}$), and the green shaded region indicates where the integrand is nonzero. The solid red lines together with the dashed green lines correspond to the region of integration $\int_{-r_0}^{r_0} dt \int_{|r-r_0+t|}^{\infty} d\tilde{r}$ used in our approximate eq. (4.6). The yellow triangle at the top shows the extra nonzero contribution included in the second integral, which is suppressed by $\frac{w^2}{r_0}$. Alternatively, the integral can be approximated by using the orange line instead of the horizontal red line.

A similar, and perhaps even more physical, approximation, albeit with more cumbersome limits of integration, can be made by considering the area of integration denoted by the red lines together with the orange line in figure 2. In that case, instead of adding the extra contribution from the yellow triangle at the top, we would subtract the area of the green triangle at the bottom. However, the results would be the same.

After this approximation, we have:

$$u_1(r, r_0) \simeq \frac{1}{2} \int_{-r_0}^{r_0} dt \int_{|r-r_0+t|}^{\infty} dr' \, S(r', t) \, . \tag{4.6}$$

Now, flipping the initial profile we get:

$$\tilde{u}_1(r, r_0) \simeq \frac{1}{2} \int_{-r_0}^{r_0} dt \int_{|r-r_0+t|}^{\infty} dr' \, \tilde{S}(r', t) \, . \tag{4.7}$$

Using the flipping property of $S(r,t)$, as discussed above, we can write:

$$\tilde{u}_1(r, r_0) \simeq -\frac{1}{2} \int_{-r_0}^{r_0} dt \int_{|r-r_0+t|}^{\infty} dr' \, S(r', -t) \, . \tag{4.8}$$

We can now change the dummy integration variable t to $-t$ and use the relation $r = 2r_0 - \tilde{r}$ to rewrite the lower integration limit, giving

$$\tilde{u}_1(r, r_0) \simeq -\frac{1}{2} \int_{-r_0}^{r_0} dt \int_{|(2r_0-\tilde{r})-r_0+t|}^{\infty} dr' \, S(r', t) \, . \tag{4.9}$$

Comparing this expression to eq. (4.6), we obtain

$$\tilde{u}_1(r, r_0) = -u_1(\tilde{r}, r_0) \, . \tag{4.10}$$

which is our desired result.

4.2 Shell width

In this subsection we will prove eq. (3.23); that is, under a spatial flip $f(x) \to \tilde{f}(x) = f(-x)$, the leading-order correction to the width is antisymmetric:

$$\Delta w \to \Delta \tilde{w} = -\Delta w \tag{4.11}$$

We assume the profile u_0 has compact support within $r_0 - w/2 < r < r_0 + w/2$, and evaluate Δw at late time $t_f = r_0$, well after the collision, at which point the left-moving and the right-moving wavepackets are far away from $r = 0$ and do not interfere with each other. In that case, when computing Δw from eq. (3.23), we can just integrate over the right-moving wavepacket; integrating over both wavepackets would just double both the the numerator and denominator in eq. (3.23), yielding the same result. The expression for the change in width of the flipped profile is then

$$\Delta \tilde{w} = \frac{\int_{r_0-w/2}^{r_0+w/2} dr \, \tilde{u}_1(r, r_0) \partial_w \tilde{u}_0(r, r_0)}{\int_{r_0-w/2}^{r_0+w/2} dr \, (\partial_w \tilde{u}_0)^2} \, . \tag{4.12}$$

It is convenient to define $y = r - r_0$, such that the spatial flip of the initial profile is given by

$$u_0(r_0 + y, r_0) \to \tilde{u}_0(r_0 + y, r_0) = u_0(r_0 - y, r_0) \, . \tag{4.13}$$

Starting with the numerator of eq. (4.12) and using the properties of u_0 and u_1 under the flip, we can write

$$\int_{r_0-w/2}^{r_0+w/2} dr \, \tilde{u}_1(r, r_0) \partial_w \tilde{u}_0(r, r_0) = \int_{-w/2}^{w/2} dy \, \tilde{u}_1(r_0 + y, r_0) \partial_w \tilde{u}_0(r_0 + y, r_0)$$

$$= -\int_{-w/2}^{w/2} dy \, u_1(r_0 - y, r_0) \partial_w u_0(r_0 - y, r_0)$$

$$= -\int_{-w/2}^{w/2} dy \, u_1(r_0 + y, r_0) \partial_w u_0(r_0 + y, r_0) \, . \tag{4.14}$$

In the third line, we changed the dummy integration variable from y to $-y$. The flip therefore changes the sign of the numerator. Following these same steps with the denominator of eq. (4.12), we can see that it is invariant under the flip. Putting these two statements together yields the desired result, $\Delta \tilde{w} = -\Delta w$.

4.3 Energy density

A similar argument holds for the leading-order change in the energy density $\Delta \rho$ at time $t_f = r_0$ due one bounce through the origin. Specifically, for $f(x) \to \tilde{f}(x) = f(-x)$, we find

$$\Delta \tilde{\rho}(r, r_0) \simeq -\Delta \rho(\tilde{r}, r_0), \tag{4.15}$$

where recall $\tilde{r} = 2r_0 - r$. The full radial energy density far from the origin is approximately the $(1+1)$-dimensional expression, as in eq. (3.17). Expanding it to the next order, we find

$$\rho_0 + \epsilon^2 \Delta \rho = 4\pi \frac{\dot{u}_0^2 + u_0'^2}{2} + 4\pi \epsilon^2 \left(\dot{u}_0 \dot{u}_1 + u_0' u_1' \right) . \tag{4.16}$$

We kept our principle of always extracting ϵ explicitly. The first term is the initial energy density given in eq. (3.17), which is actually ϵ^{-2} times the actual physical energy density. The second term is the leading change due to a single bounce.

The formula (4.1) we found for the behavior of u_1 under the flip holds at the specific time $t = r_0$, and it is not straightforward to see that the same relation holds for $\dot{u}_1$. An alternative way to proceed is to include the explicit expression for the derivatives of u_1 at $t_f = r_0$:

$$u_1'(r, r_0) = \frac{1}{2} \int_{-r_0}^{r_0} dt \; [S(r + r_0 - t) - S(r - r_0 + t)] \tag{4.17}$$

$$\dot{u}_1(r, r_0) = \frac{1}{2} \int_{-r_0}^{r_0} dt \, [S(r + r_0 - t) + S(r - r_0 + t)] . \tag{4.18}$$

Using these formulae and the expression for u_0 eq. (3.15), and omitting irrelevant constants, we can write down the explicit expression for $\Delta \rho$.

$$\Delta \rho(r, r_0) = \int_{-r_0}^{r_0} dt \; \left[f'(-r - r_0)S(r + r_0 - t, t) - f'(r - r_0)S(r - r_0 + t, t) \right] \tag{4.19}$$

Since the function $f(x)$ has compact support of width w around $x = 0$ and we consider values of r on the order of r_0, then the first term in the integrand vanishes. We can now determine how $\Delta \rho$ behaves under $f(x) \to f(-x)$:

$$\Delta \tilde{\rho}(r, r_0) = - \int_{-r_0}^{r_0} dt f'(-r + r_0)S(r - r_0 + t, -t) . \tag{4.20}$$

Changing the dummy integration variable t to $-t$ and substituting with $r = 2r_0 - \tilde{r}$, we obtain

$$\Delta \tilde{\rho}(r, r_0) = - \int_{-r_0}^{r_0} dt f'(\tilde{r} - r_0)S(-\tilde{r} + r_0 - t, t).$$

From the antisymmetry of S in its first argument, we get

$$\Delta\tilde{\rho}(r, r_0) = +\int_{-r_0}^{r_0} dt f'(\tilde{r} - r_0) S(\tilde{r} - r_0 + t, t)$$
$$= -\Delta\rho(\tilde{r}, t), \tag{4.21}$$

which is indeed eq. (4.15).

Eq. (4.15) relates the change in energy density at an arbitrary point r and its image $\tilde{r}$ under the flip. However, we are particularly interested in how the change in the maximum energy density is affected by the flip.

The energy density at the position of the maximum, r_{Max}, after one bounce, can be expanded as

$$\rho_{\text{Max}} \equiv \rho(r_{\text{Max}}) = \rho^{(0)}(r_{\text{Max}}) + \epsilon^2 \Delta\rho(r_{\text{Max}}). \tag{4.22}$$

It might be tempting to directly identify this with the change of maximum energy density. However, we should remember that in addition, the location of the maximum r_{Max} is also, in general, affected by the bounce, receiving corrections at the same order, $r_{\text{Max}} = r_{\text{Max}}^{(0)} + \epsilon^2 \Delta r_{\text{Max}}$. So, expanding to order ϵ^2, we find

$$\rho_{\text{Max}}^{(0)} + \epsilon^2 \Delta\rho_{\text{Max}} = \rho^{(0)}\left(r_{\text{Max}}^{(0)}\right) + \epsilon^2 \rho^{(0)'}\left(r_{\text{Max}}^{(0)}\right)\Delta r_{\text{Max}} + \epsilon^2 \Delta\rho\left(r_{\text{Max}}^{(0)}\right). \tag{4.23}$$

However, r_{Max} is an extremum of $\rho^{(0)}$, and so $\rho^{(0)'}\left(r_{\text{Max}}^{(0)}\right) = 0$. To leading-order, the change in the maximum is due only to a change in ρ and not a shift in the location of the maximum:

$$\Delta\rho_{\text{Max}} = \Delta\rho\left(r_{\text{Max}}^{(0)}\right). \tag{4.24}$$

Now, under a flip, $r_{\text{Max}}^{(0)}$ is mapped to $\tilde{r}_{\text{Max}}^{(0)} = 2r_0 - r_{\text{Max}}^{(0)}$, the location of the maximum of $\tilde{\rho}$; that is,

$$\Delta\tilde{\rho}_{\text{Max}} = \Delta\tilde{\rho}\left(\tilde{r}_{\text{Max}}^{(0)}\right). \tag{4.25}$$

From eq. (4.15), we can see that

$$\Delta\tilde{\rho}_{\text{Max}} = -\Delta\rho\left(r_{\text{Max}}^{(0)}\right) = -\Delta\rho_{\text{Max}}. \tag{4.26}$$

Therefore, flipping the profile reverses the direction of the change in the maximum energy density. This result nicely complements our result regarding the the change in width eq. (4.11). Both of these results indicate that there is no bias in the weak-gravity dynamics toward either increasing or decreasing the energy concentration.

5 Discussion

5.1 Phase space diagram

In section 3, we provided the recipe to compute the change in a field profile after one bounce, and it contained all the information about the functional A in eq. (2.5). In principle, one can add the resulting $\bar{\phi}_1$ to the original $\bar{\phi}_0$ to make a new initial condition, and calculate the result of the next bounce. Choosing a small ϵ and reiterating this process $\sim \epsilon^{-2}$ times is equivalent to solving eq. (2.7). In principal, this will directly reproduce the long term evolution. Unfortunately, there is one technical difficulty that we have not been able to overcome.
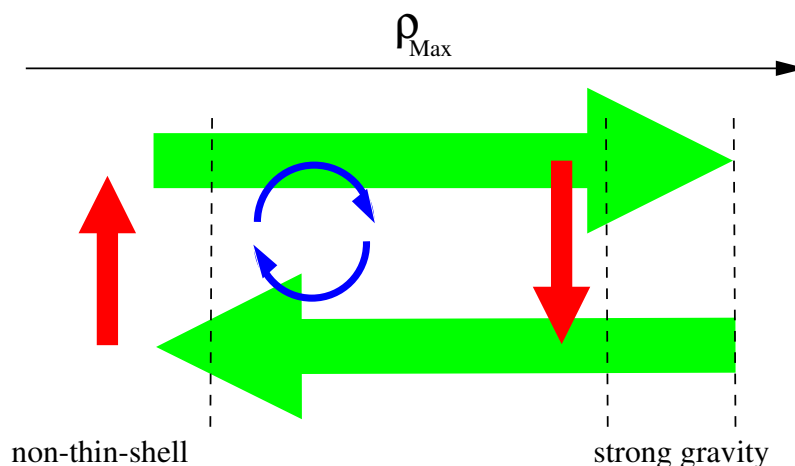
Figure 3. A two-dimensional projection within a constant-energy slice of the phase space. The horizontal axis is the peak energy density, and the big, green arrows toward left and right represent the focusing and defocusing flow due to gravitational self-interaction. Together with the upward and downward flows represented by the small, red arrows, the phase space has a circular flow pattern. The blue loop represents quasi-periodic solutions that stay within the center of this circular flow.

Our method has one disadvantage: energy conservation is by definition an approximation. We basically "turn on" a self-gravitational potential when a shell shrinks below r_0, let energy flow between it and the field kinetic terms, then turn it off when the shell expands over r_0. The amount of potential energy we turn on and off differs by $\sim \epsilon^4 w/r_0^2$. Although this is suppressed by an extra factor of w/r_0 from the quantities we care about in section 4 and so it does not invalidate our results, it is technically difficult to control. We could naïvely go to a larger r_0 for better energy conservation, which would increase the integration range required to solve eq. (3.18). However, more numerical resources would then be necessary in order to proceed.

Nevertheless, we might have learned enough about what happens in a single bounce to make a reliable extrapolation. We will attempt to do so by drawing a phase-space diagram. Since there are no gravitational degrees of freedom within spherical symmetry, the phase space of perturbations is given by all possible scalar field profiles. Due to energy conservation, we can focus on one fixed-energy, co-dimension-one surface in this infinite dimensional phase space. Within this surface, we can draw a two-dimensional projection (figure 3) and understand its structure based on our knowledge of the dynamics of one bounce.

One guiding principle of this diagram is that during one bounce, the profile changes by an infinitesimal amount $\sim \epsilon^2$, which is also an infinitesimal distance in the diagram. Within the weak-gravity time scale $\sim \epsilon^{-2}$, the evolution trajectory covers a finite distance of the diagram. In this way, the diagram directly represents the dynamical evolution in the rescaled time as given by eq. (2.7).

The horizontal axis of this two-dimensional diagram represents "how close is this profile to becoming a black hole". More technically, it is quantified by the maximum radial energy density at $r = 0$ that is reached during one AdS time. In the small-ϵ limit, the profile is

basically freely propagating, so this is a well-defined quantity. Heuristically, this maximum is reached when the highest "peak" goes through $r = 0$, and its value depends on the height of this peak, ρ_{Max}.

Note that throughout this paper, we have been referring to ρ as the rescaled energy density. In our conventions, the actual physical energy density is given by $\epsilon^2 \rho$. It is still convenient to consider the rescaled density here, since ρ_{Max} quantifies how much higher this peak is than the average, namely the relative concentration of energy. Its value increases toward the right-hand-side of figure 3. For any finite ϵ, there is a finite value of $\rho_{\text{Max}} \sim \epsilon^{-2}$ that represents a density high enough to become a black hole. It can be drawn as a vertical line. Some finite distance to the left of this line, we have another line signifying that the energy density is high enough to make gravity too strong to be described by the weak-gravity expansion.

To the left of this second line, gravity is weak enough that our analysis applies. Somewhere even further to the left, our approximation starts to fail for a different reason: we can no longer describe this peak as an isolated thin shell satisfying the hierarchy in eq. (3.5). In the small-ϵ limit, the region to apply our method always exists. This left boundary is not a very clear line. Nevertheless, in this diagram we can roughly picture it as $\rho_{\text{Max}} R_{\text{AdS}} \sim 1$, that the maximum peak density is comparable to the average density. Clearly, energy is too evenly distributed in the entire AdS space that nothing could be treated as an isolated thin shell.

The vertical axis of this diagram is not intended to represent any particular parameter of the field profile. It is merely reflecting the fact we established in section 4 that focusing and defocusing dynamics are equally likely in one bounce. This means one can always find some parameter such that the middle region is divided into two halves: in the upper half, the evolution makes the peak grow higher and moves closer to forming a black hole, and in the lower half, the peak gets lower and moves away from forming a black hole. In appendix B, we give specific numerical examples and argue the parameter controlling focusing and defocusing is the asymmetry of energy distribution: focusing occurs when the shell is is denser in the leading edge, and defocusing when it is denser in the tail.

In addition to focusing and defocusing which correspond to flowing horizontally in figure 3, what tendencies to flow in the vertical direction can we identify? Generally speaking, when ρ_{Max} is large, on the left side of figure 3, the system will tend to flow upwards. This is because a shrinking peak cannot remain the highest peak forever: a growing peak with smaller initial height will eventually take over. If that were the only vertical motion, it would lead to only two possible trajectories: starting in the upper half, ρ_{Max} would keep increasing and flow directly toward black hole collapse; alternatively, staring in the lower region, ρ_{Max} would first decrease, then bounces back to become a black hole. This directly violates many numerical results, so there must be a downward flow somewhere in figure 3.

The two numerical examples in appendix B provide tentative evidence for a downward flow. What we see is that given a symmetric profile on the boundary between the upper and lower region, after one bounce it picks up an asymmetry similar to the profiles in the lower region — its energy becomes denser in the tail. Of course, we studied only two one-
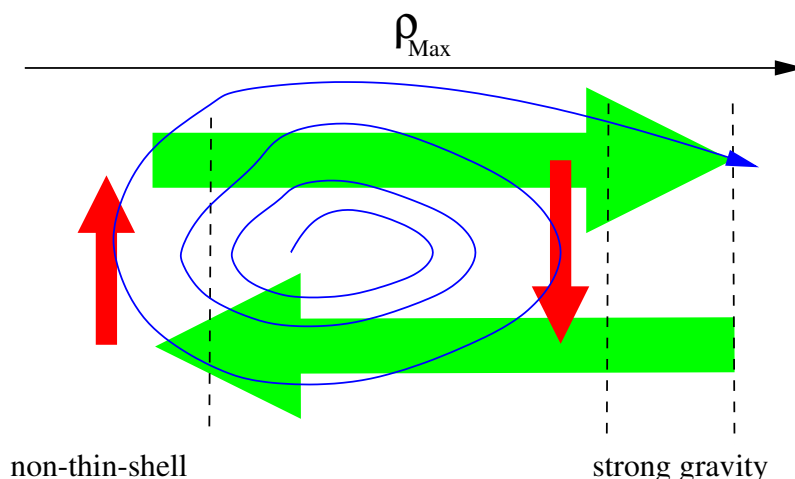
Figure 4. The same circular flow in the phase space, but the blue trajectory now represents an unstable solution. Though initially it follows the circular flow, it fluctuates to larger radius, eventually enters the strong gravity region, and collapses into a black hole.

parameter slices through an infinite dimensional phase space, so better numerical and/or analytical investigation is required to verify this. At the level of this paper, we will simply conjecture that such a downward flow exists, because the resulting circulatory flow, shown in figure 3, explains existing numerical results very well:

- The quasi-periodic solutions stay within the circular flow near the center, as in figure 3.

- The unstable solutions initially stay within the circular flow, but their radii vary wildly and eventually these solutions enter the strong gravity regime, as in figure 4

Note that the actual motion in the true phase space is still very complicated. In this is a two-dimensional projection, evolution trajectories are allowed to cross each other. Nevertheless, this circular flow allows us to better visualize the dynamics in the phase space.

We can also repeat the argument in section 2 in a more pictorial manner. As one reduces ϵ, most of this diagram does not change. Due to the scaling behavior, all trajectories to the left of the strong gravity line remain the same, and so most of the stable solutions remain stable. The trajectories for unstable solutions must cross the strong gravity line to form black holes, so they potentially can change.

Actually, the location of the strong gravity line shifts to the right when ϵ decreases. As the total energy is reduced, it needs to be increasingly focused in order for gravity to become strong, and a collapsing solution must therefore to evolve further, across the new weak-gravity regime. In the $\epsilon \to 0$ limit, black hole formation is equivalent to a weak-gravity evolution in which ρ_{Max} goes all the way to infinity.

Although the trajectory for a black hole collapse appears divergent, it does not mean that we can immediately rule out such an evolution. In fact, this divergence is an artifact of the parameter choice, and we should appreciate that $\rho_{\mathrm{Max}} = \infty$ is not an infinite distance

away in phase space. Recall that Δw in one bounce is independent of w, so the change in width need not be a small fraction of the total. It is certainly possible to have a profile such that after ϵ^{-2} bounces, Δw is negative and order one, leading to a diverging ρ_{Max}.

The real advantage of this picture is that it recasts the $\epsilon \to 0$ limit of the stability problem into the global regularity problem of determining whether ρ_{Max} diverges. Interestingly, analyzing the regularity of AdS_3 perturbations at finite ϵ below the black hole mass gap is similarly a question of determining whether the energy density diverges. In that case, there is already strong evidence to support regularity [25, 26]. One might hope to reproduce this AdS_3 result in higher dimensions in order to confirm that the instability corners indeed shrink to measure zero. We should again caution that spectral analysis can only provide necessary conditions; it can rule out an instability, but it cannot provide equally strong evidence to support one. If the power spectrum of perturbations agrees with a diverging ρ_{Max}, a long-time evolution of eq. (2.6) in position space is still required for the final answer to the AdS stability problem.

5.2 Holographic thermalization

One motivation for studying the stability properties of AdS is to try to learn something about the non-equilibrium dynamics of closed systems. This is due to the AdS/CFT correspondence, which relates this classical gravitational system to the dynamics of a strongly-coupled quantum system. Most investigations of holographic thermalization study the Poincaré patch of AdS, which has an infinite boundary (see, for example, [27–30]). In these cases, any nonzero energy density in the bulk will collapse into a black hole, corresponding to thermalization on the boundary.

Here, instead, we are considering global AdS which has a closed, spherical boundary and therefore a very different thermalization behavior. Other studies of global AdS, such as [22], implicitly assume the connection between forming a black hole in the bulk and thermalization of the boundary system. Although that is valid in some cases, we would like to highlight other possibilities. What are the possible holographic dual descriptions of the bulk story presented here?

One caveat is that explicit examples of the AdS/CFT correspondence usually contain compact extra dimensions whose sizes are comparable to R_{AdS}, for example in $AdS_5 \times S_5$. In the $\epsilon \to 0$ limit, the five-dimensional AdS-Schwarzschild black hole has a horizon radius much smaller than R_{AdS} and is therefore too small to represent the most typical states; a ten-dimensional black hole of the same total energy, which breaks the symmetry of the S^5, has even higher entropy. Therefore, the gravitational stability problem of AdS_5 does not directly translate to the thermalization problem in the boundary system. This might be an interesting direction for future work, but we will set this concern aside for now. Let us take a very optimistic point of view that the AdS/CFT correspondence can work with extra dimensions arbitrarily smaller than R_{AdS}, or even without them.

After limiting our attention to the AdS space and treating our classical field theory as a limit of a quantum gravity theory, the $\epsilon \to 0$ limit leads to a different issue. Recall the well known Hawking-Page transition [31]: a black hole does not always dominate the micro-cannonical ensemble; given low enough energy, thermal gas is the most typical state.

In this case, forming a black hole does not imply thermalization. This is the main issue we wish to clarify.

First of all, this issue highlights the importance of our position space approach. Focusing on the power spectrum, initial conditions occupying only low frequency modes must propagate to higher frequency in order to approach either a black hole or a thermal gas state. This type of turbulent cascade is a necessary condition for thermalization. However, without differentiating between the final states toward which the system could be evolving, one cannot argue unequivocally for or against thermalization.

Next, let us analyze under what circumstances the black hole or the thermal gas state will dominate the ensemble. For simplicity, we will work via dimensional analysis and ignore any order-one factors. First, note that the $\epsilon \to 0$ limit is actually the weak-gravity limit, corresponding to

$$\beta \equiv \frac{R_{\text{AdS}}}{R_{\text{BH}}} \gg 1 \; . \tag{5.1}$$

Namely, the Schwarzschild radius of the black hole made by collapsing the scalar field energy is much smaller than the AdS size. On the other hand, the most straightforward standard for trusting classical gravity is

$$\gamma \equiv \frac{R_{\text{BH}}}{l_{\text{Planck}}} \gg 1 \tag{5.2}$$

where we have restored the Planck scale $l_{\text{Planck}} = \sqrt{G\hbar}$, which has been set to one in the rest this paper. This condition implies that, at the very least, if a black hole forms, it could be described by classical gravity. For the limit we have been considering, both β and γ have been taken to infinity. We will see that whether the black hole or the thermal gas dominates depends on the details of how that limit is taken.

The entropy of a black hole with energy E is[15]

$$S_{\text{BH}} \sim (l_{\text{Planck}}E)^2 \sim \left(\frac{R_{\text{BH}}}{l_{\text{Planck}}}\right)^2 \; . \tag{5.3}$$

The entropy of a thermal gas in AdS space with the same total energy is

$$S_{\text{gas}} \sim \left(\frac{R_{\text{BH}}R_{\text{AdS}}}{l_{\text{Planck}}^2}\right)^{3/4} \; . \tag{5.4}$$

Thus, black hole states dominate the micro-canonical ensemble when

$$\left(\frac{R_{\text{BH}}}{l_{\text{Planck}}}\right) > \left(\frac{R_{\text{AdS}}}{R_{\text{BH}}}\right)^{3/2} \; ; \quad \gamma > \beta^{3/2} \; . \tag{5.5}$$

This condition is equivalent to comparing the thermal wavelength λ_{th} of the gas to the black hole radius; the black hole dominates the ensemble if

$$R_{\text{BH}} > \lambda_{\text{th}} \; . \tag{5.6}$$

---

[15]Note that we are assuming here that the spacetime is effectively AdS_4 at distance scales of order the size of the black hole; in string constructions, such as $AdS_4 \times S_7$, black holes whose radius is small compared to the AdS radius would be eleven-dimensional rather than four-dimensional, leading to different formulas.

We can see that whether the condition in eq. (5.5) is satisfied depends on how the limits of large β and large γ are taken. A classical and small-ϵ limit does not restrict the system to being dominated by either the thermal gas or black hole states.

Note that whether, and for how long, classical evolution is a good approximation depends on more details of the state. For example, even if a black hole forms which is classical according to eq. (5.2), the process by which it formed might not be. A simple rule of thumb for the validity of the classical limit is that the occupation numbers in the modes of interest have to be large. If the system is in a state where the energy is roughly equipartitioned between a number of modes up to some maximum frequency ω_{max}, we require

$$\text{energy per mode} \gg \omega_{\text{max}} . \tag{5.7}$$

The thermal gas states can never satisfy this condition because modes with frequency of order the temperature have occupation numbers of order one, yet contribute a significant fraction of the entropy of the gas. Independent of the limiting procedure and which states dominate, the thermal gas final state is never compatible with a classical description.[16]

On the other hand, a spherically symmetric collapse into a black hole can often be completely classical. Such a process only needs to excite the radial modes from the longest wavelength $\sim R_{\text{AdS}}$ to the shortest wavelength $\sim R_{\text{BH}}$, thus

$$\text{number of modes} = \frac{R_{\text{AdS}}}{R_{\text{BH}}} . \tag{5.8}$$

The condition on occupation numbers, eq. (5.7), becomes

$$\beta \ll \gamma^2 . \tag{5.9}$$

Comparing this to eq. (5.5), we see that eq. (5.9) is always true when the black hole dominates the ensemble, but it can still be true even if thermal gas dominates.[17] *Thus, the specific stability problem within classical gravity investigated in this paper, namely a spherically symmetric collapse into a black hole, is a valid dual to some boundary system, independent of whether such a process is equivalent to a efficient thermalization or not.*[18]

Furthermore, when the thermal gas dominates, if a black hole really forms in the time scale we investigated,

$$T_{\text{weak gravity}} = R_{\text{AdS}} \frac{R_{\text{AdS}}}{R_{\text{BH}}} \propto \epsilon^{-2} , \tag{5.10}$$

---

[16]Nevertheless, from the position-space viewpoint, classical evolution may still describe the "process of approaching" a thermal gas state, at least distinguishing it from approaching a black hole. In the later case energy becomes more concentrated, but in the former case it does not.

[17]Note that spherical symmetry is very important here. Without it, the number of modes would have been $\left(\frac{R_{\text{AdS}}}{R_{\text{BH}}}\right)^3$, and with that many modes, the black hole collapse would have failed to remain classical in the thermal gas-dominated regime.

[18]In this paper, "efficient" means that thermalization happens in the shortest time scale allowed by the dynamics, $\sim \epsilon^{-2}$. One should not confuse this with, for example, the much faster thermalization in the Poincaré patch of AdS, where, within one AdS time, perturbations cross the horizon, form a planar black hole, and appear to thermalize.

it could represent a significant delay to thermalization. In order to confirm this, we need to compare the naïve thermalization time $T_{\text{weak gravity}}$ to the black hole lifetime, given by the the evaporation time scale,

$$T_{\text{evaporation}} = \frac{R_{\text{BH}}^3}{l_{\text{Planck}}^2} \; .\tag{5.11}$$

When $T_{\text{evaporation}} > T_{\text{weak gravity}}$, which requires

$$\beta < \gamma \, ,\tag{5.12}$$

the system thermalizes only after forming a long-lived black hole, which eventually evaporates. This process of thermalization via a quasi-stationary thermal-like state is known as prethermalization and has been observed in finite-sized, isolated quantum systems [17, 18]. Note that eq. (5.12) is compatible with thermal gas domination and a classical collapse.

To summarize, spherically symmetric black hole formation within $T_{\text{weak gravity}}$ can have two different holographic interpretations:

- When $\gamma^2 > \beta^3$, it represents efficient thermalization of the boundary system.

- When $\beta < \gamma < \gamma^2 < \beta^3$, it represents prethermalization, which delays true thermalization (to thermal gas[19]) at a time scale $\gtrsim T_{\text{evaporation}} > T_{\text{weak gravity}}$.

For the remaining possibility, when $\gamma^2 < \beta^3$ but $\beta > \gamma$, the implication of black hole formation is inconclusive from a thermalization point of view. Black holes decay too fast to be quasi-stationary intermediate states, but their evaporation cannot guarantee reaching the thermal gas state either.

6 Summary

- By combining existing numerical data with our analysis, we have argued that for a massless scalar field in AdS space, in the small-amplitude $\epsilon \to 0$ limit, solutions remaining stable up to the interaction time scale $T \sim \epsilon^{-2}$ form an open set. This improves similar observations in finite-ϵ numerical simulations [7, 9] and argues against the conjecture that the weakly turbulent instability occurs in all but a set of measure zero in the space of initial conditions [1, 4, 14].

- One important difference between our approach and previous work is that we analyzed the problem in position space. We pointed out that only position space properties can provide necessary and sufficient conditions for the collapse into a black hole. Any analysis of the power spectrum can at most provide necessary conditions for black hole formation.

---

[19]Note that since thermal gas is never classical, we do not know exactly when will it really form. We only know that within $T_{\text{evaporation}}$, the systems was too busy forming a black hole and then remaining as one, so it cannot reach the thermal gas state yet.

- In the position space analysis, we exploited the small-amplitude $\epsilon \to 0$ limit and argued in section 3 that the only relevant dynamics are the gravitational self-interactions near $r = 0$. This argument requires a hierarchy of scales given in eq. (3.5), which is difficult to reach in realistic numerical simulations.

- We showed that gravitational self-interaction near $r = 0$ obeys an exact antisymmetry under time reversal. As a result, it is equally likely for interactions to focus or defocus energy. This equality is consistent with existing numerical results.[20] We remind the readers that gravity can be effectively repulsive: tidal forces tend to pull things apart. The possible defocusing of a radiation shell is due to such tidal effect.

- By making use of scaling symmetry, we simplified the stability problem in the $\epsilon \to 0$ limit into a global regularity problem within a finite rescaled time. The evolution was recast as a simple, first-order differential equation. We hope that this point of view, combined with the other techniques in this work and the existing literature, will allow a rigorous analysis of stability in the vanishing amplitude limit.

- Even if black holes do form in the ϵ^{-2} time scale, we point out that it does not always represent efficient thermalization of the boundary theory via AdS/CFT duality. In some cases, black hole formation describes prethermalization, and actual thermalization is delayed until this black hole evaporates.

Acknowledgments

---

[20]More specifically, one could take any numerical simulation and pause it at a moment when gravity is still weak. If one keeps the field profile but reverses the time derivative at this moment, the simulation will literally evolve backward toward the original initial profile, up to the numerical error and higher-order effects (which are small if the hierarchy of scales in eq. (3.5) is satisfied during the process).

A Analytical details

In this appendix, we will clarify some analytical details omitted in section 3. There we showed how to reach a simple differential equation for u_1, eq. (3.18), which can be solved simply by integrating the Green's function:

$$
\begin{aligned}
u_1(r, t_f) &= \frac{1}{2} \int_{-r_0}^{r_0} dt \int_{r-r_0+t}^{r+r_0-t} dr' \ S(r', t) \\
&= \frac{1}{2} \int_{-r_0}^{r_0} dt \int_{r-r_0+t}^{r+r_0-t} dr' \ \left(C(\ddot{u}_0 + u_0'') + \dot{C}\dot{u}_0 + C'\left(u_0' - \frac{u_0}{r'}\right) \right) . \quad \text{(A.1)}
\end{aligned}
$$

Here we should be careful about our method of images. A physical solution ϕ_1 is only given by a u_1 that is an odd function of r, and it is not obvious that the u_1 given by the above integral will have this property. Another potentially worrisome observation is that the lower limit of the r' integral can be negative for some positive r, but a physical answer should only invoke an integration over the physical space $r > 0$ where the quantity $C = V - M/r$ is naturally defined.

In this case, these concerns about the method of images can be easily resolved. As explained in section 3, we can generalize the definition of V and M to include the $r < 0$ region. We will find that M is an odd function of r and V is even. Together with the fact that u_0 is odd, we see that the integrand in eq. (A.1) is odd. Any integration over negative r is canceled by an equal region with positive r, so effectively the lower limit of the r' integral is $|r-r_0+t|$. Eq. (A.1) is effectively only integrating over the physical range. It is, however, more convenient to keep working in this form and avoid the confusion of taking an absolute value. An odd integrand here also guarantees that u_1 is an odd function which leads to a physical ϕ_1.

The form of eq. (A.1) clearly suggests some integrations by parts.

$$
\begin{aligned}
u_1(r, t_f) &= -\frac{1}{2} \int_{-r_0}^{r_0} dt \int_{r-r_0+t}^{r+r_0-t} dr' \ C'\frac{u_0}{r'} \ + \ \frac{1}{2} \int_{-r_0}^{r_0} dt \int_{r-r_0+t}^{r+r_0-t} dr' \ (Cu_0'' + C'u_0') \\
&\quad + \frac{1}{2} \int_{r-2r_0}^{r+2r_0} dr' \int_{-r_0}^{r_0-|r-r'|} dt \ (C\ddot{u}_0 + \dot{C}\dot{u}_0) \quad\quad\quad\quad \text{(A.2)} \\
&= -\frac{1}{2} \int_{-r_0}^{r_0} dt \int_{r-r_0+t}^{r+r_0-t} dr' \ C'\frac{u_0}{r'} \\
&\quad + \frac{1}{2} \int_{-r_0}^{r_0} dt \ C\Big[(r+r_0-t), t\Big] u_0'\Big[(r+r_0-t), t\Big] \\
&\quad - \frac{1}{2} \int_{-r_0}^{r_0} dt \ C\Big[(r-r_0+t), t\Big] u_0'\Big[(r-r_0+t), t\Big] \\
&\quad + \frac{1}{2} \int_{r-2r_0}^{r} dr' \ C\Big[r', (r_0-r+r')\Big] \dot{u}_0\Big[r', (r_0-r+r')\Big] \\
&\quad + \frac{1}{2} \int_{r}^{r+2r_0} dr' \ C\Big[r', (r_0-r'+r)\Big] \dot{u}_0\Big[r', (r_0-r'+r)\Big] \\
&\quad - \frac{1}{2} \int_{r-2r_0}^{r+2r_0} dr' \ C(r', -r_0) \dot{u}_0(r', -r_0)
\end{aligned}
$$

$$
\begin{aligned}
= \; & \frac{1}{2}\int_{-r_0}^{r_0} dt\; C\Big[(r-r_0+t),t\Big]\left(\dot{u}_0\Big[(r-r_0+t),t\Big]-u_0'\Big[(r-r_0+t),t\Big]\right) \\
& +\frac{1}{2}\int_{-r_0}^{r_0} dt\; C\Big[(r+r_0-t),t\Big]\left(\dot{u}_0\Big[(r+r_0-t),t\Big]+u_0'\Big[(r+r_0-t),t\Big]\right) \\
& -\frac{1}{2}\int_{-r_0}^{r_0} dt \int_{r-r_0+t}^{r+r_0-t} dr'\; C'\frac{u_0}{r'}-\frac{1}{2}\int_{r-2r_0}^{r+2r_0} dr'\; C(r',-r_0)\dot{u}_0(r',-r_0)\;.
\end{aligned}
$$

In the above equation, we first isolated two terms which should be integrated by parts, and for one of them we interchange the order of integration so it can be done with respect to t instead of r'. The integration by parts produces five boundary terms as line integral along five segments which we explicitly write down. Finally, two pairs of segments can combine with each other and be expressed as time integrals. We collect the remaining space integral and the only non-boundary term which cannot be integrated by parts in the end.

Note that up to this step, we have not used any approximations. We did not even use the property that C is sourced by ϕ_0. In other words, this expression could describe the change in the field profile under the influence of any other spherically symmetric gravitational effects, either apart from or on top of its self-interaction.

Our next step is to plug in eq. (3.15) and use our assumption that it represents a thin shell: we assume that u_0 is only nonzero within two narrow packages around $r=t$ and $r=-t$. This significantly simplifies eq. (A.2) to

$$
\begin{aligned}
u_1(r,t_f) = \; & -\frac{\epsilon}{\sqrt{w}}\;f'\left(\frac{r-r_0}{w}\right)\int_{-r_0}^{r_0} dt\; C\Big[(r-r_0+t),t\Big] \\
& +\frac{\epsilon}{\sqrt{w}}\;f'\left(\frac{-r-r_0}{w}\right)\int_{-r_0}^{r_0} dt\; C\Big[(r+r_0-t),t\Big] \\
& -\frac{1}{2}\int_{-r_0}^{r_0} dt \int_{r-r_0+t}^{r+r_0-t} dr'\; C'\frac{u_0}{r}-\frac{1}{2}\int_{r-2r_0}^{r+2r_0} dr'\; C(r',-r_0)\dot{u}_0(r',-r_0)\;.
\end{aligned}
\tag{A.3}
$$

Note that here the f' means a derivative with respect to the variable of f instead of a r derivative, which should always be clear from the context.

Since in the end, we are only interested in the physical range $r>0$, we can actually drop the second term because the profile f is zero there. This starts to take the promised form of eq. (3.19), and we can almost identify

$$
\Delta r = \int_{-r_0}^{r_0} dt\; C\Big[|r-r_0+t|,t\Big]\;.
\tag{A.4}
$$

Note that we have added an absolute value to the first variable in C. This makes no difference since it is even, but it helps to emphasize the fact that the integral can be strictly limited to the physical $r>0$ region.

The physical meaning of eq. (A.4) now becomes clear. When the metric includes first-order corrections, such as in eq. (3.10), a null ray actually follows

$$
\left(1+\frac{M}{r}\right)|dr| = (1+V)\,dt\;.
\tag{A.5}
$$

Thus an incoming null ray starting from $r = r_0$ and $t_i = -r_0$ does not exactly return to $r = r_0$ at $t_f = r_0$; the amount it misses is exactly given by eq. (A.4). The leading-order correction due to gravity, of course, includes the fact that geodesics are changed, and the shell simply follows the new geodesic. A geometric calculation is enough to determine how much a localized object appears to be shifted from the position predicted by the zeroth-order theory.

For any finite-sized source, the gravitational potential at large r is proportional to $1/r$, so the integral in eq. (A.4) actually had a piece proportional to $\log r_0$. Since we are have taken r_0 to be large, one might have worried that such a term would the ruin perturbation expansion. However, such the $\log r_0$ term is totally expected from the change of geodesics and does not interfere with our goal of computing the change in width or other changes.

One last concern about the position shift in eq. (A.4) is that it is a function of r. This turns out not to be a problem either, since the r-dependent part of Δr is not proportional to $\log r_0$. We can see this by taking a derivative with respect to r:

$$\partial_r \Delta r = \partial_r \left(\int_{-r_0}^{r_0-r} dt\, C\Big[(r_0 - r - t), t\Big] + \int_{r_0-r}^{r_0} dt\, C\Big[(t - r_0 + r), t\Big] \right)$$
$$= \left(-\int_{-r_0}^{r_0-r} dt\, C'\Big[(r_0 - r - t), t\Big] + \int_{r_0-r}^{r_0} dt\, C'\Big[(t - r_0 + r), t\Big] \right) . \quad (A.6)$$

According to the Einstein's equation, we have

$$C' = V' - \frac{M'}{r} + \frac{M}{r^2} = \frac{2M}{r^2} + \frac{r}{2}\left(T_{rr} - T_{tt}\right) . \quad (A.7)$$

This means that as long as we restrict the matter sources to (1) finite-sized sources that vanish beyond some fixed r and/or (2) radiation in the radial direction, $T_{rr} = T_{tt}$, then the r dependence of Δr will not have a $\log r_0$ (or any other large r_0) dependence. Furthermore, there is no small-r divergence either, since the two terms in eq. (A.6) takes opposite signs and cancel each other near $r = 0$. Pictorially, this means that different infinitesimal segments within the wavepacket "shift" differently from one another by some finite amount.

In the last line of eq. (A.3), the first term is also finite for the same reason as eq. (A.7), and the second term is finite because u_0 has compact support. These terms should be combined and understood as some perturbative deformations of the wavepacket profile. They are cleanly separated from the $\Delta r \sim \log r_0$ overall shift, which is uniquely defined by a projection:

$$\Delta r = -\frac{\int u_1 \partial_r u_0 \, dr}{\int (\partial_r u_0)^2 \, dr} . \quad (A.8)$$

We can simply remove this shift mode from eq. (3.19) and study the other deformations. A more physical way to understand the removal of this shift is letting the wavepacket evolve an extra time $\Delta t = \Delta r$ such that it really reaches position r_0; then it will be fair to compare with the zeroth-order profile at the same position.

In order to eventually form a black hole, we need the energy density to become large. Since the total energy is conserved, the most trivial way to increase the energy density is

to narrow the width of the profile. The leading-order change in width can be extracted from u_1 by the following projection:

$$\Delta w = \frac{\int u_1 \partial_w u_0 \ dr}{\int (\partial_w u_0)^2 \ dr} \ . \tag{A.9}$$

Note that the ϵ dependence was already scaled out in eq. (3.6). Interestingly, ϵ^2 has the unit of length in our conventions, and the physical change in width is $\epsilon^2 \Delta w$. Therefore, Δw is dimensionless. The width w is the only other dimensionful quantity that can potentially affect Δw in the leading order (r_0 affects only the subleading error), and so there is no way it can enter the expression for Δw.

What we really wish to determine the sign of Δw. $\Delta w < 0$ means that the shell gets narrower, and several bounces later it might form a black hole. On the other hand, $\Delta w > 0$ means that the shell gets wider, and several bounces later the energy will be more diluted, which in some sense is moving away from a black hole in phase space.

One technical point to note here is that, given our shell profile eq. (3.15), the mode $\partial_w u_0$ actually measures the scaling the profile around some center. If that center is not the center of mass, then this scaling not only changes the width but also shifts the position. A simple projection will be contaminated by the large $\sim \log r_0$ contribution from the position shift. We will avoid this by always defining the profile $f(x)$ to have its center of mass at $x = 0$. This means that, on top of the normalization

$$\int f'(x)^2 \ dx = 1 \, , \tag{A.10}$$

we also demand that

$$\int x f'(x)^2 \ dx = 0 \ . \tag{A.11}$$

This guarantees that the scaling mode $\partial_w u_0$ is orthogonal to the shift mode $\partial_r u_0$.

B Numerical examples

B.1 The asymmetry-focusing correlation

In this appendix, we numerically evaluate the change to a thin-shell profile after one bounce. Our example will be the following two one-parameter families of profiles.
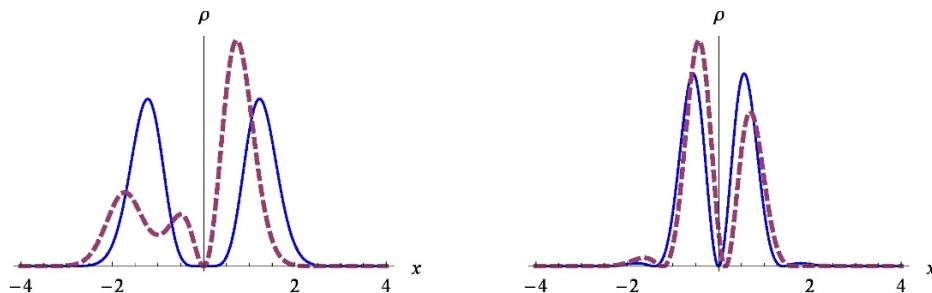
$$g_a(x) = (1 + ax \pm x^2) \ e^{-x^2} \, , \tag{B.1}$$

$$N_a = \sqrt{\int g_a'^2 \ dx} \, , \tag{B.2}$$

$$X_a = N_a^{-2} \int x g_a'^2 \ dx \, , \tag{B.3}$$

$$f_a(x) \equiv N_a^{-1} g_a(x + X_a) \ . \tag{B.4}$$

They are symmetric when $a = 0$, and varying a scans through two different directions of asymmetry. Note that the quadratic term is necessary. Without it, a small a simply means

Figure 5. The left panel shows the energy density of the profiles with the "+" sign defined in eq. (B.1), and the right panel shows the profiles with the "−" sign. The blue curves are the symmetric, $a = 0$ profiles. The red (dashed) curves are for $a = 0.5$, which is the maximally "tilted" profile in the range we scanned through. We can see that the family with the "+" sign is more sensitive to the parameter a.

an overall shift in position and the profile will be still symmetric to leading order. Our definition of $f_a(x)$ shifts the center of mass back to $x = 0$ and preserves only the asymmetry generated by a.

We plot some representative profiles in figure 5. Note that for both families, we have $f_a(x) = f_{-a}(-x)$. Scanning through positive and negative values of a can confirm our analytical proof in section 4 that flipping the profiles leads to opposite behaviors within one bounce. It will also provide a better understanding about what physical quantity really affects whether a profile becomes focused or not.

There are infinite ways to be asymmetric, and our parameter a certainly is not the unique parameter to quantify the asymmetry. It also has no reason to be the asymmetry directly responsible for focusing or defocusing the profile. However, for any family of profiles centered around a symmetric one, we can define a natural parameter to quantify the asymmetry, at least for small values of a. Here is how it goes. First of all, g_a has a center of mass shifted by X_a from g_0 by the definition in eq. (B.3). On the other hand, one can also naturally define the shift by a projection to the zero mode, which is exactly the way we defined Δr in eq. (A.8).

$$\bar{X}_a = N_0^{-2} \int [g_0(x) - g_a(x)]g_0'(x) \, dx \ . \tag{B.5}$$

These two shifts already disagree at linear order in a, therefore the amount of their disagreement, $\Delta X_a = (X_a - \bar{X}_a)$, seems to be a reasonable way to quantify the asymmetry.

Given these profiles, we solve eq. (3.10) for M and V, and then we can integrate eq. (A.1). When we scan the parameter a from -0.5 to 0.5, the change in width Δw defined in eq. (A.9) follows a pattern closely resembling the behavior of this asymmetry parameter, ΔX_a. We compare them side-by-side in figure 6. Note that they are not identical. For example, the relative slopes between the two families near $a = 0$ are not the same. Thus, although we see a rough correlation between them, we cannot claim that our asymmetry parameter directly controls focusing or widening in one bounce.
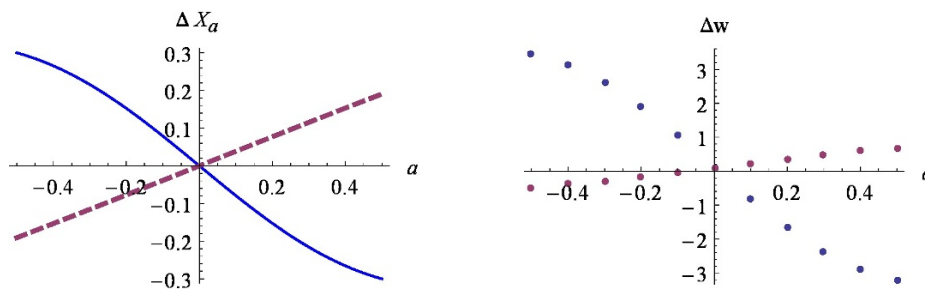
Figure 6. The left panel shows the asymmetry parameter defined as the difference between the center-of-mass shift and the field-profile shift. The blue curve is for the "+" family and the red curve for the "−" family. The right panel shows the change is width after one bounce, which qualitatively agrees with the asymmetry parameter. These are done with $r_0 = 60$ and $w = 1$. Recall that the physical change in width is actually $\epsilon^2 \Delta w$.

In our conventions, the profiles are moving toward the right in figure 5. If we compare their shapes in figure 5 to their behaviors in figure 6, we get the following impression:

- When the wavepacket is denser in the front, we get $\Delta w < 0$. The shell gets focused into a smaller region, and gravitational effect will become stronger in the next bounce.

- When the wavepacket is denser in the tail, we get $\Delta w > 0$. It profile gets wider after one bounce, and gravitational effect will become weaker in the next bounce.

Since the family of profiles with "+" sign is much more sensitive to the parameter a, we will use it to test other behaviors later in appendix B.2.

In figure 6, one might notice that for the $a = 0$ symmetric profiles, the changes in width are not exactly zero as we argued earlier. This deviation is not physical but simply an artifact of our approximation. That is because although the physical solution is symmetric in time, our technical choice breaks that symmetry by a small amount. We have set the correction to the field profile at the initial time to zero, $u_1(r, -r_0) = 0$. This is a small error since the first-order correction to the metric in eq. (3.10) would have already modified the free field profile at that time, by a small fraction $\sim \epsilon^2 V \sim \epsilon^2/r_0$.

We test this explanation by varying r_0 and verifying that Δw goes to zero in the expected way; see figure 7. We also verified that the position shift indeed has an r_0-dependent shift $\Delta r \propto \log r_0$, as discussed in appendix A.

Finally, with a symmetric profile, one can ask for a prediction for the next bounce. What we observed in these examples is that a symmetric profile will pick up a $\Delta X_a > 0$ in one bounce. This is a very tentative evidence that in the next bounce, they will have $\Delta w > 0$, namely their energy become defocused. We stress again that this is not a proof, but merely two examples we observed. A more thorough investigation is required to support the generic downward flow we conjectured in section 5.

B.2 The effect of another object

Black hole formation does not always involve all of the energy becoming concentrated into a thin shell. For example, an initially smooth field profile might start to develop one or
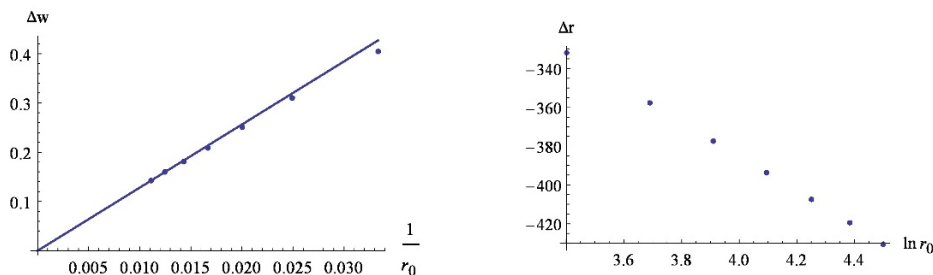
Figure 7. Results with $a = 0$ and varying r_0 from 30 to 90 in steps of 10. The left panel shows Δw, which indeed goes to zero as $1/r_0$. The right panel shows the position shift Δr defined in eq. (A.8) which has the correct $\log r_0$ dependence.

more sharp peaks. It is possible for the energy density in these peaks to become large enough to induce strong gravity and black hole collapse before or even without the average density of the entire profile ever becoming large.

In this section, we will present some evidence that sharp peaks evolve similarly to thin shells. In the perturbative regime, one is free to separate the matter into components in many ways. In particular, we can treat a smooth field profile with a sharp peak as a thin shell propagating in the background of some additional diffuse source of gravity. Our approach is convenient since eq. (A.1) and further analysis about it do not rely on the specific metric ansatz eq. (3.10). As long as the additional source are also spherically symmetric, we can simply repeat the calculation in the previous appendix.

We will start by considering a simple situation in which the additional matter sources are static. In addition to the thin shell with total mass $4\pi\epsilon^2$, we have

$$M_0(r) = 10\epsilon_{\text{star}}^2 \tanh\left(\frac{r}{w_{\text{star}}}\right)^2 , \tag{B.6}$$

$$P_r \approx 0 . \tag{B.7}$$

This is a star of roughly constant density, radius w_{star} and total mass $10\epsilon_{\text{star}}^2$. It does not interact with the massless field which forms the shell in any other way other than gravitationally, so it simply enters by altering the metric in eq. (3.10). We assume the star is stable and supports itself by a negligible amount of radial pressure (but it can have angular pressure), so it does not add any extra term to modify T_{rr}.

According to the momentum space analysis, including this additional gravitational source breaks the AdS resonance structure and should interfere with black hole formation [12]. We show that such an interpretation is not necessary to understand the dynamics of thin shells in one bounce. Remember that for a symmetric shell profile, we argued that there cannot be a change in width due to the time-reversal symmetry. Adding an extra, static source does not break that symmetry, so symmetric profiles again cannot change in size. And, it is straightforward to verify that asymmetry still focuses or defocuses in qualitatively the same way as before.

In figure 8, we demonstrate that whether the shell gets thinner or thicker has the same dependence on the asymmetry induced by the parameter a. Its magnitude does have an
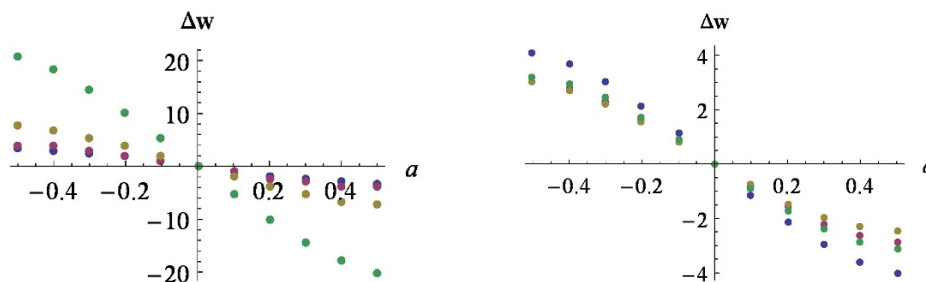
Figure 8. We plot the change in width Δw as a function of the parameter a in eq. (B.1). Both figures are with $r_0 = 30$. The left panel shows the effect of dialing the mass of the additional matter source, $\epsilon_{\mathrm{star}}^2 = 0, 10, 50, 200$, without changing its size $w_{\mathrm{star}} = 1$. The right panel shows the effect of dialing the size while keeping the same density, $w_{\mathrm{star}} = 1, 5, 20, 200$. We have removed the errant $1/r_0$ contribution by hand.

interesting dependence on the additional source. In the first set of data we fix the size of the star to be equal to the shell. The change in width turns out to grow linearly with the additional mass. On the other hand if we keep the same density and increase the size of the star, the change in width is not strongly affected.

Open Access. This article is distributed under the terms of the Creative Commons Attribution License ([CC-BY 4.0](#)), which permits any use, distribution and reproduction in any medium, provided the original author(s) and source are credited.

References

[1] P. Bizon and A. Rostworowski, *On weakly turbulent instability of anti-de Sitter space*, *Phys. Rev. Lett.* **107** (2011) 031102 [`arXiv:1104.3702`] [INSPIRE].

[2] H.P. de Oliveira, L.A. Pando Zayas and E.L. Rodrigues, *A Kolmogorov-Zakharov Spectrum in AdS Gravitational Collapse*, *Phys. Rev. Lett.* **111** (2013) 051101 [`arXiv:1209.2369`] [INSPIRE].

[3] S.L. Liebling, *Nonlinear collapse in the semilinear wave equation in AdS space*, *Phys. Rev.* **D 87** (2013) 081501 [`arXiv:1212.6970`] [INSPIRE].

[4] O.J.C. Dias, G.T. Horowitz, D. Marolf and J.E. Santos, *On the Nonlinear Stability of Asymptotically Anti-de Sitter Solutions*, *Class. Quant. Grav.* **29** (2012) 235019 [`arXiv:1208.5772`] [INSPIRE].

[5] M. Maliborski, *Instability of Flat Space Enclosed in a Cavity*, *Phys. Rev. Lett.* **109** (2012) 221101 [`arXiv:1208.2934`] [INSPIRE].

[6] A. Buchel, L. Lehner and S.L. Liebling, *Scalar Collapse in AdS*, *Phys. Rev.* **D 86** (2012) 123011 [`arXiv:1210.0890`] [INSPIRE].

[7] A. Buchel, S.L. Liebling and L. Lehner, *Boson stars in AdS spacetime*, *Phys. Rev.* **D 87** (2013) 123006 [`arXiv:1304.4166`] [INSPIRE].

[8] P. Bizoń, *Is AdS stable?*, *Gen. Rel. Grav.* **46** (2014) 1724 [`arXiv:1312.5544`] [INSPIRE].

[9] M. Maliborski and A. Rostworowski, *Lecture Notes on Turbulent Instability of Anti-de Sitter Spacetime*, *Int. J. Mod. Phys.* **A 28** (2013) 1340020 [`arXiv:1308.1235`] [INSPIRE].

[10] M. Maliborski and A. Rostworowski, *Time-Periodic Solutions in an Einstein AdS-Massless-Scalar-Field System*, *Phys. Rev. Lett.* **111** (2013) 051102 [arXiv:1303.3186] [INSPIRE].

[11] V. Balasubramanian, A. Buchel, S.R. Green, L. Lehner and S.L. Liebling, *Holographic Thermalization, Stability of Anti-de Sitter Space and the Fermi-Pasta-Ulam Paradox*, *Phys. Rev. Lett.* **113** (2014) 071601 [arXiv:1403.6471] [INSPIRE].

[12] M. Maliborski and A. Rostworowski, *What drives AdS spacetime unstable?*, *Phys. Rev.* **D 89** (2014) 124006 [arXiv:1403.5434] [INSPIRE].

[13] G.T. Horowitz and J.E. Santos, *Geons and the Instability of Anti-de Sitter Spacetime*, arXiv:1408.5906 [INSPIRE].

[14] O.J.C. Dias, G.T. Horowitz and J.E. Santos, *Gravitational Turbulent Instability of Anti-de Sitter Space*, *Class. Quant. Grav.* **29** (2012) 194002 [arXiv:1109.1825] [INSPIRE].

[15] P. Bizoń and A. Rostworowski, *Comment on "Holographic Thermalization, Stability of Anti-de Sitter Space and the Fermi-Pasta-Ulam Paradox"*, *Phys. Rev. Lett.* **115** (2015) 049101 [arXiv:1410.2631] [INSPIRE].

[16] B. Craps, O. Evnin and J. Vanhoof, *Renormalization group, secular term resummation and AdS (in)stability*, *JHEP* **10** (2014) 48 [arXiv:1407.6273] [INSPIRE].

[17] S. Trotzky, Y.A. Chen, A. Flesch, I.P. McCulloch, U. Schollwöck, J. Eisert and I. Bloch, *Probing the relaxation towards equilibrium in an isolated strongly correlated one-dimensional Bose gas*, *Nature Physics* **8** (2012) 325 [arXiv:1101.2659].

[18] M. Gring et al., *Relaxation and Prethermalization in an Isolated Quantum System*, *Science* **337** (2012) 1318 [arXiv:1112.0013].

[19] A.D. Rendall, *Convergent and divergent perturbation series and the post-Minkowskian approximation scheme*, *Class. Quant. Grav.* **7** (1990) 803.

[20] P. Bizon, T. Chmaj and A. Rostworowski, *Late-time tails of a self-gravitating massless scalar field revisited*, *Class. Quant. Grav.* **26** (2009) 175006 [arXiv:0812.4333] [INSPIRE].

[21] P. Basu, C. Krishnan and A. Saurabh, *A Stochasticity Threshold in Holography and and the Instability of AdS*, *Int. J. Mod. Phys.* **A 30** (2015) 1550128 [arXiv:1408.0624] [INSPIRE].

[22] J. Abajo-Arrastia, E. da Silva, E. Lopez, J. Mas and A. Serantes, *Holographic Relaxation of Finite Size Isolated Quantum Systems*, *JHEP* **05** (2014) 126 [arXiv:1403.2632] [INSPIRE].

[23] M.A. Amin, E.A. Lim and I.-S. Yang, *A scattering theory of ultrarelativistic solitons*, *Phys. Rev.* **D 88** (2013) 105024 [arXiv:1308.0606] [INSPIRE].

[24] M.A. Amin, E.A. Lim and I.-S. Yang, *Clash of Kinks: Phase Shifts in Colliding Nonintegrable Solitons*, *Phys. Rev. Lett.* **111** (2013) 224101 [arXiv:1308.0605] [INSPIRE].

[25] P. Bizoń and J. Jałmuzna, *Globally regular instability of AdS$_3$*, *Phys. Rev. Lett.* **111** (2013) 041102 [arXiv:1306.0317] [INSPIRE].

[26] J. Jałmuzna, *Three-dimensional Gravity and Instability of AdS$_3$*, *Acta Phys. Polon.* **B 44** (2013) 2603 [arXiv:1311.7409] [INSPIRE].

[27] S. Lin and E. Shuryak, *Toward the AdS/CFT Gravity Dual for High Energy Collisions. 3. Gravitationally Collapsing Shell and Quasiequilibrium*, *Phys. Rev.* **D 78** (2008) 125018 [arXiv:0808.0910] [INSPIRE].

[28] S. Bhattacharyya and S. Minwalla, *Weak Field Black Hole Formation in Asymptotically AdS Spacetimes*, *JHEP* **09** (2009) 034 [arXiv:0904.0464] [ɪɴSPIRE].

[29] T. Albash and C.V. Johnson, *Evolution of Holographic Entanglement Entropy after Thermal and Electromagnetic Quenches*, *New J. Phys.* **13** (2011) 045017 [arXiv:1008.3027] [ɪɴSPIRE].

[30] V. Balasubramanian et al., *Holographic Thermalization*, *Phys. Rev.* **D 84** (2011) 026010 [arXiv:1103.2683] [ɪɴSPIRE].

[31] S.W. Hawking and D.N. Page, *Thermodynamics of Black Holes in anti-de Sitter Space*, *Commun. Math. Phys.* **87** (1983) 577 [ɪɴSPIRE].