Holographic proof of the strong subadditivity of entanglement entropy

Matthew Headrick
Stanford Institute for Theoretical Physics, Stanford, California 94305-4060, USA

Tadashi Takayanagi
Department of Physics, Kyoto University, Kyoto, 606-8502, Japan

(Received 5 May 2007; published 21 November 2007)

When a quantum system is divided into subsystems, their entanglement entropies are subject to an inequality known as *strong subadditivity*. For a field theory this inequality can be stated as follows: given any two regions of space A and B, $S(A) + S(B) \geq S(A \cup B) + S(A \cap B)$. Recently, a method has been found for computing entanglement entropies in any field theory for which there is a holographically dual gravity theory. We give a simple geometrical proof of strong subadditivity employing this holographic prescription.

DOI: 10.1103/PhysRevD.76.106013 PACS numbers: 11.25.Tq, 03.67.Mn, 04.60.−m, 04.70.Dy

Entanglement entropy is an important tool in the study of quantum information (see e.g. [1]). It quantifies the extent to which the state of a given subsystem of a quantum system is correlated with that of the rest of the system. Entanglement entropy enjoys a crucial mathematical property called *strong subadditivity* [2]. Recently, Ryu and one of the authors of the present paper [3] proposed a relationship, applicable to any quantum field theory with a holographic gravity dual, between the entanglement entropy of a region of space in the quantum field theory (QFT) and the area of a certain minimal surface in the dual spacetime. An important test of the validity of this proposal is whether it has the property of strong subadditivity. This question was investigated in a variety of examples in the paper [4], always with an affirmative answer. In this note we give a general argument that it does, based only on general properties of holographic dualities. Besides giving support to the proposal, our argument provides an intuitive, geometrical way to understand strong subadditivity, a property whose formal algebraic proof is highly nontrivial.

The same argument can be used to establish a concavity property for holographically computed Wilson loop expectation values when the loops are coplanar. This is briefly discussed at the end of the paper.

The von Neumann entropy of a density matrix ρ, $S(\rho) = -\mathrm{Tr}(\rho \ln\rho)$, quantifies the extent to which the state represented by ρ fails to be a pure state. If ρ is obtained by tracing over part of the Hilbert space representing a subsystem—for example, one that is inaccessible to the experimentalist—then $S(\rho)$ is referred to as the *entanglement entropy* of the remaining subsystem. More formally, if the Hilbert space of the full system factorizes into Hilbert spaces of two subsystems, $\mathcal{H}_{\mathrm{full}} = \mathcal{H}_1 \otimes \mathcal{H}_2$, then for each subsystem we define a reduced density matrix $\rho_1 = \mathrm{Tr}_{\mathcal{H}_2}\rho_{\mathrm{full}}$, $\rho_2 = \mathrm{Tr}_{\mathcal{H}_1}\rho_{\mathrm{full}}$, and a corresponding entanglement entropy $S(\rho_1)$ and $S(\rho_2)$. On the basis of the concavity of the function $-x\ln x$ and elementary properties of Hilbert spaces, these can be shown quite generally to obey the following inequalities:

$$|S(\rho_1) - S(\rho_2)| \leq S(\rho_{\mathrm{full}}) \leq S(\rho_1) + S(\rho_2). \tag{1}$$

This property of entanglement entropy is known as *subadditivity* (the first inequality is also called the Araki-Lieb inequality [5]). In particular, if the full system is in a pure state then the two subsystems have the same entanglement entropy.

Now suppose the system is made up of more than two subsystems, $\mathcal{H}_{\mathrm{full}} = \bigotimes_i \mathcal{H}_i$. Then the inequalities (1) can be strengthened to yield [2]

$$\begin{aligned} S(\rho_{12}) + S(\rho_{23}) &\geq S(\rho_2) + S(\rho_{123}), \\ S(\rho_{12}) + S(\rho_{23}) &\geq S(\rho_1) + S(\rho_3), \end{aligned} \tag{2}$$

where ρ_{12} is the reduced density matrix for $\mathcal{H}_1 \otimes \mathcal{H}_2$, etc. These two inequalities can be shown to be equivalent by the formal device of adding a fourth subsystem such that ρ_{1234} is a pure state. This property of entanglement entropy is known as *strong subadditivity*, and its proof is highly nontrivial (although again it depends only on elementary properties of Hilbert spaces) [6]. Strong subadditivity represents the concavity of the von Neumann entropy and is a sufficiently strong property that it essentially uniquely characterizes the von Neumann entropy [7].

In the context of a quantum field theory, a natural type of subsystem to consider is that associated with a given region of space. To any region A is associated a Hilbert space $\mathcal{H}_A$, and for two disjoint regions A and B we have $\mathcal{H}_{A\cup B} = \mathcal{H}_A \otimes \mathcal{H}_B$. For the entanglement entropy associated to $\mathcal{H}_A$ we will write simply $S(A)$ [rather than $S(\rho_A)$]. Because of the infinite number of degrees of freedom involved in a field theory, $S(A)$ typically suffers from an ultraviolet divergence proportional to the surface area of A [8,9]. In order to deal with finite quantities one must impose a UV cutoff (and, if the surface area of A is infinite, an IR cutoff as well). One may also consider subtracted quantities that remain finite as the UV cutoff is removed, such as the *mutual information*, $I(A, B) = S(A) + S(B) - S(A \cup B)$, defined when A and B (and their surfaces) are disjoint. By (1) this is non-negative. By employing these quantities, Casini and Huerta [10] showed that an analogue of the c-theorem in two-dimensional QFTs can be derived from strong subadditivity.

To avoid confusion, it is important to remember that the concept of entanglement entropy refers to a specific state of the system at a specific time. Therefore all of the regions and surfaces we consider in this paper are restricted to a fixed constant-time slice of the field theory's spacetime.

Recently, a proposal has been made in [3] for how to compute the entanglement entropy of a region of space in any quantum field theory that admits a holographic gravity dual. The proposal is very simple. The gravity theory lives in a space which as usual we call the bulk, and the QFT on its conformal boundary. (To avoid confusion, we will reserve the term "boundary" for the space on which the QFT lives, and use the term "surface" for the boundaries of various regions in the bulk and boundary.) We consider all hypersurfaces [11] m in the bulk that end on ∂A, and ask for the one with minimal area. (See Fig. 1.) We then have

$$S(A) = \frac{1}{4G_{\rm N}} \min_{m:\partial m = \partial A} a(m), \qquad (3)$$

where $a(m)$ is the area of m. For the case when the bulk gravity theory lives on a static asymptotically anti–de Sitter (AdS) spacetime, Fursaev [12] has given a derivation of (3) using Euclidean quantum gravity and the basic principles [13] of the anti–de Sitter/conformal field theories (AdS/CFT) correspondence [14]. Notice that expression (3) coincides with the Bekenstein-Hawking formula of black hole entropy if we replace the minimal surface with a black hole horizon. Indeed, at high temperature the spacetime of the gravity theory generally includes a horizon; when part of the minimal surface wraps the horizon, its contribution corresponds to the usual thermal entropy. We will see an example of this situation when we come to Fig. 2 below.

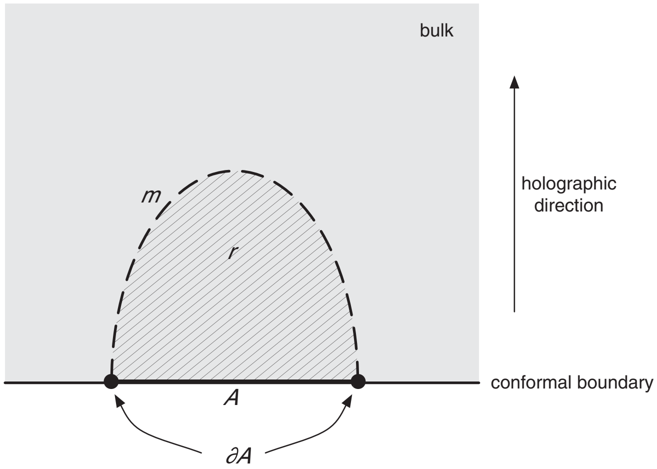

FIG. 1. A constant-time slice of a spacetime on which a gravity theory lives, and the conformal boundary on which its holographically dual field theory lives. A is a region of the boundary; m is the minimal hypersurface in the bulk ending on ∂A; and r is a region of the bulk such that $\partial r = A \cup m$.

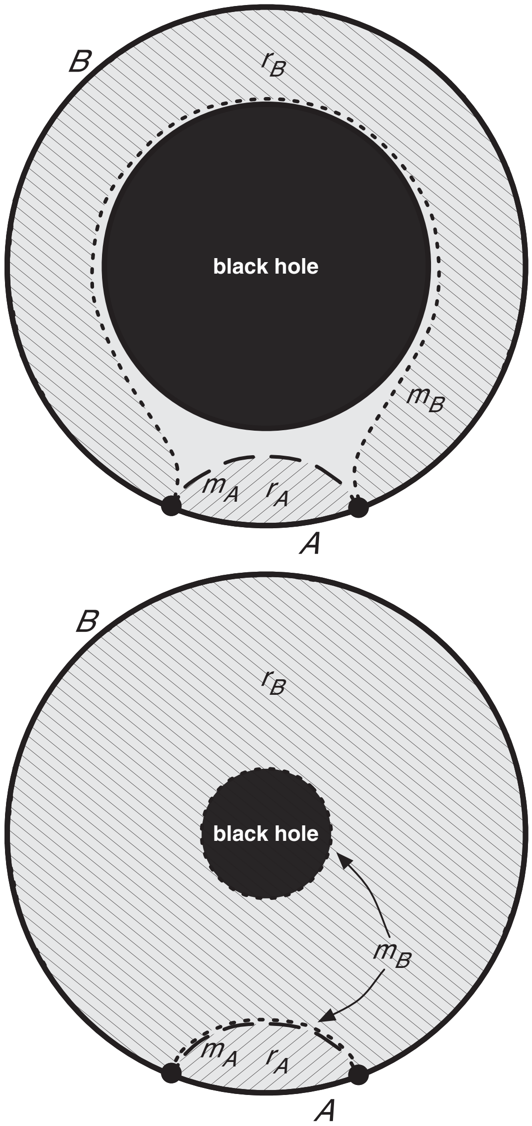

FIG. 2. Two examples in which the QFT lives on a compact space and the bulk contains a black hole. (Technically, since we are considering an eternal black hole in static coordinates, in each case the full spacetime consists of two copies of the region shown connected by an Einstein-Rosen bridge; this will not affect our discussion.) The boundary is divided into two regions A and B. Since $\partial A = \partial B$, if the bulk had trivial $(d-1)$st homology the corresponding minimal hypersurfaces m_A and m_B would be identical, and we would have $S(A) = S(B)$. However, due to the requirement that each hypersurface be homologous to the corresponding boundary region, m_B can either (top) wrap around the other side of the event horizon, or (bottom) separate into two connected components, one being m_A and the other the event horizon. In the latter case we have $S(B) = S(A) + S_{\rm BH}$, where $S_{\rm BH}$ is the black hole's Bekenstein-Hawking entropy; since $S_{\rm full} = S_{\rm BH}$, the Araki-Lieb inequality is saturated.

Three refinements should be made to (3). First, both sides are divergent; the left-hand side is ultraviolet divergent as discussed above, while the right-hand side is infrared divergent due to the infinite proper distance from any point in the bulk to the conformal boundary. It is easy to see that the latter divergence, like the former, is proportional to the surface area of A. In fact, these two divergences are the same, a manifestation of the usual UV/IR correspondence characteristic of holographic dualities. Therefore (3) is meant to apply in the presence of a UV cutoff in the QFT, corresponding to an IR cutoff in the gravity theory. The simplest such cutoff is a brute force one that cuts off the bulk space at a finite value of the holographic coordinate. The exact choice of cutoff will not be important in what we say below, and for simplicity of presentation we will leave it implicit in the discussion.

Second, there is a complication that occurs when the bulk has nontrivial $(d-1)$st homology (where d is the spatial dimension of the bulk, which is also the spacetime dimension of the boundary). This will be the case, for example, when the bulk contains a black hole. Fursaev's derivation of (3) then tells us that we should minimize a not over all hypersurfaces ending on ∂A but only over those that are homologous to A; that is, there should exist a region r of the bulk such that $\partial r = A \cup m$. See Figs. 1 and 2 for examples. (See [3,12,15] for further discussion.) This rule will be essential in what follows.

Third, formula (3) is exact in the limit that the gravity in the bulk is controlled by the Einstein-Hilbert action. Higher curvature corrections to the bulk action will lead to corrections to the functional $a(m)$. For example, Fursaev [12] showed that, if the bulk action is corrected by a Gauss-Bonnet term, then $a(m)$ is corrected by an Einstein-Hilbert term,

$$a(m) = \int_m \sqrt{h}(1 + 2\alpha R(h)) + 4\alpha \int_{\partial m} \sqrt{\gamma} K, \tag{4}$$

where h is the induced metric on m and α is the coefficient of the Gauss-Bonnet term in the bulk action (see [12] for details). In order to make the variational problem for m well defined, we have also included a Gibbons-Hawking boundary term; γ is the induced metric on ∂m, and K is the trace of its extrinsic curvature (in m).

All of these regions and surfaces—both on the boundary and in the bulk—must lie on a single constant-time slice. In order to have a well-defined notion of "constant-time slice" in the bulk, we must restrict ourselves to states for which the bulk geometry is static. A covariant generalization of (3) to time-dependent geometries will be discussed in [16]. We leave the proof of strong subadditivity in that context to future work.

In paper [4] the authors investigated in a variety of examples whether the formula (3) for the entanglement entropy satisfied the property of strong subadditivity, and in all cases studied it did. Here we will give a simple argument that it does in general.

We begin by rewriting the inequalities (2) in the forms

$$\begin{aligned} S(A) + S(B) &\geq S(A \cup B) + S(A \cap B), \\ S(A) + S(B) &\geq S(A \setminus B) + S(B \setminus A), \end{aligned} \tag{5}$$

where $A \setminus B \equiv A \cap B^c$. We will prove the first inequality; the proof of the second one is very similar and is left as an exercise to the reader.

Let m_A, m_B be the minimal hypersurfaces in the bulk ending on ∂A, ∂B respectively, and r_A, r_B be the corresponding regions of the bulk (so that $\partial r_A = A \cup m_A$, $\partial r_B = B \cup m_B$). (See top of Fig. 3.) We now define the regions $r_{A\cup B} = r_A \cup r_B$, $r_{A\cap B} = r_A \cap r_B$. We can decompose the surfaces of these regions as usual into a part on the bound-

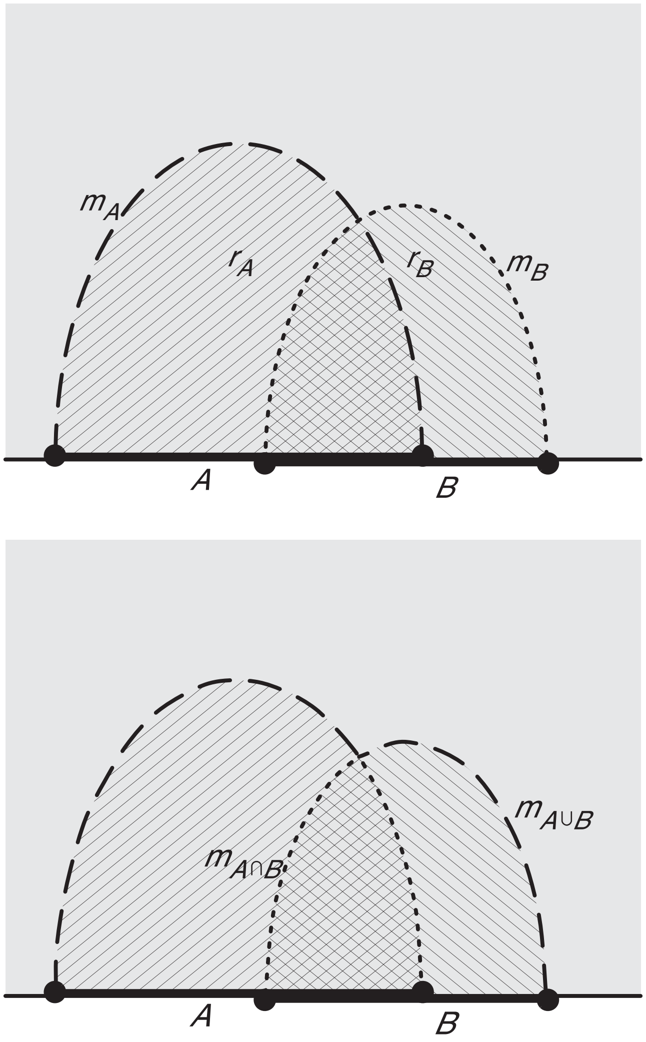

FIG. 3. Two overlapping regions A and B of the boundary, with (top) their respective minimal bulk hypersurfaces m_A, m_B and bulk regions r_A, r_B, and (bottom) their minimal hypersurfaces m_A and m_B cut up and rearranged into two new hypersurfaces $m_{A\cup B}$ (the bulk part of the surface of $r_A \cup r_B$) and $m_{A\cap B}$ (the bulk part of the surface of $r_A \cap r_B$). $m_{A\cup B}$ and $m_{A\cap B}$ end on $\partial(A \cup B)$ and $\partial(A \cap B)$ respectively (although they are not necessarily the minimal such hypersurfaces).

ary and a part in the bulk,

$$\partial r_{A\cup B} = (A \cup B) \cup m_{A\cup B}, \qquad \partial r_{A\cap B} = (A \cap B) \cup m_{A\cap B}. \tag{6}$$

(See bottom of Fig. 3.) Clearly $m_{A\cup B}$ ends on $\partial(A \cup B)$. While nothing says that it is the *minimal* hypersurface ending on $\partial(A \cup B)$, its area is an upper bound on the area of the minimal one, and therefore on $4G_N S(A \cup B)$; similarly for $A \cap B$. Now the hypersurfaces $m_{A\cup B}$ and $m_{A\cap B}$ are simply rearrangements of m_A and m_B (meaning that $m_{A\cup B} \cup m_{A\cap B} = m_A \cup m_B$), so they have the same total areas, [17]

$$a(m_{A\cup B}) + a(m_{A\cap B}) = a(m_A) + a(m_B), \tag{7}$$

which completes the proof.

Note that Eq. (7) holds not just if a is the area, but if it is any *extensive* functional of the hypersurface. This means that if m and m' are two disjoint hypersurfaces with a common boundary, $\partial m \cap \partial m' \neq \varnothing$, then we have $a(m \cup m') = a(m) + a(m')$. This is true, for example, for the Einstein-Hilbert term (with boundary term) added in (4).

In this paper we gave a simple geometric proof of strong subadditivity of entanglement entropy based on the holographic formula (3). The extra dimension in the holographic dual obviously plays an essential role in this proof. Since the strong subadditivity of entanglement entropy should be true in any quantum mechanical many-body system, our result shows that the idea of holography is consistent with any quantum system from this basic viewpoint.

It is interesting to ask when the inequalities (2) are saturated. The only examples we know in the holographic context involve only two disjoint regions, and therefore reduce to the saturation of weak subadditivity, inequalities (1). (It would be interesting to find examples where this is not the case.) The first of these, the Araki-Lieb inequality, is obviously saturated when the full system is pure; then each entanglement entropy is due only to correlations between the subsystems, rather than to the full system being in a mixed state. A system that is in a mixed state but nonetheless saturates the Araki-Lieb inequality is depicted in the bottom panel of Fig. 2. The fact that m_B is disconnected suggests that here the entanglement entropy of B has two separate and unrelated origins: the thermal entropy of the full system (S_{full}), and the correlations between A and B ($S(A)$). On the top panel of that figure, where m_B is connected, the inequality is not saturated.

As for the second inequality in (1), it is saturated (i.e. the mutual information vanishes) when two regions are sufficiently far apart that their union's minimal hypersurface does not connect them but instead is simply the union of their respective minimal hypersurfaces. This was seen in explicit examples in [4]. The mutual information vanishes if and only if the two systems are uncorrelated, i.e. $\rho_{12} = \rho_1 \otimes \rho_2$ [1]. It is interesting that the correlations can go strictly to zero in a field theory (in the large N limit).

Finally, it is useful to notice that our argument can be directly applied to the holographic derivation of a concavity property of coplanar Wilson loops [4], which is closely related to the Bachas inequality [18]. If the curves $C_A = \partial A$ and $C_B = \partial B$ lie in the same two-dimensional plane, then it is clear that the holographically computed expectation values of the corresponding Wilson loops satisfy

$$\langle W(C_A)\rangle\langle W(C_B)\rangle \le \langle W(C_{A\cap B})\rangle\langle W(C_{A\cup B})\rangle,$$
$$\langle W(C_A)\rangle\langle W(C_B)\rangle \le \langle W(C_{A\backslash B})\rangle\langle W(C_{B\backslash A})\rangle,$$

where we defined $C_{A\cap B} = \partial(A \cap B)$, etc. They are equivalent to (5) once we remember that the holographic Wilson loop expectations can also be found from the minimal surface [19]. The evidence from the gauge theory side for these relations will be discussed in [20].

M. H. is supported by the Stanford Institute for Theoretical Physics and by NSF Grant No. PHY 9870115. T. T. would like to thank the Stanford Institute for Theoretical Physics for hospitality, where the present work was initiated. T. T. is also grateful to T. Hirata, V. Hubeny, and M. Rangamani for useful comments on the draft of this paper. The work of T. T. is supported in part by JSPS Grant-in-Aid for Scientific Research No. 18840027.

[1] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information* (Cambridge University Press, Cambridge, U.K., 2000).

[2] E. H. Lieb and M. B. Ruskai, Phys. Rev. Lett. **30**, 434 (1973); J. Math. Phys. (N.Y.) **14**, 1938 (1973), with an appendix by B. Simon.

[3] S. Ryu and T. Takayanagi, Phys. Rev. Lett. **96**, 181602 (2006); J. High Energy Phys. 08 (2006) 045.

[4] T. Hirata and T. Takayanagi, J. High Energy Phys. 02 (2007) 042.

[5] H. Araki and E. H. Lieb, Commun. Math. Phys. **18**, 160 (1970).

[6] Alternative proofs, pedagogical expositions, and reviews can be found in [1,21].

[7] J. Aczél, B. Forte, and C. T. Ng, Adv. Appl. Probab. **6**, 131 (1974); W. Ochs, Rep. Math. Phys. **8**, 109 (1975); A. Wehrl, Rev. Mod. Phys. **50**, 221 (1978).

[8] L. Bombelli, R. K. Koul, J.-H. Lee, and R. D. Sorkin, Phys. Rev. D **34**, 373 (1986).

[9] M. Srednicki, Phys. Rev. Lett. **71**, 666 (1993).

[10] H. Casini and M. Huerta, Phys. Lett. B **600**, 142 (2004).

[11] We use the term "hypersurface" because m is *spatially* codimension 1; in spacetime m is of course codimension 2.

[12] D. V. Fursaev, J. High Energy Phys. 09 (2006) 018.

[13] S. S. Gubser, I. R. Klebanov, and A. M. Polyakov, Phys. Lett. B **428**, 105 (1998); E. Witten, Adv. Theor. Math. Phys. **2**, 253 (1998).

[14] J. M. Maldacena, Adv. Theor. Math. Phys. **2**, 231 (1998).

[15] R. Emparan, J. High Energy Phys. 06 (2006) 012.

[16] V. Hubeny, M. Rangamani, and T. Takayanagi, J. High Energy Phys. 07 (2007) 062.

[17] In this sentence we have assumed the generic situation that m_A and m_B intersect along (spatially) codimension 2 submanifolds. More generally we have $m_{A\cup B} \cup m_{A\cap B} \subset m_A \cup m_B$ and $a(m_{A\cup B}) + a(m_{A\cap B}) \leq a(m_A) + a(m_B)$.

[18] C. Bachas, Phys. Rev. D **33**, 2723 (1986).

[19] J. M. Maldacena, Phys. Rev. Lett. **80**, 4859 (1998); S.-J. Rey and J.-T. Yee, Eur. Phys. J. C **22**, 379 (2001).

[20] T. Hirata (unpublished).

[21] D. Petz, Rep. Math. Phys. **23**, 57 (1986); M. B. Ruskai, arXiv:quant-ph/025064; arXiv:quant-ph/0404126; M. A. Nielsen and D. Petz, Quantum Inf. Comput. **5**, 507 (2005); M. B. Ruskai, arXiv:quant-ph/0604206.