Entropy of dynamical black holes

Stefan Hollands,[1,2,*] Robert M. Wald,[3,†] and Victor G. Zhang[3,‡]
[1]*Institute for Theoretical Physics Leipzig University Brüderstrasse 16, D-04103 Leipzig, Germany*
[2]*Max-Planck-Institute MiS, Inselstrasse 22, D-04103 Leipzig, Germany*
[3]*Enrico Fermi Institute and Department of Physics University of Chicago 933 E. 56th St., Chicago, Illinois 60637 USA*

(Received 29 March 2024; accepted 13 June 2024; published 25 July 2024)

We propose a new formula for the entropy of a dynamical black hole—valid to leading order for perturbations off of a stationary black hole background—in an arbitrary classical diffeomorphism covariant Lagrangian theory of gravity in n dimensions. In stationary eras, this formula agrees with the usual Noether charge formula, but in nonstationary eras, we obtain a nontrivial correction term. In particular, in general relativity, our formula for the entropy of a dynamical black hole differs from the standard Bekenstein-Hawking formula $A/4$ by a term involving the integral of the expansion of the null generators of the horizon. We show that, to leading perturbative order, our dynamical entropy in general relativity is equal to $1/4$ of the area of the apparent horizon. Our formula for entropy in a general theory of gravity is obtained from the requirement that a "local physical process version" of the first law of black hole thermodynamics hold for perturbations of a stationary black hole. It follows immediately that for first order perturbations sourced by external matter that satisfies the null energy condition, our entropy obeys the second law of black hole thermodynamics. For vacuum perturbations, the leading-order change in entropy occurs at second order in perturbation theory, and the second law is obeyed at leading order if and only if the modified canonical energy flux is positive (as is the case in general relativity but presumably would not hold in more general theories of gravity). Our formula for the entropy of a dynamical black hole differs from a formula proposed independently by Dong and by Wall. We obtain the general relationship between their formula and ours. We then consider the generalized second law in semiclassical gravity for first order perturbations of a stationary black hole. We show that the validity of the quantum null energy condition (QNEC) on a Killing horizon is equivalent to the generalized second law using our notion of black hole entropy but using a modified notion of von Neumann entropy for matter. On the other hand, the generalized second law for the Dong-Wall entropy is equivalent to an integrated version of QNEC, using the unmodified von Neumann entropy for the entropy of matter.

DOI: 10.1103/PhysRevD.110.024070

I. INTRODUCTION

The discoveries that black holes obey laws of thermodynamics [1], radiate thermally [2], and should be assigned an entropy [3], provide some of the deepest insights we presently have on the fundamental nature of black holes within a quantum theory of gravity. A formula for the entropy of a stationary black hole in an arbitrary theory of gravity obtained from a diffeomorphism covariant Lagrangian was obtained in [4]. However, the derivation of [4] required evaluation of the entropy on the bifurcation surface, $\mathcal{B}$, of the black hole, thus restricting the validity of the formula for entropy to stationary black holes and their linear perturbations, evaluated at the "time" represented by $\mathcal{B}$. It is of considerable interest to obtain an expression for the entropy of a nonstationary black hole in a general theory of gravity at a "time" represented by an arbitrary cross section $\mathcal{C}$, since this would allow one to investigate whether, classically, black hole entropy satisfies a second law (i.e., whether it is nondecreasing with time) and whether, semiclassically, black hole entropy satisfies a generalized second law (i.e., whether the sum of black hole entropy and a matter contribution to entropy is nondecreasing). A formula for dynamical black hole entropy in a general theory of gravity on an arbitrary cross section $\mathcal{C}$ was proposed in [4]. However, that proposed dynamical entropy formula is not field redefinition

*Contact author: stefan.hollands@uni-leipzig.de
†Contact author: rmwa@uchicago.edu
‡Contact author: victorzhang@uchicago.edu

invariant, and the authors retracted their proposal in a "note in proof" in the published version of [4]. Subsequently, proposed prescriptions for dynamical black hole entropy were given independently by Dong [5] and Wall [6]. The Dong and Wall entropy formulas agree for perturbations of a stationary black hole in the cases where both have been evaluated and are expected to correspond generally at leading order in metric perturbations, although this has not been proven. We will refer to their proposal for dynamical black hole entropy as the Dong-Wall entropy.

The main aim of this paper is to apply a new strategy to the definition of dynamical black hole entropy, based upon the validity of a local, "physical process version" of the first law of black hole mechanics. Since this law is valid only at leading nontrivial order in perturbation theory about a stationary black hole, our resulting definition of dynamical black hole entropy, $S[\mathcal{C}]$, on an arbitrary cross section, $\mathcal{C}$, of the black hole horizon is intended to be applied only at leading nontrivial order in perturbation theory about a stationary black hole. Our formula does not agree with the Dong-Wall entropy, but, as we will elucidate in Sec. VI, there is a close relationship between our formula and theirs. Furthermore, in the case of general relativity, we obtain a nontrivial dynamical correction to the Bekenstein-Hawking entropy formula; namely, we obtain

$$S[\mathcal{C}] = \frac{A[\mathcal{C}]}{4} - \frac{1}{4}\int_{\mathcal{C}} V\vartheta, \tag{1}$$

where $A[\mathcal{C}]$ is the area of the cross section $\mathcal{C}$ [so $A[\mathcal{C}]/4$ is the usual Bekenstein-Hawking entropy], V is an affine parameter of the null generators of the horizon (with $V = 0$ corresponding to the bifurcation surface $\mathcal{B}$), and ϑ is the expansion of these generators with respect to this affine parametrization.

In the remainder of this introductory section, we shall give an outline of our strategy that leads to our new dynamical entropy formula and explain some of the key features of our formula. To begin, we recall that the derivation of [4] of the entropy of a stationary black hole follows from a "fundamental identity" [see (30) below] that holds in an arbitrary theory of gravity obtained from a diffeomorphism covariant Lagrangian. For source-free perturbations of a stationary black hole whose event horizon is a Killing horizon, the fundamental identity reduces to

$$d[\delta\mathbf{Q}[\xi] - \xi\cdot\boldsymbol{\theta}(\phi,\delta\phi)] = 0. \tag{2}$$

Here ξ^a denotes the horizon Killing field, $\mathbf{Q}[\xi]$ is the Noether charge associated with ξ, $\boldsymbol{\theta}$ denotes the symplectic potential, and ϕ collectively denotes all of the dynamical fields (i.e., the metric together with any matter fields appearing in the Lagrangian). Integration of this equation over a spacelike hypersurface Σ that extends from the bifurcation surface $\mathcal{B}$ of the black hole to infinity yields

$$\int_{\mathcal{B}} \delta\mathbf{Q}[\xi] = \delta M - \Omega_{\mathcal{H}}\delta J. \tag{3}$$

Here the right side is the boundary term from infinity resulting from integration of (2) and it involves the mass, M, and angular momentum, J, at spatial infinity (with $\Omega_{\mathcal{H}}$ being angular velocity of the horizon, which enters the expression when one expresses ξ^a in terms of asymptotic time translations and rotations). The left side is the boundary term at $\mathcal{B}$; very importantly, there is no contribution to this term from $\xi\cdot\boldsymbol{\theta}$ because $\xi^a = 0$ on $\mathcal{B}$. We define

$$S \equiv \frac{2\pi}{\kappa}\int_{\mathcal{B}} \mathbf{Q}[\xi], \tag{4}$$

where κ denotes the surface gravity of the black hole (so that $\kappa/2\pi$ is the Hawking temperature of the black hole). It can be shown that S is a local, geometrical quantity [4]. We thereby obtain the first law of black hole mechanics for arbitrary source-free perturbations of a stationary black hole,

$$\frac{\kappa}{2\pi}\delta S = \delta M - \Omega_{\mathcal{H}}\delta J. \tag{5}$$

This formula gives S the interpretation of representing the entropy of the black hole, valid to first order for perturbations of a stationary black hole.

However, as already mentioned above, (4) has the deficiency that the horizon boundary term is evaluated at the bifurcation surface, $\mathcal{B}$, and, thus, S has the interpretation of representing the entropy, $S[\mathcal{B}]$, of the black hole only at the "time" corresponding to $\mathcal{B}$. This deficiency is not of importance for a stationary black hole, since the entropy should be "time independent," i.e., independent of the cross section of the horizon of the black hole at which it is evaluated. However, for a nonstationary black hole, the entropy should evolve with time, and one would like to have an expression for the entropy, $S[\mathcal{C}]$, on an arbitrary cross section, $\mathcal{C}$.

An expression for $S[\mathcal{C}]$ that would satisfy the first law of black hole mechanics would be obtained if we could define $S[\mathcal{C}]$ so that

$$\delta S[\mathcal{C}] = \frac{2\pi}{\kappa}\int_{\mathcal{C}} \delta\mathbf{Q}[\xi] - \xi\cdot\boldsymbol{\theta}(\phi,\delta\phi). \tag{6}$$

If, on the event horizon $\mathcal{H}$, the pullback, $\underline{\boldsymbol{\theta}}$, of the symplectic potential $\boldsymbol{\theta}$ were of the form $\underline{\boldsymbol{\theta}} = \delta\mathbf{B}_{\mathcal{H}}$ for some quantity $\mathbf{B}_{\mathcal{H}}$ defined on $\mathcal{H}$, then the desired quantity $S[\mathcal{C}]$ could be defined by

$$S[\mathcal{C}] = \frac{2\pi}{\kappa}\int_{\mathcal{C}} (\mathbf{Q}[\xi] - \xi\cdot\mathbf{B}_{\mathcal{H}}). \tag{7}$$

However, it is not possible for $\underline{\boldsymbol{\theta}}$ to be of the form of a total variation $\delta\mathbf{B}_{\mathcal{H}}$ on $\mathcal{H}$ in general because such a form would

imply that the symplectic current flux through the horizon would vanish identically (since the symplectic current is the antisymmetrized second variation of $\boldsymbol{\theta}$), which is not the case. Indeed, it was for this reason that [4] defined the entropy via (4) only on the bifurcation surface $\mathcal{B}$.

A similar issue arises for the definition of ADM mass, M, at spatial infinity with respect to an asymptotic time translation t^a. One wishes to define M so that it is the Hamiltonian conjugate to t^a, i.e., so that it generates asymptotic time translations on phase space. This leads to the requirement that under asymptotically flat perturbations of any asymptotically flat background solution, we have [4]

$$\delta M = \int_\infty \delta\mathbf{Q}[t] - t\cdot\boldsymbol{\theta}(\phi,\delta\phi) \tag{8}$$

where "∞" denotes that the integral is taken over a sphere that approaches spatial infinity. However, in this case, as one approaches spatial infinity, the symplectic current goes to zero sufficiently rapidly that the symplectic current flux vanishes. Thus, there is no obstruction to finding a quantity $\mathbf{B}_\infty$ such that to leading asymptotic order

$$\boldsymbol{\theta} = \delta\mathbf{B}_\infty \tag{9}$$

and such a $\mathbf{B}_\infty$ can be explicitly found [4]. One may then define the ADM mass as

$$M = \int_\infty (\mathbf{Q}[t] - t\cdot\mathbf{B}_\infty) \tag{10}$$

and, thus, there is no difficulty in defining a quantity M that satisfies the desired relation (8).

The problem of defining dynamical black hole entropy is much more analogous to the problem of defining the Bondi mass, M_{B}, at null infinity $\mathcal{I}^+$. Here, the symplectic current flux does not, in general, vanish at null infinity, so an analog of (9) cannot hold in general. Nevertheless, it was found in [7] that one could find another symplectic potential $\boldsymbol{\theta}'(\phi,\delta\phi)$ for the pullback of the symplectic current to $\mathcal{I}^+$ with the property that $\boldsymbol{\theta}' = 0$ in a stationary background ϕ. The authors of [7] then defined $\mathbf{B}_{\mathcal{I}^+}$ by

$$\underline{\boldsymbol{\theta}} - \boldsymbol{\theta}' = \delta\mathbf{B}_{\mathcal{I}^+}, \tag{11}$$

where $\underline{\boldsymbol{\theta}}$ denotes the pullback of $\boldsymbol{\theta}$ to $\mathcal{I}^+$. They then defined the Bondi mass relative to an asymptotic time translation t^a on a cross section, $\mathcal{C}$, of $\mathcal{I}^+$ by

$$M_{\mathrm{B}}[\mathcal{C}] = \int_{\mathcal{C}} (\mathbf{Q}[t] - t\cdot\mathbf{B}_{\mathcal{I}^+}). \tag{12}$$

In this case, M_{B} does not satisfy (8) in general—since, in general, there is no solution to this equation—but it does satisfy (8) for perturbations off of a stationary background spacetime.

In this paper, we will apply the strategy of [7] to the definition of dynamical black hole entropy. We will prove that for first-order perturbations of a stationary black hole, there exists a quantity $\mathbf{B}_{\mathcal{H}}$ defined on the black hole horizon that satisfies[1]

$$\underline{\boldsymbol{\theta}} \mathrel{\widehat{=}} \delta\mathbf{B}_{\mathcal{H}}, \tag{13}$$

where $\underline{\boldsymbol{\theta}}$ denotes the pullback of $\boldsymbol{\theta}$ to $\mathcal{H}$ and "$\mathrel{\widehat{=}}$" denotes that the equality holds only on the horizon $\mathcal{H}$. This enables us to define dynamical black hole entropy via (7). For first-order vacuum perturbations of stationary black hole, it follows immediately from integration of (2) over a hypersurface that extends from an arbitrary cross section $\mathcal{C}$ of the horizon to spatial infinity that

$$\frac{\kappa}{2\pi}\delta S[\mathcal{C}] = \delta M - \Omega_{\mathcal{H}}\delta J, \tag{14}$$

i.e., the first law of black hole mechanics holds for an arbitrary cross section. However, it also follows that $\delta S[\mathcal{C}] = \delta S[\mathcal{B}]$, i.e., for vacuum perturbations of a stationary black hole, the entropy is "time independent" to first order and is equal to the entropy of [4]. In order to obtain a nontrivial time variation of black hole entropy, we must either work to second order in perturbation theory in the vacuum case or allow an external stress-energy, δT_{ab}, to be present in the first-order perturbation. As will be discussed in Sec. V B, the results in the vacuum case at second order are closely analogous to the first-order results with an external stress-energy, with the perturbed matter stress-energy δT_{ab} replaced by the second-order modified canonical energy [8] of the gravitational field. For simplicity, we describe only the first-order results with an external stress-energy in the discussion below.

When an external stress-energy δT_{ab} is present, the fundamental identity (2) on the black hole horizon becomes

$$d[\delta\mathbf{Q}[\xi] - \xi\cdot\boldsymbol{\theta}(\phi,\delta\phi)] = -\xi^a\delta\mathbf{C}_a, \tag{15}$$

where, restoring the form indices, $\delta\mathbf{C}_a$ is given by

$$\delta C_{aa_1\cdots a_{n-1}} = \delta T_{ae}\epsilon^e{}_{a_1\cdots a_{n-1}}, \tag{16}$$

where $\epsilon_{a_1\cdots a_n}$ is the spacetime volume n-form. The change in the dynamical entropy with "time" (i.e., its dependence on the cross section $\mathcal{C}$) can be seen by integrating (15) over the portion of the horizon, $\mathcal{H}_{12}$, bounded by cross sections $\mathcal{C}_1$ and $\mathcal{C}_2$, where we take $\mathcal{C}_2$ to lie to the future of $\mathcal{C}_1$. Using

[1]Note that this is equivalent to the Wald-Zoupas requirement for the existence of a $\boldsymbol{\theta}'$ that satisfies (11) and vanishes for a stationary background, since for a nonstationary background, one could define $\boldsymbol{\theta}' = \underline{\boldsymbol{\theta}} - \delta\mathbf{B}_{\mathcal{H}}$.

$\underline{\boldsymbol{\theta}} = \delta \mathbf{B}_{\mathcal{H}}$, we obtain

$$\frac{\kappa}{2\pi}(\delta S[\mathcal{C}_2] - \delta S[\mathcal{C}_1]) = \int_{\mathcal{H}_{12}} -\xi^a \delta \mathbf{C}_a = \int_{\mathcal{H}_{12}} \delta T_{ab} \xi^a k^b \sqrt{h} dV d^{n-2}x, \quad (17)$$

where k^a is the tangent to the affinely parametrized generators of the horizon. Thus, we obtain

$$\frac{\kappa}{2\pi} \Delta\delta S = \Delta\delta E, \quad (18)$$

where $\Delta\delta S$ is the first-order entropy difference between "times" $\mathcal{C}_2$ and $\mathcal{C}_1$, and $\Delta\delta E$ is the first-order matter energy flux (relative to the horizon Killing field ξ^a) through the horizon between these "times". Since $\kappa/2\pi$ is the temperature of the black hole, this has exactly the form of the "physical process" version of the first law of black hole mechanics [9]. Equation (18) is valid for first-order perturbations with matter sources off of a stationary black hole in an arbitrary theory of gravity obtained from a diffeomorphism covariant Lagrangian.

If the matter stress-energy satisfies the null energy condition, then $\delta T_{ab} \xi^a k^b \geq 0$. It then follows immediately from (18) that our definition of entropy satisfies the classical second law $\Delta\delta S \geq 0$ for first-order perturbations. For source-free first-order perturbations, we have $\Delta\delta S = 0$, so the leading-order change in entropy occurs at second order. As previously mentioned, we will show in Sec. V B that the entropy is nondecreasing with time at second order if and only if the modified canonical energy is non-negative at second order. This is the case in general relativity but would not be expected to hold in more general theories of gravity.

For the case of general relativity, the term $-\xi \cdot \mathbf{B}_{\mathcal{H}}$ appearing in (7) gives rise to the "dynamical correction term" to the Bekenstein-Hawking entropy given in (1). This correction term makes $S[\mathcal{C}]$ smaller than the Bekenstein-Hawking entropy. It follows immediately from (18) that, to first order, our entropy increases only when matter crosses the horizon. This contrasts with the Bekenstein-Hawking entropy, which increases before matter is thrown into the black hole since the event horizon is teleologically defined and moves outward in anticipation of matter being thrown in at a later time. We show in Appendix A that, to first order, our entropy (1) is, in fact, the area of the apparent horizon corresponding to "time" $\mathcal{C}$, and thus can be determined locally, without knowledge of the future behavior of the spacetime.

As previously mentioned, our formula for dynamical black hole entropy differs from the Dong-Wall formula. However, we show in Sec. VI that there is a close relationship between our notion of entropy and the Dong-Wall entropy S_{DW}. If, for simplicity, we consider the entropy $S[V]$ evaluated on a cross section of constant affine time V, then we have

$$S[V] = S_{\mathrm{DW}}[V] - V \frac{d}{dV} S_{\mathrm{DW}}[V]. \quad (19)$$

In particular, we have

$$\frac{d}{dV} S = -V \frac{d^2}{dV^2} S_{\mathrm{DW}}. \quad (20)$$

As we shall see in Sec. VII, the fact that the first derivative of our entropy is related to the second derivative of the Dong-Wall entropy will play an important role in the analysis of the generalized second law and its relationship to the quantum null energy condition (QNEC). We will show that with our notion of entropy, the generalized second law is equivalent to QNEC, but with a dynamical correction to the von Neumann entropy. On the other hand, with the Dong-Wall notion of entropy, the generalized second law is equivalent to an integrated version of QNEC, using the unmodified notion of von Neumann entropy of matter.

In Sec. II, we give the necessary background material on the Lagrangian formalism that we will be using throughout the paper and we also define our tetrad choice and specify our gauge conditions on the horizon. In Sec. III, we prove that for perturbations of a stationary black hole whose event horizon is a bifurcate Killing horizon, there exists a quantity $\mathbf{B}_{\mathcal{H}}$ such that $\underline{\boldsymbol{\theta}} \mathrel{\widehat{=}} \delta \mathbf{B}_{\mathcal{H}}$. This enables us to give a general definition of dynamical black hole entropy and elucidate its properties in Sec. IV. We derive the physical process version of the first law of black hole mechanics in Sec. V for both external matter perturbations and vacuum perturbations and discuss its implications for the second law. The relationship of our notion of entropy with the Dong-Wall entropy is analyzed in Sec. VI. The generalized second law and its relationship to QNEC is analyzed in Sec. VII for both our notion of entropy and the Dong-Wall notion.

We will generally follow the conventions of [10]. As we have already done above, we will use boldface to denote differential forms when the form indices are not explicitly written. Pullbacks of differential forms to the event horizon $\mathcal{H}$ will be denoted by underlining them as in (13) above. The covariant spacetime volume element is written as $\epsilon_{a_1 \ldots a_n}$, and our orientation conventions for the event horizon and horizon cross sections are given by (47) and (48) below. We will use the notation "$\mathrel{\widehat{=}}$" to denote that an equality that holds only when both sides are restricted (but not necessarily pulled back) to $\mathcal{H}$, although we will often simply use "=" for equations involving quantities that are only defined on $\mathcal{H}$, where no confusion can arise. We also set $\hbar = 1 = G$.

II. LAGRANGIAN FORMALISM, KILLING HORIZONS, TETRAD CHOICE, AND GAUGE CONDITIONS

In this section, we review some basic definitions and constructions in the Lagrangian formalism, we specify a tetrad choice on the horizon, and we impose some gauge conditions on the horizon.

A. Lagrangians, Noether charge, and the fundamental identity

We consider an arbitrary diffeomorphism covariant theory of gravity in n-dimensions derived from a Lagrangian n-form $\mathbf{L}$ of the form[2]

$$\mathbf{L} = \mathbf{L}(g_{ab}, R_{abcd}, \nabla_{a_1} R_{bcde}, \ldots, \nabla_{(a_1}\cdots\nabla_{a_m)} \times R_{bcde}, \psi, \ldots, \nabla_{(b_1}\cdots\nabla_{a_p)}\psi), \tag{21}$$

where g_{ab} is the spacetime metric, ∇ denotes the derivative operator associated with g_{ab}, R_{abcd} denotes the curvature of g_{ab}, and ψ denotes any matter fields included in the Lagrangian. We shall assume that ψ is a tensor field or fields; spinor fields can be treated via the approach developed in [12]. We collectively refer to the dynamical fields as $\phi = (g_{ab}, \psi)$. It was shown in [4] that any diffeomorphism covariant Lagrangian can be put in the form (21). Beginning in Sec. III, we restrict consideration to the case where matter fields, ψ, are not present in the Lagrangian, but we allow their presence for now.

The first variation of the Lagrangian can always be expressed in the form

$$\delta\mathbf{L} = \mathbf{E}\delta\phi + d\boldsymbol{\theta}, \tag{22}$$

where $\mathbf{E}$ is locally constructed out of ϕ and its derivatives and $\boldsymbol{\theta}(\phi, \delta\phi)$ is locally constructed out of ϕ, $\delta\phi$, and their derivatives and is linear in $\delta\phi$. In (22), contraction of the tensor indices of ϕ with the corresponding dual tensor indices of $\mathbf{E}$ is understood. The Euler-Lagrange equations of motion obtained from $\mathbf{L}$ are $\mathbf{E} = 0$. The $(n-1)$-form $\boldsymbol{\theta}$ plays the role of a *symplectic potential* [4] in that the symplectic current $(n-1)$ form is obtained from it via[3]

$$\boldsymbol{\omega}(\phi, \delta_1\phi, \delta_2\phi) = \delta_1\boldsymbol{\theta}(\phi, \delta_2\phi) - \delta_2\boldsymbol{\theta}(\phi, \delta_1\phi). \tag{23}$$

The symplectic form W is obtained by integrating the symplectic current over a Cauchy surface, Σ, of the spacetime

$$W_\Sigma(\phi, \delta_1\phi, \delta_2\phi) = \int_\Sigma \boldsymbol{\omega}(\phi, \delta_1\phi, \delta_2\phi). \tag{24}$$

Let χ^a be an arbitrary vector field on the spacetime $\mathcal{M}$. Since χ^a is the infinitesimal generator of a diffeomorphism, i.e., a local symmetry of $\mathbf{L}$, there is an associated Noether current $(n-1)$-form $\mathbf{J}$ defined by

$$\mathbf{J}(\phi) = \boldsymbol{\theta}(\phi, \mathcal{L}_\chi\phi) - \chi\cdot\mathbf{L}(\phi), \tag{25}$$

where $\mathcal{L}_\chi$ denotes the Lie derivative with respect to χ^a and we use the notation "$\cdot$" to denote the contraction of a vector field into the first index of a differential form. This definition of $\mathbf{J}$ holds for any field configuration ϕ, i.e., ϕ need not be a solution to the field equations, $\mathbf{E} = 0$. It can be shown that the Noether current can be written in the form [13]

$$\mathbf{J} = d\mathbf{Q}[\chi] + \chi^a\mathbf{C}_a. \tag{26}$$

Here, the $(n-2)$ form $\mathbf{Q}$ is referred to as the "Noether charge" and the dual vector valued $(n-1)$-form $\mathbf{C}_a$ vanishes when the equations of motion are satisfied and is referred to as the constraints of the theory associated with diffeomorphism invariance. The Noether charge takes the general form [4]

$$\mathbf{Q}[\chi] = \mathbf{W}_c(\phi)\chi^c + \mathbf{X}^{cd}(\phi)\nabla_{[c}\chi_{d]} + \mathbf{Y}(\phi, \mathcal{L}_\chi\phi) + d\mathbf{Z}(\phi, \chi), \tag{27}$$

where $\mathbf{W}_c$, $\mathbf{X}^{cd}$, $\mathbf{Y}$, and $\mathbf{Z}$ are covariant quantities, with $\mathbf{W}_c$ and $\mathbf{X}^{cd}$ locally constructed from ϕ, and $\mathbf{Y}$ and $\mathbf{Z}$ locally constructed from the indicated fields, with $\mathbf{Y}$ linear in $\mathcal{L}_\chi\phi$ and $\mathbf{Z}$ linear in χ. For the case where the matter field ψ is a single tensor field $A^{a_1\cdots a_k}{}_{b_1\cdots b_l}$ of type (k, l), the constraints take the explicit form [14]

$$C_{aa_1\cdots a_{n-1}} = \epsilon_{ca_1\cdots a_{n-1}}\Big[2(E_G)^c{}_a - \sum A^{d_1\cdots d_k}{}_{b_1\cdots a\cdots b_l}(E_M)_{d_1\cdots d_k}{}^{b_1\cdots c\cdots b_l} + \sum A^{d_1\cdots c\cdots d_k}{}_{b_1\cdots b_l}(E_M)_{d_1\cdots a\cdots d_s}{}^{b_1\cdots b_l}\Big], \tag{28}$$

where $\epsilon_{a_1\cdots a_n}$ is the spacetime volume form, $(E_G)_{ab} = 0$ are the equations of motion for g^{ab} (obtained by varying the Lagrangian with respect to g^{ab}) and $(E_M)_{a_1\cdots a_k}{}^{b_1\cdots b_l}$ are the equations of motion for the matter field $A^{a_1\cdots a_k}{}_{b_1\cdots b_l}$. The summations in (28) run over the possible index positions of a and c. The generalization of (28) to more than one matter field is straightforward.

Taking the first variation of (25) (taking the vector field χ^a to be fixed) and using (22) and (23), we obtain [4]

[2]Due to the Bianchi identities, not all of the derivatives $\nabla_{(a_1}\cdots\nabla_{a_m)}R_{bcde}$ are algebraically independent at each point. An independent set is given e.g., by $\nabla_{(a_1}\cdots\nabla_{a_m}R^b{}_{cd)}{}^e$ [11].

[3]Here we assume that the field variations $\delta_1\phi$ and $\delta_2\phi$ arise from a two-parameter variation $\phi(\lambda_1, \lambda_2)$ and thus commute. Note that (23) uses the opposite sign convention of [4], but the same sign convention as [8].

$$\delta\mathbf{J}(\phi) = -\chi\cdot(\mathbf{E}(\phi)\delta\phi) + \boldsymbol{\omega}(\phi;\delta\phi,\mathcal{L}_\chi\phi) + d[\chi\cdot\boldsymbol{\theta}(\phi,\delta\phi)], \tag{29}$$

Equating this to the first variation of (26), we obtain

$$\boldsymbol{\omega}(\phi,\delta\phi,\mathcal{L}_\chi\phi) = \chi\cdot(\mathbf{E}(\phi)\delta\phi) + \chi^a\delta\mathbf{C}_a(\phi) + d(\delta\mathbf{Q}[\chi] - \chi\cdot\boldsymbol{\theta}(\phi,\delta\phi)]). \tag{30}$$

Following Hollands and Wald [8], we shall refer to this equation as the *fundamental identity*. The fundamental identity holds for arbitrary ϕ and $\delta\phi$, i.e., they need not satisfy the equations of motion or linearized equations of motion. For the case where χ^a is a Killing field of the background ϕ so that $\mathcal{L}_\chi\phi = 0$, a second variation of (30) yields

$$\boldsymbol{\omega}(\phi,\delta\phi,\mathcal{L}_\chi\delta\phi) = \chi\cdot(\delta\mathbf{E}(\phi)\delta\phi) + \chi\cdot(\mathbf{E}(\phi)\delta^2\phi) + \chi^a\delta^2\mathbf{C}_a(\phi) + d(\delta^2\mathbf{Q}[\chi] - \chi\cdot\delta\boldsymbol{\theta}(\phi,\delta\phi)). \tag{31}$$

B. Tetrad choice on a bifurcate Killing horizon

We are interested in this paper in first- and second-order perturbations of spacetimes that contain a black hole whose event horizon is a bifurcate Killing horizon. By definition, a Killing horizon $\mathcal{H}$ is a null surface to which a Killing field ξ^a is normal. Thus, ξ^a is tangent to the null geodesic generators of $\mathcal{H}$. The surface gravity, κ, of $\mathcal{H}$ is defined by

$$\xi^b\nabla_b\xi^a = \kappa\xi^a \tag{32}$$

so κ measures the failure of Killing parametrization of the null generators to be an affine parametrization. It is necessary for κ to be constant in order that there not be a parallelly propagated curvature singularity on $\mathcal{H}$ [15]. We shall consider here only the case where κ is constant and $\kappa \neq 0$, in which case $\mathcal{H}$ corresponds to a bifurcate Killing horizon [16].

We wish to use a tetrad basis on $\mathcal{H}$ to the future of the bifurcation surface $\mathcal{B}$ that is Lie transported by the Killing field. One way of doing this would be to introduce Gaussian null coordinates, using the Killing parameter v as one of the coordinates (see, e.g., [17]). However, the introduction of Gaussian null coordinates requires an arbitrary choice of a cross section, $\mathcal{C}$, of $\mathcal{H}$, and this would raise questions in our subsequent analysis as to the extent to which our constructions depend upon our choice of $\mathcal{C}$. Instead, we will introduce a tetrad that is constructed only from ξ^a since our constructions will, in any case, depend upon ξ^a.

We do so by introducing a vector field N^a on $\mathcal{H}$ that satisfies,

$$\nabla_a\xi_b \mathrel{\widehat{=}} 2\kappa N_{[a}\xi_{b]}, \tag{33}$$

$$N^aN_a \mathrel{\widehat{=}} 0, \tag{34}$$

where $\mathrel{\widehat{=}}$ denotes equality on $\mathcal{H}$. That there exists a unique N^a satisfying these conditions can be seen as follows. From the hypersurface orthogonality of ξ^a on $\mathcal{H}$ and Frobenius's theorem, there exists a vector field w^a on $\mathcal{H}$ such that

$$\nabla_{[a}\xi_{b]} \mathrel{\widehat{=}} 2w_{[a}\xi_{b]}. \tag{35}$$

Since ξ^a is a Killing field, we have $\nabla_{[a}\xi_{b]} = \nabla_a\xi_b$ and it follows from (32) that

$$\xi^aw_a \mathrel{\widehat{=}} \kappa \tag{36}$$

so, in particular, $w_a \neq 0$. We define

$$N^a \mathrel{\widehat{=}} (\xi^bw_b)^{-1}w^a + c\xi^a = \frac{1}{\kappa}w^a + c\xi^a \tag{37}$$

with

$$c = -\frac{1}{2(\xi^bw_b)^2}w^aw_a = -\frac{1}{2\kappa^2}w^aw_a. \tag{38}$$

Then it is easily seen that N^a satisfies (33) and (34). Uniqueness follows from the fact that (33) uniquely determines N^a up to addition of a multiple of ξ^a at each point and (34) then fixes that multiple. Note that if we contract (33) with ξ^a and use (32), we find that $N^a\xi_a \mathrel{\widehat{=}} 1$. Thus, N^a is a past-directed null vector. Note also that N^a must be invariant under the action of the isometries generated by ξ^a, since this action preserves (33) and (34) but N^a is unique. Thus, we have $\mathcal{L}_\xi N^a \mathrel{\widehat{=}} 0$.

It is useful to complete ξ^a and N^a to a basis of the tangent space on $\mathcal{H}$ by introducing vectors s_i^a $(i = 1,\ldots,n-2)$ on $\mathcal{H}$ such that $s_i^a\xi_a = 0 = s_i^aN_a$, $s_i^as_{ja} = \delta_{ij}$ and $\mathcal{L}_\xi s_i^a|_{\mathcal{H}} = 0$. This can be done by choosing s_i^a on a cross section[4] so as to satisfy these properties and then Lie-transporting s_i^a along ξ^a.

It is important to understand the behavior of our tetrad $\{\xi^a, N^a, s_i^a\}$ as one approaches the bifurcation surface $\mathcal{B}$. This is most conveniently analyzed by introducing Gaussian null coordinates based on a choice of affine parameter V of the null generators of $\mathcal{H}$ (see, e.g., [18])—$V = 0$

[4]Note that there is no reason why there need exist a cross section to which N^a is orthogonal, so, unlike a Gaussian null coordinate basis, s_i^a is not assumed to be tangent to the cross section. It may not be possible to make global choices of s_i^a on a cross section, in which case our construction would have to be done in patches. Different choices of s_i^a will merely differ by rotations in the $(n-2)$-plane orthogonal to ξ^a and N^a, and it will be manifest that our results below will not depend upon the choice of s_i^a.

corresponding to the bifurcation surface $\mathcal{B}$—since such coordinates are smooth at $\mathcal{B}$. Since we require $V = 0$ at $\mathcal{B}$, the choice of V is unique up to $V \to fV$, where f may vary from generator to generator but is constant on each generator. The results below hold for any choice of V. By (32), the relationship between Killing parameter v and affine parameter V is

$$V = e^{\kappa v}, \tag{39}$$

where the (generator-dependent) scaling freedom of V corresponds to the (generator-dependent) freedom in the choice of origin of v. Thus, we have

$$\xi^a = \kappa e^{\kappa v} k^a = \kappa V k^a, \tag{40}$$

where k^a is tangent to the affinely parametrized geodesics and thus is smooth and nonvanishing on $\mathcal{B}$. Thus, we see that $\xi^a = O(V)$ at $V = 0$. Similarly, $N^a = O(V^{-1})$ at $V = 0$, i.e., the components of N^a in Gaussian null coordinates blow up as $1/V$ as $V \to 0$. Finally, the tetrad vectors s_i^a smoothly extend to $\mathcal{B}$. To see this, we note that

$$\mathcal{L}_k s_i^a = \frac{1}{\kappa} e^{-\kappa v} \mathcal{L}_\xi s_i^a + e^{-\kappa v} \xi^a s_i^b \nabla_b v = \kappa k^a s_i^b \nabla_b v. \tag{41}$$

Since $\mathcal{L}_\xi [\kappa s_i^b \nabla_b v] = 0$, the right side is smooth, which implies that s_i^a smoothly extends to $\mathcal{B}$.

The following is an important consequence of the results of the previous paragraph. Suppose that in a spacetime with a bifurcate Killing horizon we have a tensor field $S^{a_1 \cdots a_k}{}_{b_1 \cdots b_l}$ that is Lie derived by ξ^a and is smooth on $\mathcal{H} \backslash \mathcal{B}$. We can expand $S^{a_1 \cdots a_k}{}_{b_1 \cdots b_l}$ on $\mathcal{H} \backslash \mathcal{B}$ in the basis $\{\xi^a, N^a, s_i^a\}$. Since both $S^{a_1 \cdots a_k}{}_{b_1 \cdots b_l}$ and the basis elements are Lie derived by ξ^a, the coefficients appearing in the basis expansion must also be Lie derived by ξ^a. It follows immediately that if any nonvanishing term in the basis expansion has strictly more N^a's than ξ^a's, then $S^{a_1 \cdots a_k}{}_{b_1 \cdots b_l}$ cannot be extended to $\mathcal{B}$. Indeed, in smooth coordinates covering $\mathcal{B}$, some components of $S^{a_1 \cdots a_k}{}_{b_1 \cdots b_l}$ will blow up as $1/V^p$, where p is the maximum of the number of N^a's minus the number of ξ^a's appearing in any individual nonvanishing term in the basis expansion. On the other hand, if all of the nonvanishing terms in the basis expansion of $S^{a_1 \cdots a_k}{}_{b_1 \cdots b_l}$ have at least as many ξ^a's as N^a's, then $S^{a_1 \cdots a_k}{}_{b_1 \cdots b_l}$ can be smoothly extended to $\mathcal{B}$.

C. Gauge conditions on perturbations

In the remainder of this paper, we will be concerned with perturbations, δg_{ab}, of a black hole with a bifurcate Killing horizon. Without loss of generality, we will assume that the true event horizon of the perturbed spacetime coincides with $\mathcal{H}$. We take ξ^a to be fixed under the variation, i.e., we take $\delta \xi^a = 0$. We will impose the following two gauge conditions on δg_{ab} at the event horizon:

$$\xi^a \delta g_{ab} \mathrel{\hat{=}} 0, \tag{42}$$

$$\nabla_a (\xi^b \xi^c \delta g_{bc}) \mathrel{\hat{=}} 0. \tag{43}$$

Both of the above gauge conditions can be imposed without any loss of generality. This is most easily seen from the fact that they automatically hold for any perturbation in Gaussian null coordinates based on the Killing parameter v [see (146) below]. However, we will use only the above two gauge conditions in our analysis below, e.g., we need not assume that the full Gaussian null gauge conditions are imposed.

Equation (42) states that ξ^a remains the null normal to $\mathcal{H}$ under perturbations, since the condition implies that ξ^a remains null and orthogonal to any vector tangent to $\mathcal{H}$. Equation (43) implies that

$$\begin{aligned} \delta(\xi^b \nabla_b \xi^a) = \xi^b \delta \Gamma^a{}_{bc} \xi^c &\mathrel{\hat{=}} -\frac{1}{2} \xi^b \xi^c \nabla^a \delta g_{bc} \\ &\mathrel{\hat{=}} -\frac{1}{2} \nabla^a (\xi^b \xi^c \delta g_{bc}) \mathrel{\hat{=}} 0, \end{aligned} \tag{44}$$

where (42) was used to get the second and third equalities. Comparing with (32), we see that the second condition is equivalent to

$$\delta \kappa = 0, \tag{45}$$

i.e., the surface gravity of ξ^a on $\mathcal{H}$ does not change under the perturbation.

If matter fields ψ are present in the Lagrangian, then the results we will obtain below can be straightforwardly generalized if additional conditions are imposed on their perturbations, $\delta\psi$. In particular, if an electromagnetic field is present, our results can be straightforwardly generalized if the condition $\xi^a \delta A_a \mathrel{\hat{=}} 0$ is imposed, which can always be achieved without loss of generality by an electromagnetic gauge transformation. However, very recently, Wall and Yan [19] extended the definition of the Dong-Wall entropy to Proca fields without imposing additional restrictions on the Proca field. Furthermore, even more recently, Visser and Yan [20] showed that the results of the next section can be generalized to the case where general matter fields are present, without imposing additional restrictions on the perturbations. Thus, we believe that it should be possible to extend our analysis to the case where additional matter fields are present. Nevertheless, for the remainder of this paper, we shall restrict consideration to the case where the metric is the only dynamical field in the Lagrangian, i.e., no additional dynamical fields ψ are present.

III. THE PULLBACK OF THE SYMPLECTIC POTENTIAL IS A TOTAL VARIATION FOR PERTURBATIONS OF A STATIONARY BLACK HOLE

In this section, we consider perturbations of a stationary black hole whose event horizon is a bifurcate Killing horizon. We assume that the perturbations are everywhere smooth and satisfy the gauge conditions (42) and (43). However, we need not assume that the background spacetime metric g_{ab}, is a solution to the field equations, nor do we need to assume that the perturbed metric, δg_{ab} is a solution to the linearized field equations. We will prove that there exists a quantity $\mathbf{B}_{\mathcal{H}}$ defined on the black hole horizon that is locally and covariantly constructed from the metric and[5] ξ^a such that the pullback, $\underline{\boldsymbol{\theta}}$, to $\mathcal{H}$ of the symplectic potential $\boldsymbol{\theta}(g_{ab}, \delta g_{ab})$ is given by

$$\underline{\boldsymbol{\theta}} \widehat{=} \delta \mathbf{B}_{\mathcal{H}}. \tag{46}$$

As explained in the Introduction, (46) will enable us to define dynamical black hole entropy, and we shall then do so in the next section. As stated at the end of the previous section, we restrict consideration here to the vacuum case, but very recently, Visser and Yan [20] showed that (46) holds in the nonvacuum case as well. Their expression for $\mathbf{B}_{\mathcal{H}}$ is not covariant in the metric and ξ^a, but a covariant expression can be obtained using methods similar to those used in the proof of Theorem 1 below.

The horizon $\mathcal{H}$ is an $(n-1)$-dimensional null surface. Since the pullback of the spacetime metric to $\mathcal{H}$ is degenerate, it does not induce a unique volume form on $\mathcal{H}$. However, we can define a volume form $\boldsymbol{\epsilon}^{(n-1)}_{a_1\cdots a_{n-1}}$ on $\mathcal{H}$ that is Lie derived by ξ^a in the background spacetime by

$$-n\xi_{[a_1}\epsilon^{(n-1)}_{a_2\cdots a_n]} = \epsilon_{a_1 a_2 \cdots a_n}, \tag{47}$$

where $\epsilon_{a_1 a_2 \cdots a_n}$ is the spacetime volume form. It is also convenient to define an $(n-2)$-form $\epsilon^{(n-2)}_{a_1\cdots a_{n-2}}$ on $\mathcal{H}$ by

$$\epsilon^{(n-2)}_{a_1\cdots a_{n-2}} = \xi^c \epsilon^{(n-1)}_{c a_1 \cdots a_{n-2}}. \tag{48}$$

The pullback of $\boldsymbol{\epsilon}^{(n-2)}$ to any cross section, $\mathcal{C}$, of $\mathcal{H}$ will then agree with the volume element on $\mathcal{C}$ obtained from the nondegenerate metric on $\mathcal{C}$ obtained from the pullback of the spacetime metric.

The symplectic potential $\boldsymbol{\theta}$ is an $(n-1)$-form on spacetime, so its pullback $\underline{\boldsymbol{\theta}}$ to $\mathcal{H}$ must be proportional to $\epsilon^{(n-1)}_{a_1\cdots a_{n-1}}$. It follows from (47) that we have

$$\underline{\boldsymbol{\theta}} \widehat{=} \alpha^a \xi_a \boldsymbol{\epsilon}^{(n-1)}, \tag{49}$$

where we defined

$$\alpha^a \equiv \frac{1}{(n-1)!}\theta_{a_1\cdots a_{n-1}}\epsilon^{a a_1 \cdots a_{n-1}}. \tag{50}$$

Our results on the form of $\underline{\boldsymbol{\theta}}$ will be a direct consequence of the following theorem.

Theorem 1. Let $(\mathcal{M}, g_{ab})$ be an n-dimensional spacetime with a bifurcate Killing horizon $\mathcal{H}$ with horizon Killing field ξ^a. Let δg_{ab} be an arbitrary smooth perturbation of g_{ab} satisfying the gauge conditions (42) and (43). Let α^a be a vector field on $\mathcal{M}$ of the form

$$\alpha^a = \sum_{i=0}^{k} T^{a b_1 \cdots b_i c d}_{(i)} \nabla_{(b_1} \cdots \nabla_{b_i)} \delta g_{cd}, \tag{51}$$

where the tensors $T^{a b_1 \cdots b_i c d}_{(i)} = T^{a (b_1 \cdots b_i)(cd)}_{(i)}$ are smooth and are locally and covariantly constructed from the metric, curvature, and covariant derivatives of the curvature. Then on the horizon, the scalar function $\alpha_c \xi^c$ takes the form

$$\alpha_c \xi^c \widehat{=} \sum_{i=0}^{k-1} \tilde{T}^{b_1 \cdots b_i c d}_{(i)} \nabla_{(b_1} \cdots \nabla_{b_i)} \mathcal{L}_\xi \delta g_{cd}, \tag{52}$$

where the tensors $\tilde{T}^{b_1 \cdots b_i c d}_{(i)}$ are smooth on $\mathcal{H}$ and are locally and covariantly constructed from the metric, curvature, and its derivatives as well as from ξ^a and N^a [see (33) and (34)], with ξ^a and N^a appearing only algebraically.

Proof. To begin, we focus on the tensor field $T^{a b_1 \cdots b_i c d}_{(k)}$ on $\mathcal{H}$ appearing in the highest derivative term in the expansion (51) of α^a. Our aim is to show that $\xi_a T^{a b_1 \cdots b_i c d}_{(k)} \nabla_{(b_1} \cdots \nabla_{b_k)} \delta g_{cd}$ can be written in the form $\tilde{T}^{b_1 \cdots b_{k-1} c d}_{(k-1)} \nabla_{(b_1} \cdots \nabla_{b_{k-1})} \mathcal{L}_\xi \delta g_{cd}$ plus lower derivative terms of essentially the same character as appeared in the original expression for $\xi_a \alpha^a$. This will enable us to make an inductive argument to prove the theorem.

By the hypothesis of the theorem, the tensor field $T^{a b_1 \cdots b_i c d}_{(k)}$ is smooth on $\mathcal{H}$, including at the bifurcation surface $\mathcal{B}$. Since ξ^a is a Killing field of the background spacetime, $T^{a b_1 \cdots b_k c d}_{(k)}$ is Lie derived by ξ^a. Therefore, as shown in the last paragraph of Sec. II B, if we expand it in the basis $\{\xi^a, N^a, s^a_i\}$, the only nonvanishing terms in the basis expansion will have at least as many factors of ξ^a as N^a. If we then contract α^a with ξ^a, we will eliminate one factor of N^a from each term in the basis expansion, so each nonvanishing term will have at least one more factor of ξ^a than N^a. More explicitly, the basis expansion of $\xi_a T^{a b_1 \cdots b_i c d}_{(k)}$ takes the form

[5]The vector field N^a on $\mathcal{H}$ will also appear in our expression for $\mathbf{B}_{\mathcal{H}}$. However, N^a on $\mathcal{H}$ is locally and covariantly determined from the metric and ξ^a via (33) and (34).

$$\begin{aligned}\xi_a T^{ab_1\cdots b_k cd}_{(k)} \hat{=}\ & c_1 \xi^{b_1}\xi^{b_2}\cdots\xi^{b_k}\xi^c\xi^d \\ & + c_2 N^{(b_1}\xi^{b_2}\cdots\xi^{b_k)}\xi^c\xi^d + c_3\xi^{b_1}\xi^{b_2}\cdots\xi^{b_k}N^{(c}\xi^{d)} \\ & + c_4 s_{i_1}^{(b_1}\xi^{b_2}\cdots\xi^{b_k)}\xi^c\xi^d + c_5\xi^{b_1}\xi^{b_2}\cdots\xi^{b_k}s_i^{(c}\xi^{d)} \\ & + \cdots \\ & + c_6 s_{i_1}^{(b_1}s_{i_2}^{b_2}\cdots s_{i_{k-1}}^{b_{k-1}}\xi^{b_k)}s_{j_1}^{(c}s_{j_2}^{d)} + c_7 s_{i_1}^{b_1}s_{i_2}^{b_2}\cdots s_{i_k}^{b_k}\xi^{(c}s_j^{d)},\end{aligned} \tag{53}$$

where $c_1, c_2, \ldots, c_7$ are Lie derived by ξ^a. Here, the first term in the basis expansion is the term with all ξ^a's. In the second line, we replace one of the ξ^a's with an N^a. Since $T^{ab_1\cdots b_i cd}_{(k)}$ is symmetric in $b_1 \ldots b_k$ as well as in cd, there are two distinct terms that we can get via such a replacement, namely, one term where the N^a replaces a ξ^a on a b_i index and one term where it replaces a ξ^a on a c or d index. In the third line, we similarly replace a ξ^a by an s_i^a. (Since i ranges from 1 to $n-2$, there are $n-2$ distinct expressions of the form appearing in the third line.) The "…" in the fourth line represents all of the terms we get by continuing to replace ξ^a's by N^a's and s_i^a's, thereby continually decreasing the number of factors of ξ^a. The key point is that each term must have at least one more ξ^a than N^a. Thus, the final two terms in the basis expansion—shown on the fifth line—have $(k+1)$ factors of s_i^a and one factor of ξ^a.

We now consider each term in the basis expansion separately. Since at least one ξ^a appears in each term in the basis expansion, there are three possible cases, which we will consider separately:

(i) *Case (1):* There are no ξ^{b_i} factors, i.e., there are no factors of ξ^a that contract into $\nabla_{(b_1}\cdots\nabla_{b_k)}$ in (51). Furthermore, there is only one ξ^c or ξ^d factor, i.e., only one index of δg_{cd} contracts into factor of ξ^a.

(ii) *Case (2):* As in case (1), there are no ξ^{b_i} factors. However, there are ξ^c and ξ^d factors, i.e., both indices of δg_{cd} contract into factors of ξ^a.

(iii) *Case (3):* There is at least one ξ^{b_i} factor, i.e., there is at least one contraction of ξ^a with a derivative index in (51).

Case (1) is the most straightforward to analyze. Since only one factor of ξ^a occurs in this case, the only terms in the basis expansion (53) that meet the criterion of case (1) take the form of the very last term in (53). Thus, to analyze this case, we need to consider the quantity

$$s_{i_1}^{b_1}s_{i_2}^{b_2}\cdots s_{i_k}^{b_k}\xi^c s_j^d\nabla_{(b_1}\cdots\nabla_{b_k)}\delta g_{cd}, \tag{54}$$

where we dropped the symmetrization over cd in the basis term since δg_{cd} is symmetric in c and d. We now move ξ^c, through all of the derivatives $\nabla_{(b_1}\cdots\nabla_{b_k)}$ to write it as

$$\begin{aligned}& s_{i_1}^{b_1}s_{i_2}^{b_2}\cdots s_{i_k}^{b_k}\xi^c s_j^d\nabla_{(b_1}\cdots\nabla_{b_k)}\delta g_{cd} \\ & = s_{i_1}^{b_1}s_{i_2}^{b_2}\cdots s_{i_k}^{b_k}s_j^d\nabla_{(b_1}\cdots\nabla_{b_k)}(\delta g_{cd}\xi^c) \\ & \quad + \text{lower derivative terms},\end{aligned} \tag{55}$$

where the "lower derivative terms" have one or more derivatives ∇_{b_i} acting only on ξ^c and thus have no more than $k-1$ derivatives remaining to act on δg_{cd}. However, by our gauge condition (42), we have $\delta g_{cd}\xi^c \hat{=} 0$. Since derivatives of this quantity in directions tangential to the horizon must also vanish, we see that the terms occurring in case (1) are equivalent to an expression with strictly fewer than k derivatives acting on δg_{cd}.

The analysis of case (2) is similar to case (1). We move both ξ^c and ξ^d through $\nabla_{(b_1}\cdots\nabla_{b_k)}$ to obtain a term where all k derivatives act on $\delta g_{cd}\xi^c\xi^d$ plus "lower derivative terms," i.e., terms where at least one derivative acts on ξ^c or ξ^d and thus no more than $k-1$ derivatives act on δg_{cd}. However, the term $\nabla_{(b_1}\cdots\nabla_{b_k)}(\delta g_{cd}\xi^c\xi^d)$ can have at most one factor N^{b_i} contracting into it, since there are precisely two factors of ξ^a in the terms that arise in case (2) and we must have strictly more factors of ξ^a than N^a. Therefore, in the term $\nabla_{(b_1}\cdots\nabla_{b_k)}(\delta g_{cd}\xi^c\xi^d)$, there is at most one derivative of $\delta g_{cd}\xi^c\xi^d$ that is taken in a direction that is not tangential to the horizon. Consequently, this term vanishes by our gauge condition $\nabla_b(\delta g_{cd}\xi^c\xi^d)\hat{=}0$. Therefore, we are left only with the lower derivative terms, i.e., terms with $k-1$ or fewer derivatives acting on δg_{cd}.

Finally, to treat case (3), we must analyze terms of the form

$$\xi^{b_i}\nabla_{b_1}\cdots\nabla_{b_i}\cdots\nabla_{b_k}\delta g_{cd}. \tag{56}$$

To do so, we first bring ξ^{b_i} through $\nabla_{b_1}\cdots\nabla_{b_{i-1}}$ to rewrite this term as

$$\begin{aligned}& \xi^{b_i}\nabla_{b_1}\cdots\nabla_{b_i}\cdots\nabla_{b_k}\delta g_{cd} \\ & = \nabla_{b_1}\cdots\nabla_{b_{i-1}}[\xi^{b_i}\nabla_{b_i}(\nabla_{b_{i+1}}\cdots\nabla_{b_k}\delta g_{cd})] \\ & \quad + \text{lower derivative terms}.\end{aligned} \tag{57}$$

We then write

$$\xi^{b_i}\nabla_{b_i}(\nabla_{b_{i+1}}\cdots\nabla_{b_k}\delta g_{cd}) = \mathcal{L}_\xi(\nabla_{b_{i+1}}\cdots\nabla_{b_k}\delta g_{cd}) + \text{lower derivative terms.} \tag{58}$$

Finally, we use the fact that since ξ^a is a Killing field of the background metric, $\mathcal{L}_\xi$ commutes with the background derivative operator ∇_b, so we have

$$\mathcal{L}_\xi(\nabla_{b_{i+1}}\cdots\nabla_{b_k}\delta g_{cd}) = \nabla_{b_{i+1}}\cdots\nabla_{b_k}\mathcal{L}_\xi\delta g_{cd}. \tag{59}$$

Putting this all together, we have shown that

$$\xi^{b_i}\nabla_{b_1}\cdots\nabla_{b_i}\cdots\nabla_{b_k}\delta g_{cd} = \nabla_{b_1}\cdots\nabla_{b_{i-1}}\nabla_{b_{i+1}}\cdots\nabla_{b_k}\mathcal{L}_\xi\delta g_{cd} + \text{lower derivative terms.} \tag{60}$$

Combining all of the results of the analyses of cases (1), (2), and (3) above, we see that the highest derivative term in $\xi_a\alpha^a$ can be written in the form

$$\xi_a T_{(k)}^{ab_1\cdots b_k cd}\nabla_{(b_1}\cdots\nabla_{b_k)}\delta g_{cd} \hat{=} \tilde{T}_{(k-1)}^{b_1\cdots b_{k-1}cd}\nabla_{(b_1}\cdots\nabla_{b_{k-1})}\mathcal{L}_\xi\delta g_{cd} + \sum_{i=0}^{k-1} T'^{b_1\cdots b_i cd}_{(i)}\nabla_{(b_1}\cdots\nabla_{b_i)}\delta g_{cd} \tag{61}$$

Here $\tilde{T}_{(k-1)}^{b_1\cdots b_{k-1}cd}$ arises entirely from case (3), wherein one ξ^b in terms in the basis expansion of $\xi_a T_{(k)}^{ab_1\cdots b_k cd}$ has been combined with a ∇_b and converted to a $\mathcal{L}_\xi$. It follows that each term in basis expansion of $\tilde{T}_{(k-1)}^{b_1\cdots b_{k-1}cd}$ has at least as many ξ^a's as N^a's and thus is smooth on $\mathcal{H}$. The components of $\tilde{T}_{(k-1)}^{b_1\cdots b_{k-1}cd}$ are obtained algebraically from $\xi_a T_{(k)}^{ab_1\cdots b_k cd}$ together with the basis elements N^a, ξ^a and $\{s_i^a\}$. However, it is clear that $\tilde{T}_{(k-1)}^{b_1\cdots b_{k-1}cd}$ does not depend on the choice of $\{s_i^a\}$, as can be seen from the fact that we could have worked with the projector

$$q^{ab} = g^{ab} - 2\xi^{(a}N^{b)} \tag{62}$$

rather than choosing the basis $\{s_i^a\}$. Since $\xi_a T_{(k)}^{ab_1\cdots b_k cd}$ is locally and covariantly constructed from the metric, curvature, and its derivatives, we see that $\tilde{T}_{(k-1)}^{b_1\cdots b_{k-1}cd}$ is locally and covariantly constructed from the metric, curvature, and its derivatives together with ξ^a and N^a, with only algebraic dependence on ξ^a and N^a.

On the other hand $T'^{b_1\cdots b_i cd}_{(i)}$ is algebraically constructed from not only $\xi_a T_{(k)}^{ab_1\cdots b_k cd}$, ξ^a, and N^a but also from derivatives of ξ^a. However, we can use the Killing field identity,

$$\nabla_a\nabla_b\xi_c = R_{cbad}\xi^d, \tag{63}$$

to eliminate second derivatives of ξ^a in terms of ξ^a and curvature. By successively using this identity, we can rewrite all factors involving an arbitrary number of derivatives of ξ^a in terms of curvature, derivatives of curvature, ξ^a, and $\nabla_b\xi^a$. We may then use

$$\nabla_b\xi_a \hat{=} 2\kappa N_{[b}\xi_{a]} \tag{64}$$

to eliminate the first derivative of ξ^a in terms of ξ^a and N^a. In this way, we see that $T'^{b_1\cdots b_i cd}_{(i)}$ is locally and covariantly constructed from the metric, curvature, derivatives of curvature, ξ^a and N^a, with only algebraic dependence on ξ^a and N^a.

Next, we note that since ξ^a is a Killing field of the background spacetime, we have

$$\tilde{T}_{(k-1)}^{b_1\cdots b_{k-1}cd}\nabla_{(b_1}\cdots\nabla_{b_{k-1})}\mathcal{L}_\xi\delta g_{cd} = \mathcal{L}_\xi[\tilde{T}_{(k-1)}^{b_1\cdots b_{k-1}cd}\nabla_{(b_1}\cdots\nabla_{b_{k-1})}\delta g_{cd}] \tag{65}$$

$$= \xi^a\nabla_a[\tilde{T}_{(k-1)}^{b_1\cdots b_{k-1}cd}\nabla_{(b_1}\cdots\nabla_{b_{k-1})}\delta g_{cd}]. \tag{66}$$

Thus, we may rewrite (61) as

$$\sum_{i=0}^{k-1} T'^{b_1\cdots b_i cd}_{(i)}\nabla_{(b_1}\cdots\nabla_{b_i)}\delta g_{cd} = \xi_a T_{(k)}^{ab_1\cdots b_k cd}\nabla_{(b_1}\cdots\nabla_{b_k)}\delta g_{cd} - \xi^a\nabla_a[\tilde{T}_{(k-1)}^{b_1\cdots b_{k-1}cd}\nabla_{(b_1}\cdots\nabla_{b_{k-1})}\delta g_{cd}]. \tag{67}$$

The right side of (67) is smooth and vanishes on the bifurcation surface on account of the explicit factor of ξ^a in both terms. Thus, the left side of (67) has exactly the same properties as $\xi_a\alpha^a$ except that (i) there now is additional algebraic dependence on ξ^a and N^a and (ii) the highest derivative term has only $k-1$ derivatives of δg_{cd}. The first point is of no consequence for any of the arguments that led from (53) to (61). The second point enables us to prove the theorem by the following inductive argument.

We take as an inductive hypothesis in k that a slight strengthening of the statement of the theorem holds; namely, we allow the tensors appearing in (51) to depend algebraically on ξ^a and N^a as well as on the metric, curvature, and covariant derivatives of the curvature. We assume that the inductive hypothesis holds for $k = p-1$. For $k = p$, the above arguments show that the highest derivative term in $\xi_a\alpha^a$ takes the form (61) and that the lower derivative terms are such that the $k = p-1$ hypothesis can be applied to them. Therefore, the inductive hypothesis for $k = p-1$ implies the inductive hypothesis for $k = p$, and all that remains to show is that the inductive hypothesis holds for $k = 0$. However, this is just a simple

special case of cases (1) and (2) in the above proof of the proposition, and it follows immediately from our gauge conditions (42) and (43) that $\xi_a \alpha^a = 0$ when $k = 0$. Thus, our inductive hypothesis holds for $k = 0$ and the proof is completed. ∎

The main result of this section can now be obtained as follows. For α^a defined by (50), the general form of $\boldsymbol{\theta}$ given by Eq. (23) of [4] shows that (51) holds with $k = m + 1$, where m is the maximum number of derivatives of the curvature appearing in the Lagrangian (21). Therefore, by (49) and the above theorem, we have

$$\underline{\boldsymbol{\theta}} \widehat{=} \boldsymbol{\epsilon}^{(n-1)} \sum_{i=0}^{m} \tilde{T}_{(i)}^{b_1 \cdots b_i cd} \nabla_{(b_1} \cdots \nabla_{b_i)} \mathcal{L}_\xi \delta g_{cd}. \tag{68}$$

Define $\mathbf{B}_\mathcal{H}$ on $\mathcal{H}$ by

$$\mathbf{B}_\mathcal{H} = \boldsymbol{\epsilon}^{(n-1)} \sum_{i=0}^{m} \tilde{T}_{(i)}^{b_1 \cdots b_i cd} \nabla_{(b_1} \cdots \nabla_{b_i)} \mathcal{L}_\xi g_{cd}. \tag{69}$$

Then we have

$$\underline{\boldsymbol{\theta}} \widehat{=} \delta \mathbf{B}_\mathcal{H}. \tag{70}$$

Namely, when we take a first variation of $\mathbf{B}_\mathcal{H}$, we will get a factor of $\mathcal{L}_\xi g_{cd} = 0$ appearing in all terms except the terms where the variation acts on this factor. However, these terms manifestly yield $\underline{\boldsymbol{\theta}}$.

The quantities $\tilde{T}_{(i)}^{b_1 \cdots b_i cd}$ needed to obtain $\mathbf{B}_\mathcal{H}$ can be computed by following the steps used in the proof of the theorem. We will compute $\mathbf{B}_\mathcal{H}$ explicitly for general relativity and for theories with Lagrangians that are quadratic in the curvature tensor in the next section. We will also discuss the ambiguities in the definition of $\mathbf{B}_\mathcal{H}$ in the next section.

IV. DYNAMICAL BLACK HOLE ENTROPY

A. General definition of dynamical black hole entropy

In the previous section, we have proven that for a stationary black hole whose event horizon, $\mathcal{H}$, is a bifurcate Killing horizon there exists an $(n-1)$-form $\mathbf{B}_\mathcal{H}$ on $\mathcal{H}$ with the property that (70) holds for arbitrary perturbations. Following the motivation given in the Introduction, we define the *entropy $(n-2)$-form* $\mathbf{S}$ on $\mathcal{H}$ by

$$\mathbf{S} \equiv \frac{2\pi}{\kappa} (\mathbf{Q}[\xi] - \xi \cdot \mathbf{B}_\mathcal{H}). \tag{71}$$

We define the entropy, $S[\mathcal{C}]$, of the black hole at a "time" represented by an arbitrary cross section $\mathcal{C}$ by

$$S[\mathcal{C}] \equiv \int_\mathcal{C} \mathbf{S} = \frac{2\pi}{\kappa} \int_\mathcal{C} (\mathbf{Q}[\xi] - \xi \cdot \mathbf{B}_\mathcal{H}). \tag{72}$$

Since the definition of $\mathbf{S}$ does not depend on a choice of cross section and $\mathbf{S}$ is smooth on $\mathcal{H}$, it follows immediately that $S[\mathcal{C}]$ varies continuously with the cross section $\mathcal{C}$ in the sense defined in [21]. We now consider the ambiguities in our definition of $S[\mathcal{C}]$ for stationary black holes, for first-order perturbations of stationary black holes, and for second- and higher-order deviations from a stationary black hole.

For the unperturbed stationary black hole, we have $\mathbf{B}_\mathcal{H} = 0$ by (69), so we have

$$S[\mathcal{C}] = \frac{2\pi}{\kappa} \int_\mathcal{C} \mathbf{Q}[\xi]. \tag{73}$$

Thus, the only possible ambiguities in $S[\mathcal{C}]$ for stationary black holes arise from ambiguities in the definition of $\mathbf{Q}$, which are given by [4]

$$\mathbf{Q}[\xi] \to \mathbf{Q}[\xi] + \xi \cdot \boldsymbol{\mu}(\phi) + \mathbf{Y}(\phi, \mathcal{L}_\xi \phi) + d\mathbf{Z}(\phi, \xi). \tag{74}$$

Here the first two terms result from the ambiguity in $\boldsymbol{\theta}$

$$\boldsymbol{\theta}(\phi, \delta\phi) \to \boldsymbol{\theta}(\phi, \delta\phi) + \delta\boldsymbol{\mu}(\phi) + d\mathbf{Y}(\phi, \delta\phi) \tag{75}$$

arising from adding an exact term $d\boldsymbol{\mu}$ to the Lagrangian and an exact term $d\mathbf{Y}$ to $\boldsymbol{\theta}$ (see Ref. [4]). Since the $(n-1)$-form $\boldsymbol{\mu}$ is locally and covariantly constructed from the metric, we have $\mathcal{L}_\xi \boldsymbol{\mu} = 0$ in the stationary background. Since $\boldsymbol{\mu}$ smoothly extends to $\mathcal{B}$, it then follows from the type of argument given at the end of Sec. II B that the pullback of $\boldsymbol{\mu}$ to $\mathcal{H}$ vanishes in the stationary background. The $\mathbf{Y}$ term in (74) vanishes in the stationary background since $\mathcal{L}_\xi \phi = 0$. Finally, $d\mathbf{Z}$ integrates to zero over any cross section $\mathcal{C}$. Thus, $S[\mathcal{C}]$ is completely unambiguous in the stationary background. Furthermore, if the equations of motion hold in the stationary background, then $d\mathbf{Q} = \mathbf{J}$ by (26). However, (25) implies that the pullback of $\mathbf{J}$ to $\mathcal{H}$ vanishes, because $\boldsymbol{\theta}(\phi, \mathcal{L}_\xi \phi) = 0$ (since ξ^a is a Killing field) and the pullback of $\xi \cdot \mathbf{L}$ to $\mathcal{H}$ vanishes (since ξ^a is tangent to $\mathcal{H}$). Thus, we find that the pullback of $d\mathbf{Q}$ to $\mathcal{H}$ vanishes. This implies that $S[\mathcal{C}]$ does not depend on $\mathcal{C}$ for a stationary black hole, i.e., the entropy is "time independent". In particular, we may evaluate $S[\mathcal{C}]$ on the bifurcation surface $\mathcal{B}$, in which case it reduces to the definition given in [4,22]. Thus, our definition (72) reduces to the standard expression for entropy for stationary black holes on any cross section $\mathcal{C}$.

For first-order perturbations of a stationary black hole, we have

$$\delta S[\mathcal{C}] = \frac{2\pi}{\kappa} \int_\mathcal{C} (\delta\mathbf{Q}[\xi] - \xi \cdot \boldsymbol{\theta}), \tag{76}$$

where (70) was used to replace $\delta\mathbf{B}_\mathcal{H}$ by $\boldsymbol{\theta}$. The only possible ambiguities in this expression result from the ambiguities (74) and (75). Thus, the ambiguity in $\delta\mathbf{S}$ is

$$\delta\mathbf{S} \to \delta\mathbf{S} + \frac{2\pi}{\kappa}[\xi\cdot\delta\boldsymbol{\mu} + \mathbf{Y}(\phi, \mathcal{L}_\xi\delta\phi) + d\delta\mathbf{Z}(\phi,\xi) - \xi\cdot\delta\boldsymbol{\mu} - \xi\cdot d\mathbf{Y}(\phi,\delta\phi)] \tag{77}$$

$$= \delta\mathbf{S} + \frac{2\pi}{\kappa}[\mathcal{L}_\xi\mathbf{Y}(\phi,\delta\phi) + d\delta\mathbf{Z}(\phi,\xi) - \xi\cdot d\mathbf{Y}(\phi,\delta\phi)] \tag{78}$$

$$= \delta\mathbf{S} + \frac{2\pi}{\kappa}[d(\xi\cdot\mathbf{Y})(\phi,\delta\phi) + d\delta\mathbf{Z}(\phi,\xi)] \tag{79}$$

where the general identity

$$\mathcal{L}_\xi\boldsymbol{\alpha} = \xi\cdot d\boldsymbol{\alpha} + d(\xi\cdot\boldsymbol{\alpha}) \tag{80}$$

for any form $\boldsymbol{\alpha}$ was used in the last line. The exact terms in (79) integrate to zero on any cross section. Thus, for arbitrary first-order perturbations, $\delta S[\mathcal{C}]$ is entirely unambiguous. As we will see explicitly below, in the case of general relativity our formula for $\delta S[\mathcal{C}]$ does not agree with the Bekenstein-Hawking entropy formula in nonstationary eras.

For second-order perturbations, there will, in general, be substantial additional ambiguities resulting from ambiguities in defining $\mathbf{B}_\mathcal{H}$ at second order. As we shall now explain, there are two sources of these ambiguities.

The first source has to do with the essential presence of ξ^a in the formula for $\mathbf{B}_\mathcal{H}$. At first order, the perturbed metric enters the formula for $\mathbf{B}_\mathcal{H}$ only via the factor $\mathcal{L}_\xi\delta g_{ab}$, so all of the other quantities in that formula take their background values. Since we require $\delta\xi^a = 0$, ξ^a is rigidly fixed on spacetime to be the horizon Killing field of the background metric g_{ab}. Similarly, N^a on $\mathcal{H}$ is rigidly fixed by (33) and (34). However, the varied metric will not, in general, admit any Killing fields at all, and, at second and higher orders in perturbation theory, the formula for $\mathbf{B}_\mathcal{H}$ will depend upon how ξ^a is chosen. We can (and will) require as a gauge choice that ξ^a is a fixed vector field on spacetime, but unless g_{ab} is rigidly tied to ξ^a, the formula for $\mathbf{B}_\mathcal{H}$ will have a gauge dependence with regard to the metric. In order to avoid this, we must provide a prescription to rigidly fix ξ^a in terms of the metric in a situation where the metric no longer possesses a horizon Killing field. Without loss of generality, we may assume that the true event horizon of the perturbed spacetime coincides with $\mathcal{H}$, that ξ^a remains normal to $\mathcal{H}$, and that $\xi^b\nabla_b\xi^a = \kappa\xi^a$ on $\mathcal{H}$. This rigidly fixes ξ^a on $\mathcal{H}$ up to a choice of cross section $\mathcal{B}$ of $\mathcal{H}$ in the varied spacetime corresponding to the bifurcation surface. However, since as many as $m+1$ derivatives of ξ^a occur in (69), we must also give a construction of ξ^a off of $\mathcal{H}$ at least up to this order for the case of a metric that does not possess a horizon Killing field. Once we have determined ξ^a, we may define N^a on $\mathcal{H}$ by (35) and the first equalities of (37) and (38).

A simple and natural way to rigidly tie g_{ab} to ξ^a in a neighborhood of $\mathcal{H}$ would be to introduce Gaussian null coordinates (GNCs) on $\mathcal{H}$ adapted to a choice of affine parameter V on the null geodesics of $\mathcal{H}$, with $V = 0$ corresponding to $\mathcal{B}$. In these coordinates, the metric takes the form

$$ds^2 = 2dV\left(d\rho - \rho\beta_A dx^A - \frac{1}{2}\rho^2\alpha dV\right) + \gamma_{AB}dx^A dx^B, \tag{81}$$

where x^A are coordinates on a cut of constant V and ρ is an affine parameter along the null geodesics transverse to $\mathcal{H}$ that are orthogonal to these cuts. We may then define ξ^a in a neighborhood of $\mathcal{H}$ by

$$\xi^a = \kappa\left[V\left(\frac{\partial}{\partial V}\right)^a - \rho\left(\frac{\partial}{\partial\rho}\right)^a\right], \tag{82}$$

and we may then apply a suitable diffeomorphism to g_{ab}, ξ^a such that ξ^a is the same in each spacetime.[6]

For the case of a metric with a horizon Killing field with surface gravity κ and bifurcation surface $\mathcal{B}$, this definition reproduces the horizon Killing field [8], but this formula is well-defined in a neighborhood of $\mathcal{H}$ for an arbitrary metric. In addition, the metric in this gauge satisfies conditions that yield (42) and (43) for perturbations about a metric with a horizon Killing field. However, there is a rescaling freedom $V \to f(x^A)V$ in the choice of V. Although this rescaling freedom does not affect the definition (82) of ξ^a in the case where the metric possesses a horizon Killing field with surface gravity κ and bifurcation surface $\mathcal{B}$, it will affect the definition of ξ^a in the general case. Thus, even with this choice of prescription, we have ambiguities in the definition of ξ^a and corresponding ambiguities in the definition of $\mathbf{B}_\mathcal{H}$. Of course, other prescriptions for defining ξ^a off of $\mathcal{H}$ would also be possible. In Appendix C we will give an alternative construction of ξ^a.

It is not necessarily a bad thing that $\mathbf{B}_\mathcal{H}$—and thereby our definition of entropy—depends on a choice of "time translation" ξ^a. It is generally recognized that the notion of energy should be defined as conjugate to a notion of time translations. Since energy and entropy appear together in the first law of thermodynamics for near-equilibrium systems, it seems reasonable that the notion of entropy should similarly depend on a notion of time translations. Nevertheless, the ambiguities in defining a notion of "time translation" gives rise to ambiguities in our definition of entropy beginning at second order in perturbation theory about a black hole whose event horizon is a Killing horizon.

The second source of ambiguities in the formula for entropy at second order arises from the fact that $\mathbf{B}_\mathcal{H}$ was

[6]In other words, we identify different spacetimes by identifying points near $\mathcal{H}$ with the same coordinates (V,ρ,x^A).

constructed so as to satisfy (70), which is a condition that applies only to first-order perturbations. Therefore, we have freedom to add to $\mathbf{B}_{\mathcal{H}}$ terms that vanish identically for first-order perturbations about a background spacetime with a Killing horizon but do not vanish at second order. Nevertheless, based upon the construction of $\mathbf{B}_{\mathcal{H}}$ given in the previous section, it seems natural to impose the following conditions on $\mathbf{B}_{\mathcal{H}}$:

(1) We require $\mathbf{B}_{\mathcal{H}}$ to be locally and covariantly constructed out of the metric, the curvature tensor, derivatives of the curvature tensor, and the vector fields ξ^a, N^a and their covariant derivatives.[7]
(2) We require that each term in the expression for $\mathbf{B}_{\mathcal{H}}$ contains a factor[8] of $\mathcal{L}_\xi g_{ab}$ (or covariant derivatives of this quantity). In particular, we have $\mathbf{B}_{\mathcal{H}} = 0$ in the stationary background.[9]
(3) We require that (70) holds for arbitrary perturbations of a stationary background.
(4) We require that the total number of derivatives appearing in $\mathbf{B}_{\mathcal{H}}$ cannot exceed the total number of derivatives in $\boldsymbol{\theta}$, which is one fewer than the total number of derivatives appearing in the Lagrangian. (Here, a factor of the curvature tensor counts as "two derivatives" in the Lagrangian and in $\mathbf{B}_{\mathcal{H}}$. The factor of $\mathcal{L}_\xi g_{ab}$ in $\mathbf{B}_{\mathcal{H}}$ counts as "one derivative." However, the appearance of N^a does not count towards the derivative limitation in $\mathbf{B}_{\mathcal{H}}$—despite its being determined by the derivative of ξ^a—since its presence can also arise from the purely algebraic basis expansion.)
(5) We require that $\mathbf{B}_{\mathcal{H}}$ is invariant under a rescaling of ξ^a by a constant, $\xi^a \to c\xi^a$ and the corresponding rescaling $\kappa \to c\kappa$.

These conditions are quite restrictive for theories with a relatively low total number of derivatives in the Lagrangian. In particular, they uniquely determine $\mathbf{B}_{\mathcal{H}}$ in general relativity. However, by the time one reaches curvature cubed theories (i.e., a total of six derivatives in the Lagrangian), it is not difficult to construct terms—such as $\epsilon_{ab_1\cdots b_{n-1}} R_{pq} N^p N^q R^a{}_c \xi^c g^{mn} \mathcal{L}_\xi g_{mn}$—that could be added to $\mathbf{B}_{\mathcal{H}}$ so as satisfy all of the above conditions and are nonvanishing at second order. It is possible that one could impose additional conditions on $\mathbf{B}_{\mathcal{H}}$ so as to further restrict its ambiguities and/or obtain other desired properties of $\mathbf{S}$. One attractive possibility would be to impose conditions that imply that $S[\mathcal{C}]$ satisfies a second law (i.e., that it is nondecreasing with time) at second order (see Sec. V B). However, we do not believe that it would be possible to do this within our framework.

It is instructive to compare the nature of our framework and its definition freedom arising at second order with the framework and definition freedom in the approach of Hollands, Kovacs, and Reall (HKR) [18]. As will be discussed further in Appendix C 3, HKR take an effective field theory approach, wherein the Lagrangian is expanded about the general relativity Lagrangian in powers of a small parameter ℓ and one works only to some finite order in ℓ. The HKR entropy-current $(n-2)$-form $\mathbf{S}_{\mathrm{HKR}}$ is constructed so as to agree with the Dong-Wall entropy (see Sec. VI) rather than our entropy current to first order in perturbation theory about a black hole with a bifurcate Killing horizon. The HKR construction is aimed at providing an entropy satisfying the second law to second order in perturbation theory [18], or even nonperturbatively in a recent extension [23]. But it need not be covariant in the metric [24]. As discussed further in Appendix C 3, the lack of a covariant HKR entropy current $(n-2)$-form means that there is no guarantee that the HKR entropy will be "cross section continuous", [21] i.e., if we "wiggle" the cut $\mathcal{C}$ into another cut $\mathcal{C}'$ that is arbitrarily close to $\mathcal{C}$, there is no guarantee that $S_{\mathrm{HKR}}[\mathcal{C}']$ will be close to $S_{\mathrm{HKR}}[\mathcal{C}]$, as illustrated in Fig. 1. Although we will see in Sec. VI that (our and) the Dong-Wall entropy is cross section continuous, it appears that the HKR entropy will not, in general, satisfy this property. A detailed comparison of the definition freedom in entropy at second order in our approach and the HKR approach is given in Appendix C 3.

Finally, we comment on the definition of $\mathbf{S}$ and $S[\mathcal{C}]$ outside of the perturbative context, i.e., for an arbitrary dynamical black hole. We have already noted above that our approach to defining entropy is heavily reliant on the presence of a time translation vector field ξ^a. In the case of a stationary black hole whose event horizon is a Killing horizon, the horizon Killing field ξ^a provides a natural notion of time translations upon which to base notions of energy and entropy. For a black hole that is very nearly

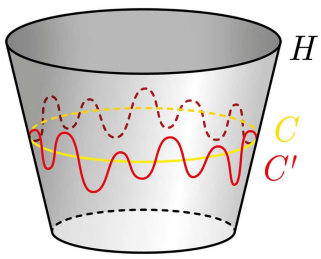

FIG. 1. Cross section continuity [21] of $S[\mathcal{C}]$ means that $S[\mathcal{C}']$ approaches $S[\mathcal{C}]$ for a cut $\mathcal{C}'$ tending to a cut $\mathcal{C}$, even if $\mathcal{C}'$ is very wiggly as illustrated.

[7]Note that N^a on $\mathcal{H}$ itself is locally and covariantly constructed from the metric, ξ^a, and the first derivative of ξ^a, since it is uniquely determined in a local manner by (35) and the first equalities of (37) and (38). It would be consistent with our construction of $\mathbf{B}_{\mathcal{H}}$ in Sec. III at first order to require that no derivatives of N^a appear in the expression for $\mathbf{B}_{\mathcal{H}}$.

[8]When matter fields are present, there may be terms proportional to the matter field contracted into ξ^a that do not contain a factor involving $\mathcal{L}_\xi$.

[9]It would be consistent with our construction of $\mathbf{B}_{\mathcal{H}}$ in Sec. III at first order to require that there be no more than m covariant derivatives of $\mathcal{L}_\xi g_{ab}$.

stationary—in particular, in first- and second-order perturbation theory about a stationary black hole—it also appears to be natural to use the definition of ξ^a given above to play the role of time translations. However, for a black hole that is far from "equilibrium", it seems quite artificial to attempt to assign a surface gravity κ and bifurcation surface $\mathcal{B}$ to the event horizon $\mathcal{H}$ so as to construct a Killing-like vector field ξ^a. Furthermore, even in ordinary physics where a notion of time translation is present, the definition of the entropy of a system that is far from equilibrium is not well defined without making a choice of coarse-grained observable [25]. Thus, we do not advocate applying our formulas for $\mathbf{S}$ and $S[\mathcal{C}]$ outside of the perturbative context around a stationary (i.e., "equilibrium") black hole.

We now evaluate our formula for $\mathbf{S}$ and $S[\mathcal{C}]$ for the case of general relativity and, more generally, for theories whose Lagrangians do not depend on derivatives of the curvature tensor.

B. Dynamical black hole entropy in general relativity

The Lagrangian for general relativity is

$$L_{a_1\cdots a_n} = \frac{1}{16\pi}\epsilon_{a_1\cdots a_n}R. \tag{83}$$

The symplectic potential $\boldsymbol{\theta}$ is [4]

$$\theta_{a_1\cdots a_{n-1}} = \frac{1}{16\pi}\epsilon^{e}{}_{a_1\cdots a_{n-1}}g^{fh}(\nabla_f\delta g_{eh} - \nabla_e\delta g_{fh}) \tag{84}$$

and the Noether charge is [4]

$$Q_{a_1\cdots a_{n-2}} = -\frac{1}{16\pi}\epsilon_{a_1\cdots a_{n-2}cd}\nabla^c\xi^d. \tag{85}$$

Using the methods employed in Sec. III to obtain $\mathbf{B}_{\mathcal{H}}$ from $\boldsymbol{\theta}$, we obtain

$$B_{\mathcal{H}a_1\cdots a_{n-1}} = -\frac{1}{16\pi}N^b\epsilon_{ba_1\cdots a_{n-1}}g^{ef}\mathcal{L}_\xi g_{ef}. \tag{86}$$

It is easily seen that $\mathbf{B}_{\mathcal{H}}$ satisfies all of the requirements enumerated above in Sec. IV A. Furthermore, $\mathbf{B}_{\mathcal{H}}$ is unique, since it is allowed to contain only one derivative—which is used up by the required factor of $\mathcal{L}_\xi g_{ef}$—and no terms of the required form can be constructed that vanish for arbitrary first-order perturbations. In terms of the affinely parametrized tangent $k^a = \xi^a/(\kappa V)$ [see (40)], we can rewrite $\mathbf{B}_{\mathcal{H}}$ as

$$B_{\mathcal{H}a_1\cdots a_{n-1}} = -\frac{1}{16\pi}n^b\epsilon_{ba_1\cdots a_{n-1}}g^{ef}\mathcal{L}_k g_{ef} = -\frac{1}{8\pi}n^b\epsilon_{ba_1\cdots a_{n-1}}\vartheta, \tag{87}$$

where ϑ is the expansion of the null generators of the horizon in the affine parametrization V and $n^a = \kappa V N^a$ (so $n^a k_a = 1$). This agrees with Eq. (6.24) in [26] (see also [27]) for the corresponding term in the symplectic potential on an arbitrary null surface needed to implement the Wald-Zoupas prescription for defining charges.

The entropy $(n-2)$-form (71) in general relativity is

$$S_{a_1\cdots a_{n-2}} = -\frac{1}{8\kappa}\epsilon_{a_1\cdots a_{n-2}cd}\nabla^c\xi^d - \epsilon^{(n-2)}_{a_1\cdots a_{n-2}}\frac{1}{4}V\vartheta, \tag{88}$$

where $\epsilon^{(n-2)}_{a_1\cdots a_{n-2}} = \xi^b N^c \epsilon_{bca_1\cdots a_{n-2}}$ was defined by (47) and (48) and is such that its pullback to any cross section $\mathcal{C}$ on the horizon yields the volume element on $\mathcal{C}$ obtained from the pullback of the metric to $\mathcal{C}$. For a given cross section $\mathcal{C}$ of the horizon, the entropy is

$$S[\mathcal{C}] = \frac{A[\mathcal{C}]}{4} - \frac{1}{4}\int_{\mathcal{C}} V\vartheta\boldsymbol{\epsilon}^{(n-2)}, \tag{89}$$

where $A[\mathcal{C}]$ is the ($(n-2)$-dimensional) area of $\mathcal{C}$ and the integral over $\mathcal{C}$ in the second term is taken with respect to the natural volume element $\epsilon^{(n-2)}_{a_1\cdots a_{n-2}}$ on $\mathcal{C}$. The first term in (89) is the usual Bekenstein-Hawking entropy. For a stationary black hole, we have $\vartheta = 0$. However, for a dynamical black hole, we have $\vartheta > 0$, so $S[\mathcal{C}]$ is *smaller* than the Bekenstein-Hawking entropy. We will show in Appendix A that for first-order perturbations, $S[\mathcal{C}]$ in general relativity is, in fact, equal to the area of the apparent horizon at the "time" corresponding to the cross section, $\mathcal{C}$, of the horizon.

Rignon-Bret [28] (see also [29]) have given an analysis of aspects of our entropy expression (89) as well as an alternative entropy expression based on the York symplectic potential that is constructed to vanish on any cross section of a light cone in Minkowski spacetime.

C. Entropy form for $L(R_{abcd})$ theories

Consider a theory of gravity obtained from a Lagrangian that depends on the metric and curvature but does not depend on derivatives of the curvature,

$$L_{a_1\cdots a_n} = \epsilon_{a_1\cdots a_n}L(g_{ab}, R_{abcd}). \tag{90}$$

The symplectic potential $(n-1)$-form can be calculated from the algorithm given in Lemma 3.1 of [4] or directly from[10] Eq. (4.5) of [30]. We obtain,

$$\theta_{a_1\cdots a_{n-1}} = 2\epsilon_{ma_1\cdots a_{n-1}}\left[\frac{\partial L}{\partial R_{mbcd}}\nabla_d\delta g_{bc} - \nabla_d\left(\frac{\partial L}{\partial R_{mbcd}}\right)\delta g_{bc}\right]. \tag{91}$$

The Noether charge $(n-2)$-form can be obtained either

[10]The sign difference between our formula and Eq. (4.5) of [30] appears to be a result of differences in conventions.

from the Noether current $\mathbf{J}$ via the algorithm given in Lemma 1 of [31] or directly from Eq. (4.6) of [30],

$$Q_{a_1\cdots a_{n-2}} = \epsilon_{cda_1\cdots a_{n-2}}\left(-\frac{\partial L}{\partial R_{cdef}}\nabla_{[e}\xi_{f]} - 2\nabla_f\left(\frac{\partial L}{\partial R_{cdef}}\right)\xi_e\right). \tag{92}$$

From the construction in Sec. III, we obtain

$$B_{\mathcal{H}a_1\cdots a_{n-1}} = 2N^f\epsilon_{fa_1\cdots a_{n-1}}\xi_m\frac{\partial L}{\partial R_{mbcd}}N_d\mathcal{L}_\xi g_{bc}. \tag{93}$$

The entropy $(n-2)$-form is then

$$S_{a_1\cdots a_{n-2}} = \frac{2\pi}{\kappa}\left[\epsilon_{cda_1\cdots a_{n-2}}\left(-\frac{\partial L}{\partial R_{cdef}}\nabla_{[e}\xi_{f]} - 2\nabla_f\left(\frac{\partial L}{\partial R_{cdef}}\right)\xi_e\right) + 2\epsilon^{(n-2)}_{a_1\cdots a_{n-2}}\xi_m\frac{\partial L}{\partial R_{mbcd}}N_d\mathcal{L}_\xi g_{bc}\right]. \tag{94}$$

We now explicitly evaluate the above quantities for the case of a general curvature squared Lagrangian, i.e.,

$$L_{a_1\cdots a_n} = \epsilon_{a_1\cdots a_n}(\alpha_1 R_{abcd}R^{abcd} + \alpha_2 R_{ab}R^{ab} + \alpha_3 R^2), \tag{95}$$

where α_1, α_2, α_3 are arbitrary constants. The symplectic potential $(n-1)$-form is[11]

$$\begin{aligned}\theta_{a_1\cdots a_{n-1}} = \epsilon_{ma_1\cdots a_{n-1}}\{&4\alpha_1[R^{mbcd}\nabla_d\delta g_{bc} - \nabla_d R^{mbcd}\delta g_{bc}]\\ &+ \alpha_2[2g^{c[m}R^{b]d}\nabla_d\delta g_{bc} + 2R^{c[m}\nabla^{b]}\delta g_{bc} + g^{c[b}\nabla^{m]}R\delta g_{bc} + 2\nabla^{[m}R^{b]c}\delta g_{bc}]\\ &+ 2\alpha_3[Rg^{mc}g^{bd}(\nabla_d\delta g_{bc} - \nabla_c\delta g_{bd}) - 2g^{c[m}\nabla^{b]}R\delta g_{bc}]\}.\end{aligned} \tag{96}$$

The Noether charge $(n-2)$-form is

$$\begin{aligned}Q_{a_1\cdots a_{n-2}} = -\epsilon_{cda_1\cdots a_{n-2}}&[(2\alpha_1 R^{cdef} + 2\alpha_2 g^{ce}R^{fd} + 2\alpha_3 Rg^{ce}g^{df})\nabla_{[e}\xi_{f]}\\ &+ (4\alpha_1\nabla_f R^{cdmf} + \alpha_2 g^{cm}\nabla^d R - 2\alpha_2\nabla^c R^{md} + 4\alpha_3\nabla^d Rg^{cm})\xi_m].\end{aligned} \tag{97}$$

The $\mathbf{B}_\mathcal{H}$ $(n-1)$-form is

$$B_{\mathcal{H}a_1\cdots a_{n-1}} = N^f\epsilon_{fa_1\cdots a_{n-1}}(4\alpha_1 R^{abcd}\xi_a N_d - \alpha_2 R^{bc} - \alpha_2 R_{ad}\xi^a N^d g^{bc} - 2\alpha_3 Rg^{bc})\mathcal{L}_\xi g_{bc}. \tag{98}$$

The entropy $(n-2)$-form is

$$\begin{aligned}S_{a_1\cdots a_{n-2}} = -\frac{2\pi}{\kappa}(&\epsilon_{cda_1\cdots a_{n-2}}[(2\alpha_1 R^{cdef} + 2\alpha_2 g^{ce}R^{fd} + 2\alpha_3 Rg^{ce}g^{df})\nabla_{[e}\xi_{f]}\\ &+ (4\alpha_1\nabla_f R^{cdmf} + \alpha_2 g^{cm}\nabla^d R - 2\alpha_2\nabla^c R^{md} + 4\alpha_3\nabla^d Rg^{cm})\xi_m]\\ &- \epsilon^{(n-2)}_{a_1\cdots a_{n-2}}(4\alpha_1 R^{abcd}\xi_a N_d - \alpha_2 R^{bc} - \alpha_2 R_{ad}\xi^a N^d g^{bc} - 2\alpha_3 Rg^{bc})\mathcal{L}_\xi g_{bc}).\end{aligned} \tag{99}$$

As will be discussed in Sec. VI, this yields a formula for $S[\mathcal{C}]$ that differs from the Dong-Wall entropy in a manner analogous to how our formula for $S[\mathcal{C}]$ in general relativity differs from the Bekenstein-Hawking entropy.

V. THE PHYSICAL PROCESS FIRST LAW AND THE SECOND LAW

As explained in the Introduction, the main motivation for our new definition of dynamical black hole entropy is to obtain a local in time version of the physical process version of the first law of black hole mechanics. In this section, we derive the physical process version of the first law and discuss the second law of black hole mechanics. As stated at the end of Sec. II, we restrict consideration to Lagrangians for which the only dynamical field is the metric. However, in this context, there are two distinct cases that we consider.

The first case is where an external stress-energy is present in the first-order perturbation. Here, by an external stress-energy, we mean a source term in the first-order gravitational equations arising from matter that is not represented in the Lagrangian. In this case, as we shall see, there can be nontrivial time dependence of the dynamical black hole entropy S at first order when the external matter crosses the horizon. We show that the

[11]The choice of $\alpha_1 = 1$, $\alpha_2 = -4$, $\alpha_3 = 1$ corresponds to Gauss-Bonnet gravity. There is a typo—previously noted in [32]—in the sign of the second-to-last term in the formula for $\boldsymbol{\theta}$ for Gauss-Bonnet gravity given in Eq. (70) of [4].

entropy change is related to the energy flux by the first law of black hole mechanics. As an immediate corollary, we see that if the external matter satisfies the null energy condition, then the second law of black hole mechanics holds, i.e., the entropy is nondecreasing with time in first-order perturbation theory. These results hold for an arbitrary theory of gravity.

The second case is that of pure vacuum perturbations. In this case, S does not change with time at first order, so we have to go to second order in perturbation theory to investigate the time dependence of S. We find that at second order, results similar to the external matter case hold, with the first-order energy flux of external matter replaced by the second-order flux of modified canonical energy [8]. The flux of modified canonical energy is positive in general relativity—it is proportional to the square of the first-order perturbed shear of the horizon—so the second law of black hole mechanics holds at second order in general relativity. However, the flux of modified canonical energy would not be expected to be positive in general theories of gravity, so we would not expect a second law to hold for vacuum perturbations in general theories.

A. Perturbations with external matter sources

For an arbitrary Lagrangian theory of gravity where the only dynamical field is the metric, we consider a stationary black hole with horizon Killing field ξ^a that satisfies the vacuum field equations $(E_G)_{ab} = 0$. We consider perturbations, δg_{ab}, of this spacetime that are sourced by a matter stress-energy tensor, δT_{ab}, that is conserved, $\nabla^a \delta T_{ab} = 0$. Thus, δg_{ab} satisfies

$$2\delta(E_G)_{ab} = \delta T_{ab}. \tag{100}$$

The perturbed matter energy flux $(n-1)$-form $\mathbf{e}$ on the horizon is given by

$$\delta e_{a_1\cdots a_{n-1}} \equiv -\xi^b \delta T_b{}^c \epsilon_{ca_1\cdots a_{n-1}} = \delta T_{bc}\xi^b\xi^c \epsilon^{(n-1)}_{a_1\cdots a_{n-1}}, \tag{101}$$

where $\epsilon^{(n-1)}_{a_1\cdots a_{n-1}}$ was defined by (47). We have the following theorem:

Theorem 2 (Physical process version of the first law [9]). Let $\mathcal{C}_1$ and $\mathcal{C}_2$ be arbitrary cross sections of the horizon $\mathcal{H}$ with $\mathcal{C}_2$ lying to the future of $\mathcal{C}_1$. Let $\Delta\delta S$ denote the perturbed entropy difference between the cross sections

$$\Delta\delta S = \delta S[\mathcal{C}_2] - \delta S[\mathcal{C}_1] \tag{102}$$

and let $\Delta\delta E$ denote the perturbed total energy flux between the cross sections

$$\Delta\delta E = \int_{\mathcal{H}_{12}} \delta\mathbf{e} = -\int_{\mathcal{H}_{12}} \xi^b \delta T_b{}^c \epsilon_{ca_1\cdots a_{n-1}} = \int_{\mathcal{H}_{12}} \delta T_{bc}\xi^b\xi^c \epsilon^{(n-1)}_{a_1\cdots a_{n-1}} \tag{103}$$

where $\mathcal{H}_{12}$ denotes the portion of $\mathcal{H}$ bounded by $\mathcal{C}_1$ and $\mathcal{C}_2$. Then we have

$$\frac{\kappa}{2\pi}\Delta\delta S = \Delta\delta E. \tag{104}$$

Proof. The pullback to the horizon of the fundamental identity (30) with χ^a taken to be the horizon Killing vector field ξ^a yields

$$d[\delta\mathbf{Q}[\xi] - \xi\cdot\boldsymbol{\theta}] = -\xi^a \delta\mathbf{C}_a. \tag{105}$$

Using the definition (71) of $\mathbf{S}$ together with (70), we can rewrite the left side of (105) as

$$d[\delta\mathbf{Q}[\xi] - \xi\cdot\boldsymbol{\theta}] = \frac{\kappa}{2\pi} d\delta\mathbf{S}. \tag{106}$$

Using the general formula (28) with $E_M = 0$ for $\mathbf{C}_a$ together with (100), we can rewrite the right side of (105) as

$$-\xi^a\delta\mathbf{C}_a = \delta\mathbf{e} \tag{107}$$

so we have

$$\frac{\kappa}{2\pi} d\delta\mathbf{S} = \delta\mathbf{e}. \tag{108}$$

Integration of this equation over the region $\mathcal{H}_{12}$ of the horizon bounded by $\mathcal{C}_1$ and $\mathcal{C}_2$ then immediately yields (104) ∎.

Remark. In the case of general relativity, (104) yields

$$\frac{\kappa}{8\pi}\Delta\left[A - \int_{\mathcal{C}} V\vartheta\boldsymbol{\epsilon}^{(n-2)}\right] = \Delta E. \tag{109}$$

This result can also be derived by multiplying the Raychaudhuri equation by κV and integrating the resulting expression between the two cross sections [20,33].

There are two important immediate consequences of Theorem 2. First, if δT_{bc} satisfies the null energy condition (i.e., if $\delta T_{bc}n^b n^c \geq 0$ for all null n^a), then it follows immediately from (103) that $\Delta\delta E \geq 0$. It then follows immediately from Theorem 2 that $\Delta\delta S \geq 0$, i.e., $\delta S[\mathcal{C}_2] \geq \delta S[\mathcal{C}_1]$ whenever $\mathcal{C}_2$ lies to the future of $\mathcal{C}_1$. Thus, in an arbitrary theory of gravity, the second law of black hole mechanics holds to first order for perturbations of a black hole with external matter satisfying the null energy condition.

Second, Theorem 2 shows that, to first order, the entropy of a black hole changes between cross sections $\mathcal{C}_1$ and $\mathcal{C}_2$ if and only if there is a net energy flux of matter between $\mathcal{C}_1$ and $\mathcal{C}_2$. Thus, the increase in entropy of the black hole has an entirely local cause. In the case of general relativity, this contrasts sharply with the properties of the Bekenstein-Hawking entropy, $A/4$. The event horizon of a black hole is "teleological" in nature, i.e., its location is determined by properties of the spacetime in the asymptotic future. If one is going to throw matter into a black hole at a later time, then at early times the event horizon area will have already slightly increased in anticipation of the later arrival of the matter. In other words, if one makes a firm decision to throw a baseball into a black hole, the Bekenstein-Hawking entropy of the black hole will have increased while the baseball is still in one's hand. Our definition of the entropy of a dynamical black hole does not have this property. Although the $A/4$ term in (89) will have increased while the baseball is still in one's hand, the horizon will also be expanding at that time, so the second term in (89) will contribute negatively. Theorem 2 shows that these two terms cancel at first order, so there is no change in the black hole entropy while the baseball is still in one's hand. The entropy of the black hole changes only when the baseball crosses the horizon.

B. Vacuum perturbations

If we have no external matter sources, $\delta T_{ab} = 0$, then Theorem 2 states that $\Delta\delta S = 0$, i.e., there is no change of black hole entropy with time at first order. Thus, we must go to second order in perturbation theory to obtain the leading-order dynamical behavior of black hole entropy. To do so, we use the varied fundamental identity (31) with $\chi^a = \xi^a$. Since we assume that the vacuum equations of motion hold, the first three terms on the right side of (31) vanish.[12] We obtain[13]

$$\boldsymbol{\omega}(g, \delta g, \mathcal{L}_\xi \delta g) = d[\delta^2 \mathbf{Q}[g; \xi] - \xi \cdot \delta\boldsymbol{\theta}(g, \delta g)]. \tag{110}$$

We can rewrite this equation as

$$\begin{aligned}&\boldsymbol{\omega}(g, \delta g, \mathcal{L}_\xi \delta g) + d[\xi \cdot \delta\boldsymbol{\theta}(g, \delta g) - \xi \cdot \delta^2 \mathbf{B}_{\mathcal{H}}(g))] \\ &\quad = d[\delta^2 \mathbf{Q}[g; \xi] - \xi \cdot \delta^2 \mathbf{B}_{\mathcal{H}}(g)] = \frac{\kappa}{2\pi} d\delta^2 \mathbf{S}(g).\end{aligned} \tag{111}$$

The term $\boldsymbol{\omega}(g, \delta g, \mathcal{L}_\xi \delta g)$ is known as the *canonical energy* $(n-1)$-form [8] of the metric perturbation δg. We refer to entire left side of (111) as the *modified canonical energy* $(n-1)$-form $\mathbf{e}_G$,

$$\begin{aligned}\mathbf{e}_G(g; \delta g, \delta g) \equiv\ &\boldsymbol{\omega}(g, \delta g, \mathcal{L}_\xi \delta g) \\ &+ d[\xi \cdot \delta\boldsymbol{\theta}(g, \delta g) - \xi \cdot \delta^2 \mathbf{B}_{\mathcal{H}}(g))].\end{aligned} \tag{112}$$

Note that $\mathbf{e}_G$ is quadratic in δg_{ab} and does not depend on $\delta^2 g_{ab}$. The modified canonical energy was introduced in [8] for the case of general relativity as a means of obtaining a quantity with a locally positive flux whose total flux through the black hole horizon agrees with the total canonical energy flux. For the case of general relativity, the modified canonical energy flux is given by

$$\mathbf{e}_G = \frac{1}{4\pi}(\kappa V)^2 \delta\sigma_{cd} \delta\sigma^{cd} \boldsymbol{\epsilon}^{(n-1)}, \tag{113}$$

[see Eq. (105) of [8]], where V is an affine parameter, σ_{cd} is the shear of the generators of the horizon with respect to V, and we have used the fact [8] that $\delta\vartheta|_{\mathcal{H}} = 0$ for vacuum perturbations. Although the modified canonical energy was introduced in an entirely *ad hoc* way in [8], we now see that it naturally enters the formula for the local change in dynamical black hole entropy at second order.

From (111) we see that the situation for vacuum perturbations at second order corresponds to the situation for external matter perturbations at first order, with the modified canonical energy flux $\mathbf{e}_G$ replacing the matter stress-energy flux $\mathbf{e}$. In particular, for vacuum perturbations we have

$$\frac{\kappa}{2\pi} \Delta\delta^2 S = \int_{\mathcal{H}_{12}} \mathbf{e}_G(g; \delta g, \delta g). \tag{114}$$

It follows immediately that, in an arbitrary theory of gravity, the second law of black hole mechanics will hold at second order for vacuum perturbations if and only if the modified canonical energy flux through the horizon is positive. This is the case in general relativity. However, it seems unlikely that this will be the case in more general theories of gravity even if we allow ourselves to modify the entropy making use of the permissible ambiguities (see Appendix C 3 for further discussion).

VI. RELATIONSHIP WITH THE DONG-WALL ENTROPY

In 2013, Dong [5] proposed a formula for dynamical black hole entropy for theories obtained from a Lagrangian that depends upon the curvature but not derivatives of the curvature. In 2015, Wall [6] proposed a formula for dynamical black hole entropy for general Lagrangian theories, valid to linear order for perturbations of a stationary black hole. Wall's approach for obtaining dynamical black hole entropy does not have any obvious relationship with Dong's. Nevertheless, for the case of a Lagrangian that depends upon the curvature but not its derivatives, Wall's formula was found to agree [6] with the

[12]The first two terms would vanish in any case under pullback to $\mathcal{H}$.

[13]Here δ^2, the second variation, refers to the second derivative with respect to a parameter s of a family of metrics $g^{(s)}_{ab} = g_{ab} + s\delta g_{ab} + \frac{1}{2}s^2\delta^2 g_{ab} + O(s^3)$.

linearization of Dong's formula. In Appendix D, we show that when we apply the ideas of Wall's method as elucidated in this section to calculate entropy in dilaton gravity models, it agrees with the linearization of the entropy obtained by Dong and Lewkowycz [34] using Dong's approach. Thus, we believe that Wall's entropy agrees with the linearization of Dong's entropy in general. We will therefore refer to this entropy as the Dong-Wall entropy and will denote it by $S_{\rm DW}$. Nevertheless, it should be kept in mind that general agreement has not been proven. In the event that they do not agree, the entropy defined in this section would correspond to that proposed by Wall [6].

In this section, we will give a derivation of the Dong-Wall entropy based on the fundamental identity (30). We will then obtain a simple, general relationship between the Dong-Wall entropy and our entropy, valid up to linear order in perturbations about a stationary black hole. As we shall see, the relationship between the Dong-Wall entropy and our entropy in general theories of gravity is very similar to the relationship (89) between the Bekenstein-Hawking entropy and our entropy in general relativity.

The starting point of our derivation of the Dong-Wall entropy is the formula (28) for the pullback of the linearized constraints to the horizon, which take the form

$$\xi^a \delta \mathbf{C}_a = -2(\delta E_G)_{ab}\xi^a\xi^b \boldsymbol{\epsilon}^{(n-1)}, \tag{115}$$

where $\boldsymbol{\epsilon}^{(n-1)}$ was defined by (47). The linearized equations of motion $(\delta E_G)_{ab}$ take the general form

$$\delta E_G^{ab} = \sum_{i=0}^{k} U_{(i)}^{abc_1\cdots c_i de}\nabla_{(c_1}\cdots\nabla_{c_i)}\delta g_{de}, \tag{116}$$

where the tensors $U_{(i)}^{abc_1\cdots c_i de} = U_{(i)}^{(ab)(c_1\cdots c_i)(de)}$ are smooth and are locally and covariantly constructed from the metric, curvature, and covariant derivatives of the curvature. This is closely analogous to the quantity α^a in (51) except that there are two free indices on δE_G^{ab} instead of the single free index on α^a. Both of these free indices on δE_G^{ab} are contracted into ξ^a in (115), as opposed to the single contraction that occurs in (52). Thus, one might expect that a strengthened form of Theorem 1 should hold for $(\delta E_G)_{ab}\xi^a\xi^b$. We show in Appendix B that this is the case. Specifically, we show that $(\delta E_G)_{ab}\xi^a\xi^b$ can be written in the form

$$(\delta E_G)_{ab}\xi^a\xi^b \hat{=} \sum_{i=0}^{k-2} \tilde{U}_{(i)}^{c_1\cdots c_i de}\nabla_{(c_1}\cdots\nabla_{c_i)}(\mathcal{L}_\xi^2 - \kappa\mathcal{L}_\xi)\delta g_{de}, \tag{117}$$

where the tensors $\tilde{U}_{(i)}^{c_1\cdots c_i de}$ are smooth on $\mathcal{H}$ and are locally and covariantly constructed from the metric, curvature, and its derivatives as well as from ξ^a, and N^a [see (33) and (34)], with ξ^a and N^a appearing only algebraically. Thus, we see that $(\delta E_G)_{ab}\xi^a\xi^b$ has the form

$$\begin{aligned}(\delta E_G)_{ab}\xi^a\xi^b &\hat{=} \kappa\mathcal{L}_\xi P(g,\xi;\delta g) - \mathcal{L}_\xi^2 P(g,\xi;\delta g)\\ &= (\kappa - \mathcal{L}_\xi)\delta P(g,\xi;\mathcal{L}_\xi g)\end{aligned} \tag{118}$$

with

$$P(g,\xi;\delta g) = -\sum_{i=0}^{k-2}\tilde{U}_{(i)}^{c_1\cdots c_i de}\nabla_{(c_1}\cdots\nabla_{c_i)}\delta g_{de}. \tag{119}$$

We define the Dong-Wall entropy $(n-2)$-form by[14]

$$\mathbf{S}_{\rm DW} = \mathbf{S} + \frac{4\pi}{\kappa}P(g,\xi;\mathcal{L}_\xi g)\boldsymbol{\epsilon}^{(n-2)}, \tag{120}$$

where our entropy $(n-2)$-form $\mathbf{S}$ was defined by (71) and $\boldsymbol{\epsilon}^{(n-2)}$ was defined in (48). It is clear that $\mathbf{S}_{\rm DW} = \mathbf{S}$ in the stationary background, wherein both are given by (73). Since $\delta\mathbf{S}$ is uniquely defined for arbitrary first-order perturbations and δP is uniquely determined by (118), we see that $\delta\mathbf{S}_{\rm DW}$ is uniquely defined for arbitrary first-order perturbations. However, there will be ambiguities in $\mathbf{S}_{\rm DW}$ at second and higher orders resulting from the ambiguities in $\mathbf{S}$ discussed at the end of Sec. IV A (see also Appendix C) as well as from the additional ambiguities in defining P at second and higher orders. The Dong-Wall entropy of a cross section $\mathcal{C}$ is defined by

$$S_{\rm DW}[\mathcal{C}] = \int_{\mathcal{C}}\mathbf{S}_{\rm DW}. \tag{121}$$

Since $\mathbf{S}_{\rm DW}$ does not depend on a choice of cross section and $\mathbf{S}_{\rm DW}$ is smooth on $\mathcal{H}$, it follows immediately that $S_{\rm DW}$ is cross section continuous, as defined in [21].

A simple relationship between $\mathbf{S}_{\rm DW}$ and $\mathbf{S}$ for first-order perturbations can now be obtained as follows. As in the proof of Theorem 2, we have

$$\frac{\kappa}{2\pi}d\delta\mathbf{S} = -\xi^a\delta\mathbf{C}_a. \tag{122}$$

Using (115) and (118), we obtain

$$\frac{\kappa}{2\pi}d\delta\mathbf{S} = 2(\kappa - \mathcal{L}_\xi)[\delta P(g,\xi;\mathcal{L}_\xi g)\boldsymbol{\epsilon}^{(n-1)}]. \tag{123}$$

Applying the general identity (80) to $\mathcal{L}_\xi[\delta P(g,\xi;\mathcal{L}_\xi g)\boldsymbol{\epsilon}^{(n-1)}]$, we obtain

[14]Reanalyzing Wall's derivation [6], the authors of [35] have emphasized the fact that the entropy integrand has a current part, i.e., it is an $(n-2)$-form. However, we have not been able to find a relationship between their proposed formula and our formula (120).

$$\mathcal{L}_\xi[\delta P\boldsymbol{\epsilon}^{(n-1)}] = d[\delta P\xi\cdot\boldsymbol{\epsilon}^{(n-1)}] + \xi\cdot d[\delta P\boldsymbol{\epsilon}^{(n-1)}] = d[\delta P\boldsymbol{\epsilon}^{(n-2)}], \tag{124}$$

where the second term in the first equality vanishes because $\boldsymbol{\epsilon}^{(n-1)}$ is an $(n-1)$-form on the $(n-1)$-dimensional manifold $\mathcal{H}$, so its "d" vanishes. Thus, bringing the $\mathcal{L}_\xi\delta P$ term on the right side of (123) to the left side and using (124) and the definition (120) of $\mathbf{S}_{\rm DW}$, we obtain

$$\frac{\kappa}{2\pi}d\delta\mathbf{S}_{\rm DW} = 2\kappa\delta P(g,\xi;\mathcal{L}_\xi g)\boldsymbol{\epsilon}^{(n-1)}. \tag{125}$$

Canceling the factors of κ and contracting both sides with ξ^a, we obtain

$$\frac{1}{2\pi}\xi\cdot d\delta\mathbf{S}_{\rm DW} = 2\xi\cdot[\delta P(g,\xi;\mathcal{L}_\xi g)\boldsymbol{\epsilon}^{(n-1)}] = 2\delta P(g,\xi;\mathcal{L}_\xi g)\boldsymbol{\epsilon}^{(n-2)}. \tag{126}$$

By the general identity (80), we have

$$\xi\cdot d\delta\mathbf{S}_{\rm DW} = \mathcal{L}_\xi\delta\mathbf{S}_{\rm DW} - d[\xi\cdot\delta\mathbf{S}_{\rm DW}]. \tag{127}$$

Thus, we obtain

$$4\pi\delta P(g,\xi;\mathcal{L}_\xi g)\boldsymbol{\epsilon}^{(n-2)} = \mathcal{L}_\xi\delta\mathbf{S}_{\rm DW} - d[\xi\cdot\delta\mathbf{S}_{\rm DW}]. \tag{128}$$

By the definition (120) of $\mathbf{S}_{\rm DW}$, we have

$$\delta\mathbf{S} = \delta\mathbf{S}_{\rm DW} - \frac{4\pi}{\kappa}\delta P(g,\xi;\mathcal{L}_\xi g)\boldsymbol{\epsilon}^{(n-2)}, \tag{129}$$

and substituting in (128) we have[15]

$$\delta\mathbf{S} = \delta\mathbf{S}_{\rm DW} - \frac{1}{\kappa}\mathcal{L}_\xi\delta\mathbf{S}_{\rm DW} + \frac{1}{\kappa}d[\xi\cdot\delta\mathbf{S}_{\rm DW}]. \tag{131}$$

The last term is exact, so it does not contribute when one integrates over a cross section $\mathcal{C}$ to get the Dong-Wall entropy, $\delta S_{\rm DW}[\mathcal{C}]$, of the cross section. Thus, we obtain the desired general relationship between our entropy and the Dong-Wall entropy for an arbitrary cross section $\mathcal{C}$

$$\delta S[\mathcal{C}] = \delta S_{\rm DW}[\mathcal{C}] - \frac{1}{\kappa}\int_{\mathcal{C}}\mathcal{L}_\xi\delta\mathbf{S}_{\rm DW}. \tag{132}$$

It is worth noting that if we multiply (118) by $\boldsymbol{\epsilon}^{(n-2)}$ and use (128) we obtain

$$(\delta E_G)_{ab}\xi^a\xi^b\boldsymbol{\epsilon}^{(n-2)} \widehat{=} \frac{1}{4\pi}(\kappa-\mathcal{L}_\xi)[\mathcal{L}_\xi\delta\mathbf{S}_{\rm DW} - d(\xi\cdot\delta\mathbf{S}_{\rm DW})]. \tag{133}$$

Integrating over an arbitrary cross section $\mathcal{C}$, we obtain

$$\int_{\mathcal{C}}(\delta E_G)_{ab}\xi^a\xi^b\boldsymbol{\epsilon}^{(n-2)} = \frac{1}{4\pi}\int_{\mathcal{C}}(\kappa-\mathcal{L}_\xi)\mathcal{L}_\xi\delta\mathbf{S}_{\rm DW}. \tag{134}$$

Let V be an affine parameter for null geodesic generators of $\mathcal{H}$ with $V=0$ corresponding to the bifurcation surface. Let k^a be the corresponding affinely parametrized tangent to the null generators. Then $\xi^a = \kappa V k^a$ on $\mathcal{H}$ and on differential forms we have $\kappa\mathcal{L}_\xi - \mathcal{L}_\xi^2 = -\kappa^2V^2\mathcal{L}_k^2$ modulo exact terms. We therefore obtain

$$\int_{\mathcal{C}}V^2(\delta E_G)_{ab}k^ak^b\boldsymbol{\epsilon}^{(n-2)} = -\frac{1}{4\pi}\int_{\mathcal{C}}V^2\mathcal{L}_k^2\delta\mathbf{S}_{\rm DW}. \tag{135}$$

If $\mathcal{C}$ is a cross section of constant V (which always can be achieved for any given cross section using the rescaling freedom of V), then we obtain

$$\int_{\mathcal{C}}(\delta E_G)_{ab}k^ak^b\boldsymbol{\epsilon}^{(n-2)} = -\frac{1}{4\pi}\frac{d^2}{dV^2}\delta S_{\rm DW}(V). \tag{136}$$

This equation is the defining property[16] of the entropy given by Wall [see Eqs. (7) and (8) of [6]].

For the case of general relativity [i.e. using the Lagrangian (83)], we have

$$(\delta E_G)_{ab}k^ak^b = \frac{1}{16\pi}\delta R_{ab}k^ak^b = -\frac{1}{16\pi}\mathcal{L}_k\delta\vartheta, \tag{137}$$

where the linearized Raychaudhuri equation was used in the second equality. Thus, we have

[15]Note that we can combine (131) and (122) into the equation,

$$2\pi k^a\delta\mathbf{C}_a = d(\mathcal{L}_k\delta\mathbf{S}_{\rm DW}), \tag{130}$$

to first order, where $\xi^a = \kappa V k^a$ and V is an affine parameter vanishing on $\mathcal{B}$ such that $k^a\nabla_a V = 1$. This is identical in form to the defining equation (to first order) of the Dong-Wall entropy as given in Eq. (120) in Ref. [18] if we replace l^a there with k^a, and the entropy-current $(n-2)$-form $\boldsymbol{s}$ defined by Eq. (119) of Ref. [18] with $\boldsymbol{S}_{\rm DW}$, noting that $\boldsymbol{F}_l$ in Eq. (120) [18] vanishes up to and including first-order variations. Also note that $\boldsymbol{s}$ is defined only to first order by Eq. (120) in Ref. [18], and it is shown in Proposition 1 in Ref. [18] that it can be modified at second order if necessary so as to be covariant. By contrast, our derivation automatically delivers a covariant form of $\boldsymbol{S}_{\rm DW}$.

[16]Note that there is a missing factor of $1/2\pi$ on the left side of Eq. (7) (the first law) in [6]. Putting this factor back in and taking two derivatives in affine parameter of that equation, one obtains Eq. (8) of [6] with a factor of 2π on the left side instead of on the right side. To match with (136), note that the Wall's equations of motion tensor H_{ab} (defined in Eq. (2) in [6]) is given by $H_{ab} = 2(E_G)_{ab}$.

$$(\delta E_G)_{ab}\xi^a\xi^b = -\frac{1}{16\pi}\kappa^2 V^2 \mathcal{L}_k \delta\vartheta = \frac{1}{16\pi}(\kappa - \mathcal{L}_\xi)(\kappa V \delta\vartheta). \tag{138}$$

Comparing with (118), we obtain

$$\delta P(g, \xi; \mathcal{L}_\xi g) = \frac{1}{16\pi}\kappa V \delta\vartheta. \tag{139}$$

Substituting this result in (120), we find that for general relativity, we have

$$\delta \mathbf{S}_{\rm DW} = \delta \mathbf{S} + \frac{1}{4} V \delta\vartheta \boldsymbol{\epsilon}^{(n-2)}. \tag{140}$$

Thus, in general relativity the term arising from P in $\mathbf{S}_{\rm DW}$ cancels the term arising from $\mathbf{B}_{\mathcal{H}}$ in our formula (71) for $\mathbf{S}$, and we are left with only the contribution from $\mathbf{Q}$, which yields the Bekenstein-Hawking entropy. Thus, the Dong-Wall entropy equals the Bekenstein-Hawking entropy in general relativity.

The type of cancellation between the $\mathbf{B}_{\mathcal{H}}$ and P terms that occurs in the calculation of the Dong-Wall entropy in general relativity does not occur for higher derivative theories. In particular, for the Lagrangian

$$\mathbf{L} = R_{ab}R^{ab}\boldsymbol{\epsilon}, \tag{141}$$

the equations of motion are

$$\begin{aligned}(E_G)^{ab} = {} & R^{(a}{}_{cde}(g^{b)d}R^{ec} - R^{b)e}g^{cd}) \\ & - \nabla_d\nabla_e(g^{d(a}R^{b)e} - g^{de}R^{ab} - g^{ab}R^{de} + g^{e(a}R^{b)d}) \\ & - \frac{1}{2}R_{de}R^{de}g^{ab}.\end{aligned} \tag{142}$$

This can be put in the form

$$(\delta E_G)_{ab}\xi^a\xi^b = \kappa\mathcal{L}_\xi P(g, \xi; \delta g) - \mathcal{L}_\xi^2 P(g, \xi; \delta g) \tag{143}$$

with

$$\begin{aligned}P(g, \xi; \delta g) = {} & 2\xi^n N^m \delta R_{mn} - \frac{1}{2}g^{bd}q^{mn}\nabla_m\nabla_n\delta g_{bd} + g^{bd}q^{mn}\nabla_m\nabla_b\delta g_{nd} \\ & - R_{fmqe}N^f\xi^m N^q\xi^e g^{bd}\delta g_{bd} - g^{bd}q^{mn}R_{fbme}\xi^e N^f\delta g_{nd} - q^{mn}R^f{}_{bme}N^b\xi^e\delta g_{nf} \\ & - 2N^d\xi^f q^{mn}\nabla_m\nabla_d\delta g_{nf} + N^m\xi^f q^{bd}\nabla_b\nabla_d\delta g_{fm} - q^{pn}R_{mnpe}N^m\xi^e g^{ab}\delta g_{ab},\end{aligned} \tag{144}$$

where $q^{ab} = g^{ab} - 2\xi^{(a}N^{b)}$ is the orthogonal projector to ξ^a and N^a. In this case, it is easily seen that contribution to $\mathbf{S}_{\rm DW}$ from P does not cancel the contribution of $\mathbf{B}_{\mathcal{H}}$ to $\mathbf{S}$ [see (98) with $\alpha_1 = \alpha_3 = 0$ and $\alpha_2 = 1$].

We now evaluate the perturbed Dong-Wall entropy $\delta S_{\rm DW}[\mathcal{C}]$ for the Lagrangian (141), evaluated on an arbitrary cross section $\mathcal{C}$. Taking the Lie derivative of (144), we find, after some algebra, that $\mathcal{L}_\xi P(g, \xi; \delta g)$ can be put in the form

$$\begin{aligned}\mathcal{L}_\xi P(g, \xi; \delta g) \hat{=} {} & 2\xi^n N^m \mathcal{L}_\xi \delta R_{mn} + 2q^{mn}\nabla_m(\delta R_{nf}\xi^f) \\ & + \frac{1}{2}q^{mn}\nabla_m\nabla_n(g^{ab}\mathcal{L}_\xi\delta g_{ab}) - q^{pn}R_{mnpe}N^m\xi^e g^{ab}\mathcal{L}_\xi\delta g_{ab} + g^{ab}\mathcal{L}_\xi\delta g_{ab}R_{mn}\xi^m N^n.\end{aligned} \tag{145}$$

We now substitute (145) and (99) (with $\alpha_1 = \alpha_3 = 0$ and $\alpha_2 = 1$) into (120) and integrate the resulting $(n-2)$-form over $\mathcal{C}$. The computations needed for this are most readily done using Gaussian null coordinates based upon the cross section $\mathcal{C}$ and the Killing parametrization. The metric in these coordinates takes the form (see e.g., [36])[17]

$$ds^2 = 2dvdr - 2r\tilde{\alpha}dv^2 - 2r\beta_A dvdx^A + \gamma_{AB}dx^A dx^B \tag{146}$$

with $\xi^a = (\partial/\partial v)^a$ and $\tilde{\alpha} \hat{=} \kappa$. Here $\mathcal{H}$ corresponds to $r = 0$ and the cross section $\mathcal{C}$ of $\mathcal{H}$ corresponds to $v = 0$. We can express N^a and q^{ab} on $\mathcal{H}$ in terms of ξ^a, $l^a \equiv (\partial/\partial r)^a$, and $\beta_a = \beta_A(dx^A)_a$ as follows:

$$N^a \hat{=} l^a + \frac{1}{2\kappa}\beta^a - \frac{1}{8\kappa^2}\beta^p\beta_p\xi^a, \tag{147}$$

$$q^{ab} \hat{=} \gamma^{ab} - \frac{1}{\kappa}\xi^{(a}\beta^{b)} + \frac{1}{4\kappa^2}\beta_p\beta^p\xi^a\xi^b. \tag{148}$$

Using these relations for N^a and q^{ab} to rewrite (145), we find

[17]The Killing Gaussian null coordinates v, r in (146) are related to the affine Gaussian null coordinates V, ρ of (81) by $V = e^{\kappa v}$, $\rho = \kappa^{-1}e^{-\kappa v}r$. The quantities β_A and γ_{AB} of (146) are the same quantities that appear in (81), but $\tilde{\alpha}$ is a different quantity from α.

$$\frac{4\pi}{\kappa}\mathcal{L}_\xi P(g,\xi;\delta g)\boldsymbol{\epsilon}^{(n-2)} = 2\pi\mathcal{L}_\xi\delta[\boldsymbol{\epsilon}^{(n-2)}(4\xi^m l^c R_{mc} - g^{ab}\mathcal{L}_\xi g_{ab}\gamma^{pq}\nabla_p l_q)] + \frac{2\pi}{\kappa}D_q\Bigg(\boldsymbol{\epsilon}^{(n-2)}\Bigg[2\delta R_{nf}\xi^f\gamma^{qn} - \frac{1}{\kappa}\beta^q\xi^m\xi^f\delta R_{mf} + \frac{1}{2}D^q(g^{ab}\mathcal{L}_\xi\delta g_{ab}) - \frac{1}{2}\beta^q g^{ab}\mathcal{L}_\xi\delta g_{ab}\Bigg]\Bigg). \tag{149}$$

The last two terms on the right hand side are a total spatial divergence, which vanish when integrated over $\mathcal{C}$. The entropy $(n-2)$-form for the Lagrangian (141) is given by (99) with $\alpha_1 = \alpha_3 = 0$ and $\alpha_2 = 1$, i.e.,

$$S_{a_1\cdots a_{n-2}} = -\frac{2\pi}{\kappa}(\epsilon_{cda_1\cdots a_{n-2}}[2g^{ce}R^{df}\nabla_{[e}\xi_{f]} + (g^{cm}\nabla^d R - 2\nabla^c R^{md})\xi_m] + \epsilon^{(n-2)}_{a_1\cdots a_{n-2}}(R_{ad}\xi^a N^d g^{bc} + R^{bc})\mathcal{L}_\xi g_{bc}). \tag{150}$$

Taking a variation of this quantity, pulling it back to $\mathcal{C}$, and evaluating it in Gaussian null coordinates in a manner similar to what was done above for P, we find, after some algebra, that

$$\delta\mathbf{S} \hat{=}\ 2\pi\Bigg(\delta[\boldsymbol{\epsilon}^{(n-2)}(4R_{mn}\xi^m l^n - g^{ab}\mathcal{L}_\xi g_{ab}\gamma^{pq}\nabla_p l_q)] - \frac{1}{\kappa}\mathcal{L}_\xi\delta[\boldsymbol{\epsilon}^{(n-2)}(4R_{mn}\xi^m l^n - g^{ab}\mathcal{L}_\xi g_{ab}\gamma^{pq}\nabla_p l_q)] - \frac{2}{\kappa}D_m[\boldsymbol{\epsilon}^{(n-2)}\delta R_{dn}\xi^d\gamma^{mn}]\Bigg). \tag{151}$$

Note that the form of this equation—where the second term is minus the Lie derivative of the first term and the third term is exact—is to be expected in view of (131). The last line of (151) vanishes when integrated over $\mathcal{C}$. Adding (149) and (151) and integrating over $\mathcal{C}$, we find that the Dong-Wall entropy is

$$\delta S_{\mathrm{DW}}[\mathcal{C}] = 2\pi\delta\int_{\mathcal{C}}\boldsymbol{\epsilon}^{(n-2)}(4R_{mc}\xi^m l^c - 2K\bar{K}), \tag{152}$$

where

$$K = \frac{1}{2}g^{ab}\mathcal{L}_\xi g_{ab}, \tag{153}$$

$$\bar{K} = \gamma^{pq}\nabla_p l_q, \tag{154}$$

are the traces of the extrinsic curvatures in the ξ and l directions, respectively. This agrees with Wall's formula [6] and with the linearization of Dong's formula [5].

Finally, we note that if we integrate (125) between a cross section $\mathcal{C}_1$ and another cross section $\mathcal{C}_2$ lying to the future of $\mathcal{C}_1$, we obtain

$$\delta S_{\mathrm{DW}}[\mathcal{C}_2] - \delta S_{\mathrm{DW}}[\mathcal{C}_1] = 4\pi\int_{\mathcal{H}_{12}}\delta P\boldsymbol{\epsilon}^{(n-1)}, \tag{155}$$

where $\mathcal{H}_{12}$ denotes the portion of $\mathcal{H}$ bounded by $\mathcal{C}_1$ and $\mathcal{C}_2$ and we have written δP instead of $\delta P(g,\xi;\mathcal{L}_\xi g)$ to make the notation less cumbersome. Thus, the Dong-Wall entropy will satisfy the second law of black hole mechanics at first order if and only if $\delta P(x) \geq 0$ for all $x \in \mathcal{H}$. However, from (118), we see that if the matter stress-energy tensor satisfies the null energy condition, then we have

$$-e^{\kappa v}\mathcal{L}_\xi(e^{-\kappa v}\delta P) = (\kappa - \mathcal{L}_\xi)\delta P = (\delta E_G)_{ab}\xi^a\xi^b = \frac{1}{2}(\delta T_{ab})\xi^a\xi^b \geq 0, \tag{156}$$

where v is the Killing parameter along the null geodesics (see (39)). Thus, we have

$$\mathcal{L}_\xi(e^{-\kappa v}\delta P) \leq 0. \tag{157}$$

For any $x \in \mathcal{H}$, let γ denote the null geodesic of $\mathcal{H}$ passing through x. Integrating this equation along γ with respect to v from x to infinity and making the additional assumption that $\lim_{v\to\infty} e^{-\kappa v}\delta P = 0$, we obtain

$$\delta P(x) \geq 0. \tag{158}$$

Thus the Dong-Wall entropy also satisfies the second law of black hole mechanics. However, as with the Bekenstein-Hawking entropy as discussed at the end of Sec. V A, the Dong-Wall entropy may increase in anticipation of matter that will be thrown into the black hole at a later time.

VII. THE GENERALIZED SECOND LAW AND QNEC

The *generalized second law* (GSL) of thermodynamics states that the sum of the entropy of a black hole and the entropy of matter outside the black hole never decreases with time. The GSL was originally proposed by Bekenstein [3] in order to rescue the second law of thermodynamics in view of the fact that the total entropy of matter in the Universe will decrease if matter falls into a black hole; by assigning an entropy to a black hole, Bekenstein argued that the total (generalized) entropy would be nondecreasing

even if the entropy of matter decreases. Bekenstein's proposal was originally made within the context of classical black hole physics, where it is not consistent because, classically, black holes have vanishing thermodynamic temperature, but they would have to be assigned a nonzero temperature in order to satisfy the first law of thermodynamics with a finite entropy. However, shortly thereafter, Hawking [2] discovered that within the context of semiclassical physics, black holes radiate thermally at a temperature $T = \kappa/2\pi$. This not only removed a major inconsistency of Bekenstein's proposal but it made the GSL the only viable candidate for a thermodynamic second law, since quantum fields can violate all local energy conditions and the second law of black hole mechanics will be violated if there is a flux of negative energy into the black hole as occurs during the Hawking radiation process. Thus, matter entropy and black hole entropy can individually decrease in various circumstances; only their sum has a possibility of being nondecreasing in all circumstances.

There is considerable circumstantial evidence in support of the GSL. In particular, gedankenexperiments attempting to violate the GSL fail [37] in a manner reminiscent of the failure of Gedankenexperiments designed to violate the ordinary second law. Of course, one would like to obtain a precise formulation and proof of the GSL rather than just having circumstantial evidence in support of it. However, even in the context of nongravitational physics, there is no precise formulation and proof of the ordinary second law of thermodynamics. Rather, there is a basic argument for the ordinary second law, which traces back to Boltzmann. This argument involves the introduction of a coarse-grained observable to define the entropy of matter. In the context of classical physics, the entropy of a state is defined to be the logarithm of the volume of phase space corresponding to the value of the coarse-grained observable in that state. If the entropy of a state is not at its maximum value, dynamical evolution is very likely to evolve that state to a state of higher entropy. If the entropy of a state is at its maximum, it is very likely to remain in a maximum entropy state for much longer than any realistic observation time. Thus, this argument suggests that if one starts a system in a special (i.e., a low entropy) state, one will observe an entropy increase until it reaches an equilibrium (i.e., maximum entropy) state, at which point one will observe no further changes. A similar argument applies in the quantum context, with the phase space volume of the value of the coarse-grained observable replaced by the dimension of the eigensubspace of the eigenvalue of the coarse-grained observable [25]. It should be emphasized that the von Neumann entropy of the quantum state does not play any role in this argument for justifying the validity of the ordinary second law in the quantum context [25]. Indeed, for any isolated system in any quantum state, the von Neumann entropy is constant, but one would still expect to observe an entropy increase in such a system if one started it in a special state.

It is therefore quite remarkable that a precise formulation and argument/proof of the GSL has been given for semiclassical general relativity, using von Neumann entropy as the notion of entropy of matter outside the black hole. In particular, Wall [38] has shown that

$$\frac{d}{dV}\left(\frac{A}{4} + S_{\mathrm{vN}}\right) \geq 0. \tag{159}$$

Here $A(V)/4$ is the Bekenstein-Hawking entropy of the black hole on the cross section $\mathcal{C}$ corresponding to affine time V and $S_{\mathrm{vN}}(V)$ is the von Neumann entropy of the quantum field matter outside the black hole at a "time" corresponding to a spacelike hypersurface extending from $\mathcal{C}$ to spatial infinity. In fact, $S_{\mathrm{vN}}(V)$ is ill-defined on account of the infinite entanglement of the quantum field across the horizon. However, von Neumann entropy differences between states should be well defined, and that is what should be needed to make sense of dS_{vN}/dV in the GSL (see Ref. [39] for a recent proposal on how to rigorously define entropy differences in quantum field theory via a cross product construction).

In this section, we will investigate the validity of the GSL for first-order perturbations of a stationary black hole in a general theory of gravity, using both our definition of entropy, $S[\mathcal{C}]$, and the Dong-Wall definition, $S_{\mathrm{DW}}[\mathcal{C}]$. Our analysis will be based upon the QNEC for quantum fields in fixed classical background, first introduced in [40]. QNEC asserts that if $\mathcal{C}$ is a cross section of a null hypersurface $\mathcal{N}$ such that the expansion and shear of $\mathcal{N}$ vanish on $\mathcal{C}$, then for any point p on $\mathcal{C}$ we have [40]

$$\sqrt{h}\langle T_{ab}\rangle k^a k^b \geq \frac{1}{2\pi}\frac{\delta^2}{\delta V(p)^2} S_{\mathrm{vN}}, \tag{160}$$

where h is the determinant of the induced metric on $\mathcal{C}$, $V(y)$ is the affine parameter (a function of the coordinates in the transverse direction y), and S_{vN} is the von Neumann entropy of the quantum field outside the null hypersurface (where "outside" means on a spacelike hypersurface whose future Cauchy horizon coincides with $\mathcal{N}$). For simplicity, we will restrict consideration here to the case where the deformation of $\mathcal{C}$ is a uniform translation in V. Integration of (160) over $\mathcal{C}$ yields

$$\int_{\mathcal{C}} \langle T_{ab}\rangle k^a k^b \geq \frac{1}{2\pi}\frac{d^2}{dV^2} S_{\mathrm{vN}}. \tag{161}$$

QNEC has been rigorously proven/formulated in [41] within the general framework of half-sided modular inclusions and relative entropy, and this setting is expected to cover the case of a bifurcate Killing horizon in curved spacetime as considered here. We omit a more precise

technical discussion of the setting and, for purposes of readability, we stick to informal notations such as (161). In the remainder of this section, we will assume that (161) holds for an arbitrary cross section $\mathcal{C}$ of a bifurcate Killing horizon.

Equation (161) holds in a fixed background spacetime. We assume that for an unperturbed stationary black hole with a quantum field in a Hartle-Hawking-like vacuum state, both sides of (161) vanish. We now consider perturbing the quantum state, taking back-reaction into account by treating $\delta\langle T_{ab}\rangle$ as an "external matter perturbation" as in Sec. V A. Then we have

$$2\delta(E_G)_{ab}k^a k^b = \delta\langle T_{ab}\rangle k^a k^b. \tag{162}$$

Consequently, when back-reaction is taken into account, QNEC in the form (161) yields

$$\int_{\mathcal{C}} 2\delta(E_G)_{ab}k^a k^b \geq \frac{1}{2\pi}\frac{d^2}{dV^2}\delta S_{\rm vN}. \tag{163}$$

The GSL for the Dong-Wall definition of black hole entropy can now be obtained as follows. Using (136), we obtain

$$-\frac{1}{2\pi}\frac{d^2}{dV^2}\delta S_{\rm DW} \geq \frac{1}{2\pi}\frac{d^2}{dV^2}\delta S_{\rm vN}, \tag{164}$$

i.e.,

$$\frac{d^2}{dV^2}[\delta S_{\rm DW} + \delta S_{\rm vN}] \leq 0. \tag{165}$$

Integrating this equation from an initial cross section to $V = \infty$ and making the additional assumption that $\lim_{V\to\infty} d\delta S_{\rm DW}/dV = \lim_{V\to\infty} d\delta S_{\rm vN}/dV = 0$, we obtain the desired GSL

$$\frac{d}{dV}[\delta S_{\rm DW} + \delta S_{\rm vN}] \geq 0. \tag{166}$$

This derivation corresponds closely to Wall's derivation of the GSL for general relativity [38], although he based his argument on the monotonicity of the relative entropy rather than QNEC. The above argument shows that this form of the GSL corresponds to an integrated form of QNEC.

The GSL for our definition of black hole entropy can be obtained by evaluating the physical process version of the first law (104) on cross sections $\mathcal{C}$ of constant V and differentiating with respect to V. We obtain

$$\frac{\kappa}{2\pi}\frac{d}{dV}\delta S = \int_{\mathcal{C}} \delta\langle T_{ab}\rangle k^a \xi^b = \kappa V \int_{\mathcal{C}} \delta\langle T_{ab}\rangle k^a k^b \geq \frac{\kappa}{2\pi} V \frac{d^2}{dV^2}\delta S_{\rm vN}. \tag{167}$$

This relation can be put in a GSL form by defining a *modified von Neumann entropy* of matter, $\tilde{S}_{\rm vN}$, by

$$\tilde{S}_{\rm vN} \equiv S_{\rm vN} - V\frac{d}{dV}S_{\rm vN}. \tag{168}$$

Thus, $\tilde{S}_{\rm vN}$ agrees with $S_{\rm vN}$ in stationary eras but has a dynamical correction when $S_{\rm vN}$ changes with time. Since

$$\frac{d}{dV}\tilde{S}_{\rm vN} = -V\frac{d^2}{dV^2}S_{\rm vN}, \tag{169}$$

we see that (167) takes the GSL form[18]

$$\frac{d}{dV}[\delta S + \delta\tilde{S}_{\rm vN}] \geq 0. \tag{170}$$

This version of the GSL directly corresponds to QNEC, rather than an integrated form of QNEC.

ACKNOWLEDGMENTS

We thank Tom Faulkner, Igor Klebanov, Harvey Reall, and Jon Sorce for discussions. S. H. and R. M. W. thank the Erwin Schrödinger Institut in Vienna, where part of this work has been completed, for its hospitality and support. We thank Thomas Endler from MPI-MiS for creating Fig. 1 based on our suggestions and for giving us permission to include it in this publication. S. H. is grateful to the Max-Planck Society for supporting the collaboration between MPI-MiS and Leipzig U., Grant Project No. Bez. M.FE.A.MATN0003. The research of R. M. W. and V. G. Z. was supported in part by NSF Grant No. PHY-2105878 and Templeton Foundation Grant No. 62845 to the University of Chicago.

APPENDIX A: S IS THE AREA OF THE APPARENT HORIZON TO FIRST ORDER IN GENERAL RELATIVITY

The apparent horizon on an arbitrary Cauchy surface Σ is the boundary of the region containing outer trapped surfaces that lie within Σ. The apparent horizon is a marginally outer trapped surface, i.e., its outgoing null expansion vanishes. If the null energy condition holds and if weak cosmic censorship is valid, then an apparent horizon must lie within a black hole, so apparent horizons

[18]The idea of defining a modified von Neumann entropy by (168) so as to be able to write (167) in the form (170) was suggested to us by Jon Sorce. It would be interesting to see if this definition of modified von Neumann entropy can be motivated by additional, independent arguments. In this regard, Igor Klebanov has pointed out to us that the expression $V\frac{d}{dV}S_{\rm vN}$ is formally equal to the monotonic c-function for two-dimensional CFTs defined by [42]. The lightray algebras of a Killing horizon have connections with a chiral CFT, so this might be more than a coincidence.

are very useful for determining the presence of a black hole given only data on Σ. For a stationary black hole, the apparent horizon on any Cauchy surface Σ is given by the intersection of Σ with the event horizon $\mathcal{H}$.

For a nonstationary black hole, the location of an apparent horizon will depend on the choice of Cauchy surface, and the area of an apparent horizon at a time (i.e., choice of Cauchy surface) corresponding to a cross section $\mathcal{C}$ of the horizon will not, in general, be a meaningful concept. However, for first-order perturbations of a stationary black hole with bifurcate Killing horizon, the notion of the area of an apparent horizon at a time corresponding to $\mathcal{C}$ is a well-defined concept. To see this, we note that, as proven in Sec. 2.1 of [8], for a stationary black hole with bifurcate Killing horizon, any horizon cross section $\mathcal{C}$ is a strictly stably outermost marginally trapped surface in the terminology of [43]. This implies that given any function, f, on $\mathcal{C}$, there exists a unique function, ψ on $\mathcal{C}$ such that if we deform $\mathcal{C}$ infinitesimally by ψl^a (with l^a a fixed vector field tranverse to the horizon), then to first order, the outgoing expansion of $\delta\mathcal{C}$ will be f. Now suppose we perturb the black hole so that its perturbed expansion at $\mathcal{C}$ is positive, $\delta\vartheta > 0$. To find the apparent horizon at first order, we infinitesimally perturb $\mathcal{C}$ inwards by ψl^a and denote the first order change of the outgoing null-expansion ϑ by $\delta_{\psi l}\vartheta$. If we now choose ψ so that $\delta_{\psi l}\vartheta = -\delta\vartheta$, then to first order, this uniquely locates an apparent horizon lying along the flow of ψl^a from $\mathcal{C}$. To first order, the area of the apparent horizon will not depend on the choice of l^a.

The computation of the location of the apparent horizon is most conveniently done using Gaussian null coordinates based upon the cross section $\mathcal{C}$ and the Killing parametrization, as previously introduced in (146). We take $l^a = (\partial/\partial r)^a$, so l^a is the (past directed) ingoing null normal to $\mathcal{C}$, normalized so that $l^a\xi_a = 1$. The perturbed expansion of $\mathcal{C}$ along the horizon direction with respect to ξ^a is then $\delta\tilde{\vartheta} = \kappa V\delta\vartheta$. The equation for ψ that we must solve is

$$\delta_{\psi l}\tilde{\vartheta} = -\delta\tilde{\vartheta} = -\kappa V\delta\vartheta. \tag{A1}$$

The variation $\delta_{\psi l}\tilde{\vartheta} =: W\psi$ is given by a linear operator W intrinsic to $\mathcal{C}$ acting on ψ which has been given in [43] (see also Eqs. (5.12) and (5.34) of [44]) as[19]

$$W\psi = -D^aD_a\psi + \beta^aD_a\psi + \left(G_{ab}\xi^al^b + \frac{1}{2}\left(R[\gamma] - \frac{1}{2}\beta^a\beta_a + D_a\beta^a\right) - \tilde{\vartheta}\bar{\tilde{\vartheta}}\right)\psi, \tag{A2}$$

where $R[\gamma]$ denotes the scalar curvature of the induced derivative D_a on $\mathcal{C}$ associated with the induced metric γ_{ab}, $\tilde{\vartheta} = \frac{1}{2}\gamma^{cd}\mathcal{L}_\xi\gamma_{cd}$ is the outgoing expansion of $\mathcal{C}$, and $\bar{\tilde{\vartheta}} = \frac{1}{2}\gamma^{cd}\mathcal{L}_l\gamma_{cd}$ the ingoing expansion of $\mathcal{C}$.

We now use Einstein's equation $G_{ab} = 0$ and the fact that $\mathcal{H}$ has vanishing outgoing shear and expansion in the stationary background. Using the fact that $\tilde{\alpha} \widehat{=} \kappa$ and using Einstein's equation again, the scalar curvature of $\mathcal{C}$ is found to be [see also Eq. (82) of [36]]

$$R[\gamma] = D_a\beta^a + \frac{1}{2}\beta^a\beta_a + 2\kappa\bar{\tilde{\vartheta}}. \tag{A3}$$

Thus, ψ is determined by

$$-D^aD_a\psi + \beta^aD_a\psi + (D_a\beta^a)\psi + \kappa\bar{\tilde{\vartheta}}\psi = -\kappa V\delta\vartheta. \tag{A4}$$

The second and third terms on the left side combine to a total divergence $D_a(\psi\beta^a)$ intrinsic to $\mathcal{C}$. Integrating over $\mathcal{C}$, we see that ψ satisfies the relation

$$\kappa\int_{\mathcal{C}}\bar{\tilde{\vartheta}}\psi\boldsymbol{\epsilon}^{(n-2)} = -\kappa\int_{\mathcal{C}}V\delta\vartheta\boldsymbol{\epsilon}^{(n-2)}. \tag{A5}$$

On the other hand, the change in area resulting from displacing $\mathcal{C}$ infinitesimally along ψl^a is

$$\delta A = \int_{\mathcal{C}}\psi\bar{\tilde{\vartheta}}\boldsymbol{\epsilon}^{(n-2)}. \tag{A6}$$

Thus, we see that the area of the apparent horizon differs from the area of $\mathcal{C}$ by

$$\delta A = \int_{\mathcal{C}}\psi\bar{\tilde{\vartheta}}\boldsymbol{\epsilon}^{(n-2)} = -\int_{\mathcal{C}}V\delta\vartheta\boldsymbol{\epsilon}^{(n-2)}. \tag{A7}$$

Comparing with (89), we see that up to first order, we have

$$S[\mathcal{C}] = \frac{A_{\mathrm{app}}[\mathcal{C}]}{4} \tag{A8}$$

where $A_{\mathrm{app}}[\mathcal{C}]$ is the area of the apparent horizon at the time corresponding to the cross section $\mathcal{C}$ of the horizon.

APPENDIX B: GENERALIZATION OF THEOREM 1 TO THE CASE OF A TENSOR DOTTED WITH TWO (OR MORE) ξ's

Consider the hypotheses of Theorem 1, but with α^a in (51) replaced by

[19]The differences between formula (A2) for our operator defined by $\delta_{\psi l}\tilde{\vartheta} =: W\psi$ and the corresponding Eq. (1) of [43] result from the facts that (i) we are considering a null displacement rather than a spacelike displacement and (ii) they set $\tilde{\vartheta} = 0$ since they are considering a marginally outer trapped surface. Equation (A2) is valid for the change in the expansion of the orthogonal null geodesics of any $n-2$-dimensional surface $\mathcal{C}$ under an infinitesimal displacement of the surface by ψl^a, where l^a is the other null normal field to $\mathcal{C}$, normalized to have unit inner product with the null normal defining the expansion.

$$\beta^{ab} = \sum_{i=0}^{k} T_{(i)}^{abc_1\cdots c_i de} \nabla_{(c_1} \cdots \nabla_{c_i)} \delta g_{de}. \tag{B1}$$

We wish to obtain a strengthening of the conclusion (52) of Theorem 1 that is valid for the form of $\beta_{ab}\xi^a\xi^b$. Specifically, we will show that $\beta_{ab}\xi^a\xi^b$ can be written in the form

$$\beta_{ab}\xi^a\xi^b \widehat{=} \sum_{i=0}^{k-2} \tilde{T}_{(i)}^{c_1\cdots c_i de} \nabla_{(c_1} \cdots \nabla_{c_i)} (\mathcal{L}_\xi - \kappa) \mathcal{L}_\xi \delta g_{de} \tag{B2}$$

where the tensors $\tilde{T}_{(i)}^{b_1\cdots b_i cd}$ are smooth on $\mathcal{H}$ and are locally and covariantly constructed from the metric, curvature, and its derivatives as well as from ξ^a, and N^a, with ξ^a and N^a appearing only algebraically.

To show this, we start with the highest derivative term in $\beta_{ab}\xi^a\xi^b$ and proceed as in the proof of Theorem 1. Each term in the basis expansion of $\xi_a\xi_b T_{(k)}^{abc_1\cdots c_i de}$ corresponding to (53) will now have at least two "extra" ξ's, i.e., there will be at least 2 more ξ's than N's in each term. We can perform the exactly the same manipulations with one of these "extra" ξ's as was done to derive (61) in the proof of Theorem 1. We thereby obtain

$$\begin{aligned} &\xi_a\xi_b T_{(k)}^{abc_1\cdots c_k de} \nabla_{(c_1} \cdots \nabla_{c_k)} \delta g_{de} \\ &\widehat{=} \tilde{T}_{(k-1)}^{c_1\cdots c_{k-1} de} \nabla_{(c_1} \cdots \nabla_{c_{k-1})} \mathcal{L}_\xi \delta g_{de} \\ &\quad + \sum_{i=0}^{k-1} T'^{c_1\cdots c_i de}_{(i)} \nabla_{(c_1} \cdots \nabla_{c_i)} \delta g_{de} \end{aligned} \tag{B3}$$

with $\tilde{T}_{(k-1)}^{c_1\cdots c_{k-1} de}$ and $T'^{c_1\cdots c_i de}_{(i)}$ satisfying the same properties as enumerated in Theorem 1 but with $\tilde{T}_{(k-1)}^{c_1\cdots c_{k-1} de}$ now satisfying the additional property that there is at least one "extra" ξ^a appearing in each term in its basis expansion. We may now perform the same manipulations with this additional ξ^a to obtain

$$\begin{aligned} &\xi_a\xi_b T_{(k)}^{abc_1\cdots c_k de} \nabla_{(c_1} \cdots \nabla_{c_k)} \delta g_{de} \\ &\widehat{=} \tilde{T}_{(k-2)}^{c_1\cdots c_{k-2} de} \nabla_{(c_1} \cdots \nabla_{c_{k-2})} \mathcal{L}_\xi^2 \delta g_{de} \\ &\quad + \sum_{i=0}^{k-1} T''^{c_1\cdots c_i de}_{(i)} \nabla_{(c_1} \cdots \nabla_{c_i)} \delta g_{de} \end{aligned} \tag{B4}$$

where $T''^{c_1\cdots c_i de}_{(i)}$ incorporates all of the additional lower derivative terms arising from this second set of manipulations. Now, the left side of (B4) is $O(V^2)$ as $V \to 0$, where V denotes an affine parameter on the horizon (with $V = 0$ corresponding to the bifurcation surface). However, the first term on the right side is only $O(V)$ since

$$\begin{aligned} &\tilde{T}_{(k-2)}^{c_1\cdots c_{k-2} de} \nabla_{(c_1} \cdots \nabla_{c_{k-2})} \mathcal{L}_\xi^2 \delta g_{de} \\ &= \mathcal{L}_\xi^2 [\tilde{T}_{(k-2)}^{c_1\cdots c_{k-2} de} \nabla_{(c_1} \cdots \nabla_{c_{k-2})} \delta g_{de}] \\ &= \kappa V \mathcal{L}_k (\kappa V \mathcal{L}_k [\tilde{T}_{(k-2)}^{c_1\cdots c_{k-2} de} \nabla_{(c_1} \cdots \nabla_{c_{k-2})} \delta g_{de}]) \\ &= \kappa^2 (V^2 \mathcal{L}_k^2 + V \mathcal{L}_k) [\tilde{T}_{(k-2)}^{c_1\cdots c_{k-2} de} \nabla_{(c_1} \cdots \nabla_{c_{k-2})} \delta g_{de}] \end{aligned} \tag{B5}$$

and the factor with $V\mathcal{L}_k$ is only $O(V)$. We can rectify this by replacing $\mathcal{L}_\xi^2 \delta g_{de}$ in (B4) by $[\mathcal{L}_\xi^2 - \kappa\mathcal{L}_\xi]\delta g_{de}$ and compensating for this by adding the quantity $\tilde{T}_{(k-2)}^{c_1\cdots c_{k-2} de} \nabla_{(c_1} \cdots \nabla_{c_{k-2})} \kappa \mathcal{L}_\xi \delta g_{de}$ to the second term on the right side of (B4). Since this quantity involves at most $(k-1)$ derivatives of δg_{de}, it does not change the character of this term. Thus, have shown that

$$\begin{aligned} &\xi_a\xi_b T_{(k)}^{abc_1\cdots c_k de} \nabla_{(c_1} \cdots \nabla_{c_k)} \delta g_{de} \\ &\widehat{=} \tilde{T}_{(k-2)}^{c_1\cdots c_{k-2} de} \nabla_{(c_1} \cdots \nabla_{c_{k-2})} (\mathcal{L}_\xi - \kappa) \mathcal{L}_\xi \delta g_{de} \\ &\quad + \sum_{i=0}^{k-1} T'''^{c_1\cdots c_i de}_{(i)} \nabla_{(c_1} \cdots \nabla_{c_i)} \delta g_{de} \end{aligned} \tag{B6}$$

where $\tilde{T}_{(k-2)}$ and $T'''_{(i)}$ are smooth and are locally and covariantly constructed from the metric, curvature, ξ^a, and N^a, with only algebraic dependence on ξ^a and N^a. But now the left side and the first term on the right side are both $O(V^2)$ as $V \to 0$. Therefore, the last term on the right side is also $O(V^2)$ and therefore $T'''^{c_1\cdots c_i de}_{(k-1)}$ must also have two "extra" ξ's in each term of its basis expansion. This allows us to make an inductive argument to prove (B2) in parallel with the proof of Theorem 1.

Remark. For the case of a tensor

$$\gamma^{a_1\cdots a_p} = \sum_{i=0}^{k} T_{(i)}^{a_1\cdots a_p c_1\cdots c_i de} \nabla_{(c_1} \cdots \nabla_{c_i)} \delta g_{de}, \tag{B7}$$

contracted into p factors of ξ^a, the corresponding result is

$$\gamma_{a_1\cdots a_p}\xi^{a_1}\cdots\xi^{a_p} \widehat{=} \sum_{i=0}^{k-p} \tilde{T}_{(i)}^{c_1\cdots c_i de} \nabla_{(c_1} \cdots \nabla_{c_i)} (\mathcal{L}_\xi - \kappa(p-1)) \cdots (\mathcal{L}_\xi - 2\kappa)(\mathcal{L}_\xi - \kappa)\mathcal{L}_\xi \delta g_{de} \tag{B8}$$

where the tensors $\tilde{T}_{(i)}^{c_1\cdots c_i de}$ are smooth on $\mathcal{H}$ and are locally and covariantly constructed from the metric, curvature, and its derivatives as well as from ξ^a, and N^a, with ξ^a and N^a appearing only algebraically.

APPENDIX C: AMBIGUITIES OF S BEYOND FIRST ORDER

1. Nondynamical background structures

As discussed in Sec. IVA, the definition of S suffers from ambiguities at second order and beyond. One source of ambiguities arises from the presence of ξ^a in our formula for $\mathbf{B}_{\mathcal{H}}$ and the fact that, going beyond linear order in perturbation theory, we can no longer treat ξ^a as the exact Killing field on the stationary background. The approach taken in Sec. IVA was to tie the metric g_{ab} to ξ^a by postulating in effect that g_{ab} is in a Gaussian null gauge (81) and is related to ξ^a by (82), in such a way that $\mathcal{B}$ is at $\rho = 0 = V$.

In this section, we will present another possible approach wherein the class of metrics g_{ab} considered (i.e. effectively the gauge near $\mathcal{H}$), as well as the form of ξ^a off of $\mathcal{H}$ is determined by a nondynamical background structure given only *on* $\mathcal{H}$, as opposed to *off of* $\mathcal{H}$ as in (82). In particular, off of $\mathcal{H}$, though not on $\mathcal{H}$, the ξ^a considered in this appendix will depend on g_{ab}, unlike in Sec. IVA.

Our main reason for considering a different approach here is that with the approach given below, it will be possible to make a precise comparison between the ambiguities and structure of S in our approach and that of recent proposal by [18] (see Sec. C 3). In fact, while [18] also works in a Gaussian null gauge, just as we did to tie the gauge of g_{ab} to ξ^a in Sec. IVA, the gauge considered by [18] is tied to the cross section $\mathcal{C}$ on which one wants to define $S[\mathcal{C}]$. By contrast, in the present paper we have ξ^a which is not tied from the outset to a particular $\mathcal{C}$ on which we anticipate evaluating S.

In order to characterize the ambiguities in S precisely, it is necessary to specify the geometric structures underlying our construction. The rigidly fixed, nondynamical background structure considered in this appendix consists of (a) an $(n-1)$-dimensional hypersurface $\mathcal{H}$ of $\mathcal{M}$ having the topological structure of a trivial fiber bundle with fibers $\mathbb{R}$ and compact base; (b) a preferred cross section, $\mathcal{B}$, on $\mathcal{H}$; (c) a vector field ξ^a tangential to the fibers which changes its direction at $\mathcal{B}$ (and vanishes only on $\mathcal{B}$); (d) a positive real number κ; (e) a vector field N^a transverse to $\mathcal{H}$ that is defined for each point in $\mathcal{H}\backslash\mathcal{B}$, and is nonzero where defined.

Definition 1. Given our rigidly fixed, nondynamical background structure $(\mathcal{H}, \mathcal{B}, \xi^a, N^a, \kappa)$, we consider compatible spacetime metrics g_{ab} on $\mathcal{M}$ in the following sense. We require that $\mathcal{H}$ is null with normal ξ^a and, furthermore, is such that $\xi^a\nabla_a\xi^b = \kappa\xi^b$, $N^aN_a = 0$, and $g_{ab}\xi^bN^a = 1$ on $\mathcal{H}$. We call this space of metrics $\mathcal{P}$. $\boldsymbol{B}_{\mathcal{H}}$, as well as $\boldsymbol{S}$ are functionals on $\mathcal{P}$.

The space $\mathcal{P}$ does not imply any physical restrictions onto the metric. Indeed, let $\tilde{\mathcal{H}}$ be any null surface with compact cross sections ruled by affinely parametrized null geodesics with tangent k^a. Let V be any affine parameter along the null geodesics vanishing on some cut, $\mathcal{B}$ and set $\tilde{\xi}^a = \kappa V k^a$. Then we have $\tilde{\xi}^a\nabla_a\tilde{\xi}^b = \kappa\tilde{\xi}^b$ and $\tilde{\xi}^a$ is obviously null and normal to $\tilde{\mathcal{H}}$. Furthermore, let $\tilde{N}^a$ be a second null field defined on $\tilde{\mathcal{H}}$ such that $\tilde{\xi}^a\tilde{N}_a = 1$. By applying a diffeomorphism ϕ, we can clearly achieve that $\phi^*\tilde{N}^a = N^a, \phi^*\tilde{\xi}^a = \xi^a$ and $\phi[\tilde{\mathcal{H}}] = \mathcal{H}$. Thus, ϕ^*g_{ab} is in $\mathcal{P}$.

Due to the derivatives acting on ξ^a in (69), we first define N^a off of $\mathcal{H}$ by demanding that, besides relation (C1) and $N^a\xi_a = 1, N^aN_a = 0$ on $\mathcal{H}$, we have $N^a\nabla_aN^b = 0$. Since ξ^a is hypersurface orthogonal, we have $\nabla_{[a}\xi_{b]} \hat{=} 2\kappa w_{[a}\xi_{b]}$ for some w_a. In fact, we can and will adjust the extension such that $w_a = N_a$. To see this, note that we are allowed to change $\xi^a \to \xi^a + fp^a$ where f is any function on $\mathcal{M}$ vanishing on $\mathcal{H}$ and p^a is any vector field. Since ξ^a is normal to $\mathcal{H}$, we must have $\nabla_af = c\xi_a$ on $\mathcal{H}$, where c is a function that we can choose as we please. This means that $w_a \to w_a - cp_a/\kappa$ which we use to change w_a to N_a.

Our requirement $\nabla_{[a}\xi_{b]} \hat{=} 2\kappa N_{[a}\xi_{b]}$ fixes the first derivative of ξ^a on $\mathcal{H}$ only partially. Note that we are not free to fix the symmetrized derivative of ξ^a e.g. by (33) since that relation used the Killing property of ξ^a, and we are no longer requiring at this stage that ξ^a is a Killing vector field with respect to g_{ab}. However, we may, and will, demand without loss of generality that ξ^a has been extended off of $\mathcal{H}$ such that

$$\nabla_{[a}\xi_{b]} \hat{=} 2\kappa N_{[a}\xi_{b]}, \quad (\nabla_a\xi_b)N^aN^b \hat{=} 0, \quad (\nabla_{(a}\xi_{b)})\xi^b \hat{=} 0. \tag{C1}$$

We are allowed to change $\xi^a \to \xi^a + f\xi^a$, where f is any function on $\mathcal{M}$ vanishing on $\mathcal{H}$, without affecting the first relation in (C1). Since $\nabla_af = c\xi_a$ on $\mathcal{H}$, we can thereby change $\nabla_{(a}\xi_{b)} \to \nabla_{(a}\xi_{b)} + c\xi_a\xi_b$. Since c is a function that we can choose as we please and since $N^a\xi_a = 1$, we can clearly impose $(\nabla_a\xi_b)N^aN^b \hat{=} 0$. The condition $(\nabla_{(a}\xi_{b)})\xi^b \hat{=} 0$ easily follows from $\nabla_{[a}\xi_{b]} \hat{=} 2\kappa N_{[a}\xi_{b]}$ and $\xi^a\nabla_a\xi^b \hat{=} \kappa\xi^b$. Finally, using $N^a\nabla_aN^b = 0$ and $(\nabla_{(a}\xi_{b)})\xi^b \hat{=} 0 \hat{=} (\nabla_a\xi_b)N^aN^b$, we have

$$N^a\nabla_a(N^b\xi_b) \hat{=} 0, \qquad \nabla_a(\xi^b\xi_b) \hat{=} -2\kappa\xi_a. \tag{C2}$$

Transvecting $\nabla_{[a}\xi_{b]} \hat{=} 2\kappa N_{[a}\xi_{b]}$ with N^a, we get $\mathcal{L}_N\xi_a \hat{=} -2\kappa N_a$, which clearly uniquely determines the first-order change of ξ^a off of $\mathcal{H}$.

As we have argued, (C1) defines ξ^a to first order off of $\mathcal{H}$, but we would like to define it to arbitrary orders off of $\mathcal{H}$. In order to have a definition that gives us the Killing vector field ξ^a in the special case that $\mathcal{H}$ is a Killing horizon for g_{ab}, we require ξ^a to satisfy the geodesic deviation equation,

$$N^a\nabla_a(N^b\nabla_b\xi_c) = R_{cbad}N^bN^a\xi^d, \tag{C3}$$

which is a viewed as a second-order ordinary differential equation for ξ^a whose initial conditions are posed on $\mathcal{H}$. Since the geodesic deviation equation implies that $N^a\nabla_a(N^b\nabla_b(\xi^c N_c)) = 0$, the initial conditions $\xi^a N_a \widehat{=} 1$, $N^c\nabla_c(N^a\xi_a) \widehat{=} 0$, imply that $N^a\xi_a = 1$ in a neighborhood of $\mathcal{H}$.

Consider now a metric $\tilde{g}_{ab}$ having a stationary bifurcate Killing horizon $\mathcal{H}$ with Killing field $\tilde{\xi}^a$ that vanishes on $\mathcal{B}$. We have constructed in Eqs. (33) to (38) a vector field $\tilde{N}^a$ at points of $\mathcal{H}$ from the Killing field $\tilde{\xi}^a$ and the metric $\tilde{g}_{ab}$. We can align these with ξ^a, N^a as in definition 1 by a diffeomorphism ϕ, so that $g_{ab} = \phi^*\tilde{g}_{ab} \in \mathcal{P}$. Since any Killing field satisfies (63), we automatically have $N^a\nabla_a(N^b\nabla_b\xi_c) = R_{cbad}N^bN^a\xi^d$, where the derivatives and Riemann tensor refer to g_{ab}. Thus, our extension procedure produces precisely the horizon Killing field if there is one.

We likewise require that any variation δg_{ab} be compatible with our the fixed (not varying) background structure $(\mathcal{H}, \mathcal{B}, \xi^a, N^a, \kappa)$ in this appendix, in the sense that it is given as the derivative of a 1-parameter family of metrics in the space $\mathcal{P}$ (definition 1), i.e. as an element in the tangent space of $\mathcal{P}$. Then we have $\delta g_{ab}\xi^a = 0$ on $\mathcal{H}$. Furthermore, since $\delta\kappa = 0$ as κ is part of the rigidly fixed background structure, the last equation in (C2) implies that $0 = \delta[\nabla_a(\xi^b\xi_b)] = \nabla_a(\delta g_{bc}\xi^c\xi^b)$. Thus, the previously considered gauge conditions (42), (43) hold generally, not just off a stationary (bifurcate Killing-) horizon.

Our main first-order variation formula for S was (76) for any perturbation δg_{ab} satisfying gauge conditions (42), (43). In the construction leading to (76), we previously considered ξ^a as fixed not only on $\mathcal{H}$, but also off of $\mathcal{H}$. By contrast, the algorithm we have just given above for extending ξ^a, N^a off of $\mathcal{H}$ depended on the chosen metric, and therefore while $\delta\xi^a \widehat{=} 0$, the derivatives of $\delta\xi^a$ off of $\mathcal{H}$ will not in general vanish. Nevertheless, by a lengthy argument, we have shown that (76) continues to hold. As a consequence we still have the formulas expressing the second law obtained within the previous framework of nondynamical background structure in Secs. V A and V B. Similarly, we still have the second-order variation formulas (111) and (112) in the present setting with a varying ξ^a, with only minor modifications to the definition of the boundary terms in the modified canonical energy (112). However, we shall not present the details of these modifications here.

2. Structure and ambiguities of S

As in the approach taken in Sec. IV A, there is substantial potential ambiguity in S beyond first order. The first ambiguity arises because it may happen that two metrics g_{ab} and ϕ^*g_{ab} may both be in $\mathcal{P}$, where ϕ is a nontrivial diffeomorphism fixing $\mathcal{H}$ and hence ξ^a on $\mathcal{H}$. We therefore have $(\phi^*g_{ab})\xi^aN^b = 1$ by definition 1, so g_{ab} satisfies the definition 1 also for $\tilde{N}^a = \phi^{-1*}N^a$. Since g_{ab} and ϕ^*g_{ab} must be regarded as physically equivalent, this is telling us that we have, in effect, changed $N^a \to \tilde{N}^a$ while keeping g_{ab} fixed, analogous to the freedom of choosing a different affine parametrization in of $\mathcal{H}$ in Sec. IV A.

The second ambiguity arises from the fact that $\mathbf{B}_\mathcal{H}$, which enters S, was constructed to satisfy (70), which is a condition that applies only to first-order perturbations. We may add many terms to $\mathbf{B}_\mathcal{H}$ (and hence to $\mathbf{S} = \mathbf{Q} - \xi\cdot\mathbf{B}_\mathcal{H}$) that vanish identically at first order but not second. The following lemma will enable us to give a precise classification of such terms.

Lemma 1. Let $\mathbf{B}_\mathcal{H}$ be an $(n-1)$-form valued functional satisfying 1.–5. in Sec. IV A. Then it can be written as a sum of monomials of the following factors (i)—(iii), contracted with N^a, ξ^a, s_i^a's and multiplied by powers of κ.

(i) $s_{i_1}^{a_1}\cdots s_{i_j}^{a_j}\nabla_{(a_1}\cdots\nabla_{a_j)}\mathcal{L}_\xi g_{ab}$;

(ii) $s_{i_1}^{a_1}\cdots s_{i_j}^{a_j}\nabla_{(a_1}\cdots\nabla_{a_{j-1})}(\mathcal{L}_\xi)^s N_{a_j}$, $s \geq 0$;

(iii) $\nabla_{(a_1}\cdots\nabla_{a_j}R^b{}_{a_{j+1}a_{j+2})}{}^c$, $j > 0$, R_{abcd}.

Each contracted monomial in $\mathbf{B}_\mathcal{H}$ is invariant under a rescaling $\xi^a \to c\xi^a$, $N^a \to N^a/c$, $\kappa \to c\kappa$. There must be at least one factor of type (i), (ii) with $s > 0$ or a fully contracted factor of type (iii) with more ξ^a's than N^a's. The total number of derivatives in each monomial must be less than the number of derivatives in the Lagrangian, with each occurrence of κ counting as one derivative.

Proof. By assumptions 1.–5. $\mathbf{B}_\mathcal{H}$ is a sum of monomials consisting of the following factors: (a) $\nabla_{a_1}\ldots\nabla_{a_j}N_b$, (b) $\nabla_{a_1}\ldots\nabla_{a_j}\xi_b$, (c) $\nabla_{a_1}\ldots\nabla_{a_j}R_{abcd}$, which are fully contracted into the legs of the tetrad N^a, ξ^a, s_i^a. We will now show by induction in $j \geq 0$ that each such term can be converted into one of the terms (i)–(iii) listed in lemma 1. For term (c) it follows from the results of [11]. Thus, we only have to deal with the terms (a) and (b). For $j = 0$, there is nothing to show.

Assuming that the statement is true up to $j-1$ derivatives, let us prove it for j derivatives. We begin by considering all possible ways of dotting the legs of our null tetrad into a term of the form (b), $\nabla_{a_1}\ldots\nabla_{a_j}\xi_b$. By commuting derivative operators at the expense of Riemann tensors (giving terms of the form (b) with fewer derivatives and terms of the form (c)), and by moving any N^a factors into the derivative operators at the expense of new terms of the form (a) with fewer than j derivatives, we may bring the expression into the form $\nabla_{a_1}\ldots\nabla_{a_{j-k}}(N^c\nabla_c)^k\xi_b$, dotted into ξ^a's and s_i^a's. For $k \geq 2$, we may use (C3) to lower k at the expense of terms of the form (c), so only the case $k = 0, 1$ needs to be considered. As an example of such a term for $k = 1$, consider $s_{i_1}^{a_1}\ldots s_{i_s}^{a_j}\xi^{a_{s+1}}\ldots\xi^{a_{j-1}}\nabla_{a_1}\ldots\nabla_{a_{j-1}}(N^c\nabla_c\xi_b)$. At the expense of terms of the form (a) when a derivative hits N^c, we can move all derivatives tangent to ξ^d onto ξ_b and then use $\xi^d\nabla_d\xi_b \widehat{=} \kappa\xi_b$, since all derivatives are at this stage tangent to $\mathcal{H}$. Thus, we can effectively assume that the term we are dealing with is of the form

$$\text{term(b)} = s_{i_1}^{a_1} \ldots s_{i_{j-1}}^{a_{j-1}} \nabla_{a_1} \ldots \nabla_{a_{j-1}} (N^c \nabla_c \xi_b)$$
$$= s_{i_1}^{a_1} \ldots s_{i_{j-1}}^{a_{j-1}} \nabla_{a_1} \ldots \nabla_{a_{j-1}} \left(N^c \left(\frac{1}{2} \mathcal{L}_\xi g_{cb} + 2\kappa N_{[c} \xi_{b]} \right) \right). \tag{C4}$$

The first term on the right side is of the form (i) in the lemma, whereas the second has fewer than j derivatives acting onto ξ_b and at most j derivatives acting onto N^c, i.e. it is a term of the form (a), after symmetrizing derivatives at the expense of terms of the form (c) with Riemann tensors. The case $k = 0$ is treated similarly.

Next, we consider all possible ways of dotting the legs of our null tetrad into a term of the form (a), $\nabla_{a_1} \ldots \nabla_{a_j} N_b$. By commuting derivative operators at the expense of Riemann tensors (giving terms of the form (a) with fewer derivatives and terms of the form (c)), and by moving any N^a factors into the derivative operators at the expense of new terms of the form (a) with fewer than j derivatives, we may bring the expression into the form $\nabla_{a_1} \ldots \nabla_{a_{j-k}} (N^c \nabla_c)^k N_b$, dotted into ξ^a's and s_i^a's. Such a term vanishes for $k > 0$ because of $N^c \nabla_c N_b = 0$. We can move all derivatives tangent to ξ^d onto N_b at the expense of terms of the form (b) with fewer than j derivatives. Thus, we can effectively assume that the term we are dealing with is of the form

$$\text{term(a)} = s_{i_1}^{a_1} \ldots s_{i_{j-s}}^{a_{j-s}} \xi^{a_{j-s+1}} \ldots \xi^{a_j} \nabla_{a_1} \ldots \nabla_{a_j} N_b. \tag{C5}$$

We can now gradually express the number of derivatives into the ξ^a-direction in terms of Lie derivatives $\mathcal{L}_\xi$ and terms such that each factor has fewer than j derivatives. So, we have reduced the (a) terms to terms of the form

$$\text{term(a)} = s_{i_1}^{a_1} \ldots s_{i_{j-s}}^{a_{j-s}} \nabla_{a_1} \ldots \nabla_{a_{j-s}} (\mathcal{L}_\xi)^s N_b, \tag{C6}$$

and we can symmtrize the a_i-indices at the expense of terms of the type (c). If we dot ξ^b into this expression, we get terms of type (b), since $\xi^b N_b = 1$. If we dot N^b into this expression, we get terms of the type (a) with fewer derivatives since $N^b N_b = 0$ in a neighborhood of $\mathcal{H}$. So we have effectively reduced attention to terms of the form claimed in item (ii). ∎

The main consequence of this lemma is as follows: If $\mathcal{H}$ is a stationary bifurcate Killing horizon with Killing vector field ξ^a, then any factor of type (i), (ii) with $s > 0$ or any fully contracted factor of type (iii) with more ξ^a's than N^a's vanishes. By contrast, a factor of type (ii) with $s = 0$, or a fully contracted factor of type (iii) with no more ξ^a's than N^a's factors will not, in general, vanish. It is thus clear that the ambiguity in $\mathbf{B}_\mathcal{H}$ consists in adding monomials with at least *two* factors of either type (i), (ii) with $s > 0$, or a fully contracted factor of type (iii) with more ξ^a's than N^a's.

3. Comparison with HKR Proposal

In [18] Hollands, Kovacs, and Reall (HKR) constructed an entropy-current $(n-2)$-form $\mathbf{S}_{\text{HKR}}$ in an "effective field theory" (EFT) framework

$$L = \frac{1}{8\pi} (2\ell^{-2}\Lambda + R + \ell^2 L_4 + \ell^4 L_6 + \ldots), \tag{C7}$$

where each L_N is a Lagrangian that is locally and covariantly constructed out of the metric and N derivatives. EFT is understood to mean that one restricts attention to only those solutions in the theory truncated at some given order N such that, roughly speaking, the higher-order terms[20] in the equations of motion are locally small near $\mathcal{H}$. $\mathbf{S}_{\text{HKR}}$ is by construction equal to the Dong-Wall entropy-current $(n-2)$-form [5,6] $\mathbf{S}_{\text{DW}}$ to first order off a solution with a stationary bifurcate Killing horizon. [18,23] showed that the ambiguities can also be exploited, for each N, so that $\mathbf{S}_{\text{HKR}}$ satisfies the 2nd law in the EFT sense, i.e., up to terms that are of the same order in ℓ as the terms neglected when truncating the EFT order N.

A comparison between the ambiguities in the HKR approach and our approach may now be given as follows: In Lemma 2.3 in Ref. [18] it was shown that the admissible terms in $\mathbf{S}_{\text{HKR}}$ must be monomials in Gaussian null coordinate (GNC) components [see (81)] of the following factors of total boost weight zero (see definition 2.2 in Ref. [18]):

(i) $D_{(a_1} \cdots D_{a_j)} K_{ab}$;
(ii) $D_{(a_1} \cdots D_{a_{j-1})} \beta_{a_j}$, $D_{(a_1} \cdots D_{a_j)} \bar{K}_{ab}$;
(iii) $\nabla_{(a_1} \cdots \nabla_{a_j} R^b{}_{a_{j+1} a_{j+2})}{}^c$, R_{abcd};

where $n^a = (\partial_\rho)^a$. To compare with lemma 1, we have

$$N^a \mathrel{\widehat{=}} \kappa^{-1} \left[\nabla^a \log V + \frac{1}{2} \beta^a - \frac{1}{8} V \beta^b \beta_b k^a \right],$$
$$\xi^a \mathrel{\widehat{=}} \kappa V k^a,$$
$$s_i^a \mathrel{\widehat{=}} e_i^a - \frac{1}{2} V e_i^b \beta_b k^a,$$
$$\frac{1}{2} \mathcal{L}_\xi g_{ab} \mathrel{\widehat{=}} \kappa V K_{ab} - \frac{1}{4} V^3 \kappa K_{cd} \beta^c \beta^d k_a k_b, \tag{C8}$$

where $\sum_i e_i^a e_i^b = \gamma^{ab}$. The scaling requirement in lemma 1 is equivalent to the zero total boost weight requirement[21] in [18]. Furthermore, comparing (i')–(iii') in the light of (C8) with the terms (i)–(iii) in Lemma 1, one can see some broad similarities between (iii) and (iii') and (i) and (i'), if we consider x^A GNC components as analogous to s_i^a tetrad components. Items (ii) and (ii') are also similar, in the

[20]The cosmological constant term is also considered higher order, in the sense that $\Lambda \ll 1$, cf. the cosmological constant problem.

[21]Here it must be understood that the boost weights of κ, V be $+1, -1$ respectively. Note that in the HKR scheme, no explicit factors of κ nor V appear.

following sense. If in (ii), we have $s = 0$ (no Lie-derivative in the ξ^a-direction), then one can see from Eqs. (C8) that we basically get a specific combination of the *two* types of terms in (ii'). When $s \geq 1$ in the term (ii), we also get certain combinations of the *two* types of terms in (ii'). To see this, we first consider one Lie-derivative in the ξ^a-direction on N_b (C8), to get

$$\mathcal{L}_\xi N_a \hat{=} \frac{1}{2} V \mathcal{L}_k \beta_a + \frac{1}{4} V^2 (\beta^c \mathcal{L}_k \beta_c + K_{bc} \beta^b \beta^c) k_a. \tag{C9}$$

This already is of the type of terms of the form (i'), (ii') except for the term involving $\mathcal{L}_k \beta_a$. However, by equations given in [[18], sec. A.1], we have

$$\mathcal{L}_k \beta_a + K_{ab} \beta^b \hat{=} 2 R_{bcda} n^b k^c k^d, \tag{C10}$$

so $\mathcal{L}_\xi N_a$ is in fact a combination of the terms in (i'), (ii'), (iii') not involving $\bar{K}_{ab}$. This remains true by an inductive argument for $(\mathcal{L}_\xi)^s N_a$, and more generally for all our terms of type (ii) in lemma 1. Thus, our terms (i), (ii), (iii) in lemma 1 correspond to specific combinations of the terms (i'), (ii'), (iii').

Thus, we see that the ambiguities present in our approach are more restrictive than in the HKR approach. Correspondingly, our $S[\mathcal{C}]$ is more rigidly fixed than $S_{\mathrm{HKR}}[\mathcal{C}]$. We do not believe that we would have sufficient re-definition freedom in our approach to ensure that our entropy flux (114) satisfies the 2nd law beyond general relativity. On the other hand, it may still be possible to ensure that the second law holds for specific EFTs, which may thus provide an interesting selection criterion to discriminate unphysical EFTs.

It appears very unlikely that the more general freedom allowed by HKR will, in general, yield an entropy $S_{\mathrm{HKR}}[\mathcal{C}]$ that is cross section continuous [21]. To appreciate why, recall that, as shown in Sec. 2.1 of [18], under a change of affine parameter $V = V'\psi(x^A)$, where $\psi > 0$, the quantities associated with a corresponding GNC system [see (81)] change as

$$\begin{aligned} \beta'_a &\hat{=} \beta_a + 2 D_a \log\psi - 2 V K_a{}^b D_b \log\psi, \\ \bar{K}'_{ab} &\hat{=} \psi^{-1} (\bar{K}_{ab} - V D_a D_b \log\psi - V (D_a \log\psi) D_b \log\psi - V \beta_{(a} D_{b)} \log\psi), \\ K'_{ab} &\hat{=} \psi K_{ab}, \end{aligned} \tag{C11}$$

where K_{ab} and $\bar{K}_{ab}$ are the extrinsic curvatures along V respectively ρ. Let $\mathcal{C}$ be a cut of constant V and $\mathcal{C}'$ a cut of constant V' and consider a ψ that is extremely close to 1 but with extremely large angular derivatives (see Fig. 1). As an example, consider now the Ricci-squared theory, with Dong-Wall entropy given by (152). It is easy to see that the first term in (152) is cross section covariant because the definition of the Ricci tensor does not involve any notion of foliation. Let us next focus on the $K\bar{K}$-term in (152). Using (C11), it is seen to change as

$$K' \bar{K}' \hat{=} K (\bar{K} - V' D_a D^a \psi - V' \beta^a D_a \psi) \tag{C12}$$

under a change of affine parameter. Therefore, since $\gamma'_{ab} \hat{=} \gamma_{ab}$, we obtain

$$\int_{\mathcal{C}'} K' \bar{K}' \boldsymbol{\epsilon}'^{(n-2)} = \int_{\mathcal{C}'} K (\bar{K} - V' D_a D^a \psi - V' \beta^a D_a \psi) \boldsymbol{\epsilon}^{(n-2)}. \tag{C13}$$

The terms in this last expression involving derivatives of ψ threaten to blow up. However, these terms crucially are *linear* in ψ, and since V' is constant on $\mathcal{C}'$, we may pull it out of the integral and then integrate the angular derivatives D_a by parts. Hence, it is clear that also the $K\bar{K}$ term in the Dong-Wall entropy given by (152) is cross section continuous. On the other hand, it seems highly unlikely that $\mathbf{S}_{\mathrm{HKR}}$ would only produce terms that are linear in ψ, since, in particular, it contains noncovariant terms starting at sufficiently high EFT order [24]. By contrast, since our entropy is obtained from an entropy $(n-2)$-form that is covariant in the metric and ξ^a, it automatically yields a cross section continuous entropy.

Of course, in an EFT framework, it is not sensible to consider surfaces, like $\mathcal{C}'$, that are wiggly on a scale that is comparable to or smaller than the cutoff scale ℓ of the EFT. More precisely, in the EFT setting, we fix a GNC system with respect to which the EFT assumptions [18] are formulated, and we would not be allowed to perform a change of GNCs with a ψ such that e.g., $\ell \bar{K}'$ as in (C11) failed to remain small. By the results of [18] and their generalization [23] on the second law for S_{HKR}, if we do chose to impose such a restriction, this would preclude the possibility that reparametrization noninvariant terms could spoil the second law. However, one can see from (C11) that this would imply a restriction on how large V can be, which is somewhat against the EFT spirit, since that is not a restriction on the UV behavior. If we merely insist that the wiggles in $\mathcal{C}'$ were on a scale considerably larger than ℓ, i.e., if we merely require

$|D_{a_1}\cdots D_{a_k}\psi| \ll O(\ell^{-k})$ but we allow V to be arbitrarily large, then a violation of the second law of S_{HKR} [18,23] appears to be possible due to reparametrization noninvariance.

APPENDIX D: APPLICATION TO DILATON GRAVITY THEORIES WITH SCALAR FIELDS

Dong and Lewkowycz [34] have analyzed a dilaton gravity theory coupled to two scalar fields and have obtained an entropy formula for this theory using Dong's approach. In this appendix we show that the linearization of their entropy expression agrees with what would be calculated from (120). This lends further support to the belief that Dong's entropy agrees in general with the entropy that would be obtained from Wall's approach.

The Lagrangian of dilaton gravity in two dimensions coupled with two scalar fields, σ and ω, considered in [34] is

$$L_{ab} = -\frac{1}{2}\epsilon_{ab}[\phi R - (\nabla\sigma)^2 - (\nabla\omega)^2 + \lambda\omega\nabla_c\nabla_d\sigma\nabla^c\nabla^d\sigma], \tag{D1}$$

where λ is a constant. The equations of motion for g_{ab} are

$$(E_G)_{ab} = -\frac{1}{2}\Big\{\nabla_a\sigma\nabla_b\sigma + \nabla_a\omega\nabla_b\omega + \nabla_a\nabla_b\phi - g_{ab}(g^{ed}\nabla_e\nabla_d\phi + \frac{1}{2}[(\nabla\sigma)^2 + (\nabla\omega)^2 - \lambda\omega(\nabla\nabla\sigma)^2]) + \lambda[\nabla^m(\omega\nabla_m\nabla_b\sigma)\nabla_a\sigma + (a\leftrightarrow b)] - \lambda\nabla^c(\omega\nabla_a\nabla_b\sigma\nabla_c\sigma)\Big\} \tag{D2}$$

The pullback of the constraints to the horizon can be put in the form

$$\xi^a C_{ab_2} = 2\epsilon^{(1)}_{b_2}[\kappa\mathcal{L}_\xi P - \mathcal{L}^2_\xi P] \tag{D3}$$

where

$$P(\psi,\xi;\delta\psi) = \frac{1}{2}\kappa(\delta\phi - \lambda\omega\mathcal{L}_\xi\delta\sigma N^a\nabla_a\sigma) \tag{D4}$$

where ψ collectively refers to the dynamical fields g,ϕ,σ,ω.

The Noether charge 0-form is

$$Q = -\frac{1}{2}\epsilon_{ae}A^{aebd}\nabla_{[b}\xi_{d]} + [\epsilon_{ae}\nabla_f A^{fbae} + \epsilon_{af}B^{a(fb)}]\xi_b \tag{D5}$$

where

$$A^{abcd} = -\frac{1}{2}\phi(g^{ac}g^{bd} - g^{ad}g^{bc}) \tag{D6}$$

$$\tilde{B}^{abc} = \frac{1}{2}\nabla_d\phi(g^{ac}g^{bd} - g^{ad}g^{bc}) + \frac{1}{2}\lambda\omega\nabla_{(m}\nabla_{n)}\sigma\nabla_t\sigma(2g^{ma}g^{nc}g^{bt} - g^{mb}g^{nc}g^{at}). \tag{D7}$$

The symplectic potential $\boldsymbol{\theta}$ is given by

$$\theta_{b_2} = \epsilon_{ab_2}(A^{abcd}\nabla_d\delta g_{bc} + \tilde{B}^{abc}\delta g_{bc} + D^a\delta\sigma + F^{af}\nabla_f\delta\sigma + M^a\delta\omega) \tag{D8}$$

where

$$D^a = \nabla^a\sigma + \lambda g^{me}g^{na}\nabla_e(\omega\nabla_{(m}\nabla_{n)}\sigma) \tag{D9}$$

$$F^{af} = -\lambda\omega g^{ma}g^{nf}\nabla_{(m}\nabla_{n)}\sigma \tag{D10}$$

$$M^a = \nabla^a\omega \tag{D11}$$

The pullback of $\boldsymbol{\theta}$ to the horizon takes the form

$$\underline{\boldsymbol{\theta}} \hat{=} \delta\mathbf{B}_{\mathcal{H}} \tag{D12}$$

where

$$B_{\mathcal{H}b} = \frac{1}{2}\epsilon^{(1)}_b(2\lambda\omega N^a\nabla_a\sigma\mathcal{L}_\xi\sigma) \tag{D13}$$

Our entropy 0-form S (see (71) is given by

$$S = \frac{2\pi}{\kappa}(Q - \xi\cdot\mathbf{B}_{\mathcal{H}}) \tag{D14}$$

$$= \frac{2\pi}{\kappa}[-\kappa\phi + \kappa\lambda\omega N^a\nabla_a\sigma\mathcal{L}_\xi\sigma - \mathcal{L}_\xi(-\phi + \lambda\omega N^a\nabla_a\sigma\mathcal{L}^2_\xi\sigma)] \tag{D15}$$

The Dong-Wall entropy (see (120) is

$$S_{\mathrm{DW}} = \frac{2\pi}{\kappa}(Q - \xi\cdot\mathbf{B}_{\mathcal{H}}) - \frac{4\pi}{\kappa}P(\psi,\xi;\mathcal{L}_\xi\psi) \tag{D16}$$

$$= 2\pi[-\phi + \lambda\omega N^a\nabla_a\sigma\mathcal{L}_\xi\sigma], \tag{D17}$$

This agrees[22] with the linearization of (A.44) of [34].

[22]To compare, one must take into account—as one also must do in comparing Eq. (14) of [6] with Eq. (1.3) of [5]—that Dong works in Euclidean signature and uses a complex z coordinate.

[1] J. M. Bardeen, B. Carter, and S. W. Hawking, The four laws of black hole mechanics, Commun. Math. Phys. **31**, 161 (1973).

[2] S. W. Hawking, Particle creation by black holes, Commun. Math. Phys. **43**, 199 (1975).

[3] J. D. Bekenstein, Black holes and entropy, Phys. Rev. D **7**, 2333 (1973).

[4] V. Iyer and R. M. Wald, Some properties of the noether charge and a proposal for dynamical black hole entropy, Phys. Rev. D **50**, 846 (1994).

[5] X. Dong, Holographic entanglement entropy for general higher derivative gravity, J. High Energy Phys. 01 (2014) 044.

[6] A. C. Wall, A second law for higher curvature gravity, Int. J. Mod. Phys. D **24**, 1544014 (2015).

[7] R. M. Wald and A. Zoupas, General definition of, "conserved quantities" in general relativity and other theories of gravity, Phys. Rev. D **61**, 084027 (2000).

[8] S. Hollands and R. M. Wald, Stability of black holes and black branes, Commun. Math. Phys. **321**, 629 (2013).

[9] S. Gao and R. M. Wald, "Physical process version" of the first law and the generalized second law for charged and rotating black holes, Phys. Rev. D **64**, 084020 (2001).

[10] R. M. Wald, *General Relativity* (Chicago University Press, Chicago, USA, 1984).

[11] U. Müller, C. Schubert, and A. van de Ven, A closed formula for the Riemann normal coordinate expansion, Gen. Relativ. Gravit. **31**, 1759 (1999).

[12] K. Prabhu, The first law of black hole mechanics for fields with internal gauge freedom, Classical Quantum Gravity **34**, 035011 (2017).

[13] V. Iyer and R. M. Wald, Comparison of the noether charge and euclidean methods for computing the entropy of stationary black holes, Phys. Rev. D **52**, 4430 (1995).

[14] M. D. Seifert and R. M. Wald, General variational principle for spherically symmetric perturbations in diffeomorphism covariant theories, Phys. Rev. D **75**, 084029 (2007).

[15] I. Râcz and R. M. Wald, Extensions of spacetimes with killing horizons, Classical Quantum Gravity **9**, 2643 (1992).

[16] I. Rácz and R. M. Wald, Global extensions of spacetimes describing asymptotic final states of black holes, Classical Quantum Gravity **13**, 539 (1996).

[17] A. A. Rahman and R. M. Wald, Black hole memory, Phys. Rev. D **101**, 124010 (2020).

[18] S. Hollands, Á. D. Kovács, and H. S. Reall, The second law of black hole mechanics in effective field theory, J. High Energy Phys. 08 (2022) 258.

[19] A. C. Wall and Z. Yan, Linearised second law for higher curvature gravity and non-minimally coupled vector fields, arXiv:2402.05411.

[20] M. R. Visser and Z. Yan, Properties of dynamical black hole entropy, arXiv:2403.07140.

[21] P.-N. Chen, D. E. Paraizo, R. M. Wald, M.-T. Wang, Y.-K. Wang, and S.-T. Yau, Cross-section continuity of definitions of angular momentum, Classical Quantum Gravity **40**, 025007 (2022).

[22] R. M. Wald, Black hole entropy is the noether charge, Phys. Rev. D **48**, R3427 (1993).

[23] I. Davies and H. S. Reall, A non-perturbative second law of black hole mechanics in effective field theory, Phys. Rev. Lett. **132**, 171402 (2024).

[24] I. Davies and H. S. Reall, Dynamical black hole entropy in effective field theory, J. High Energy Phys. 05 (2023) 006.

[25] R. M. Wald, Entropy and black-hole thermodynamics, Phys. Rev. D **20**, 1271 (1979).

[26] V. Chandrasekaran, É. É. Flanagan, and K. Prabhu, Symmetries and charges of general relativity at null boundaries, J. High Energy Phys. 11 (2018) 125.

[27] K. Shi, X. Wang, Y. Xiu, and H. Zhang, Covariant phase space with null boundaries, Commun. Theor. Phys. **73**, 125401 (2021).

[28] A. Rignon-Bret, Second law from the noether current on null hypersurfaces, Phys. Rev. D **108**, 044069 (2023).

[29] L. Ciambelli, L. Freidel, and R. G. Leigh, Null raychaudhuri: Canonical structure and the dressing time, J. High Energy Phys. 01 (2024) 166.

[30] T. Azeyanagi, G. Compère, N. Ogawa, Y. Tachikawa, and S. Terashima, Higher-derivative corrections to the asymptotic Virasoro symmetry of 4d extremal black holes, Prog. Theor. Phys. **122**, 355 (2009).

[31] R. M. Wald, On identically closed forms locally constructed from a field, J. Math. Phys. (N.Y.) **31**, 2378 (1990).

[32] P. Bueno, P. A. Cano, V. S. Min, and M. R. Visser, Aspects of general higher-order gravities, Phys. Rev. D **95**, 044010 (2017).

[33] A. Mishra, S. Chakraborty, A. Ghosh, and S. Sarkar, On the physical process first law for dynamical black holes, J. High Energy Phys. 09 (2018) 034.

[34] X. Dong and A. Lewkowycz, Entropy, extremality, euclidean variations, and the equations of motion, J. High Energy Phys. 01 (2018) 081.

[35] S. Bhattacharyya, P. Dhivakar, A. Dinda, N. Kundu, M. Patra, and S. Roy, An entropy current and the second law in higher derivative theories of gravity, J. High Energy Phys. 09 (2021) 169.

[36] S. Hollands, A. Ishibashi, and R. M. Wald, A higher dimensional stationary rotating black hole must be axisymmetric, Commun. Math. Phys. **271**, 699 (2007).

[37] W. G. Unruh and R. M. Wald, Acceleration radiation and the generalized second law of thermodynamics, Phys. Rev. D **25**, 942 (1982).

[38] A. C. Wall, Proof of the generalized second law for rapidly changing fields and arbitrary horizon slices, Phys. Rev. D **85**, 104049 (2012).

[39] J. Kudler-Flam, S. Leutheusser, A. A. Rahman, G. Satishchandran, and A. J. Speranza, A covariant regulator for entanglement entropy: Proofs of the Bekenstein bound and QNEC, arXiv:2312.07646.

[40] R. Bousso, Z. Fisher, S. Leichenauer, and A. C. Wall, Quantum focusing conjecture, Phys. Rev. D **93**, 064044 (2016).

[41] F. Ceyhan and T. Faulkner, Recovering the QNEC from the ANEC, Commun. Math. Phys. **377**, 999 (2020).

[42] H. Casini and M. Huerta, A c-theorem for the entanglement entropy, J. Phys. A **40**, 7031 (2007).

[43] L. Andersson, M. Mars, and W. Simon, Local existence of dynamical and trapping horizons, Phys. Rev. Lett. **95**, 111102 (2005).

[44] N. Engelhardt and A. C. Wall, Coarse graining holographic black holes, J. High Energy Phys. 05 (2019) 160.