11 Black hole entropy as a Noether charge

In lecture 9 we have discussed black hole spacetimes and their symmetries, while in lecture 10 we used covariant phase space formalism to study the phase space of asymptotically flat spacetimes and identify the asymptotic notions of energy, angular momentum, and electric charge with the Noether charges associated with the asymptotic symmetry generators. In this lecture, we combine some of the aspects discussed in the previous lecture for asymptotically flat spacetimes with the presence of a bifurcate Killing horizon (as it is the case for stationary black hole solutions) and take a further step forward before moving to the discussion of the laws of black holes mechanics and their thermodynamical interpretation, which will be the subject of the next lecture.

Specifically, in this lecture, we consider a general classical theory of gravity in D dimensions, described by a diffeomorphism invariant Lagrangian and, assuming only that the theory admits stationary black hole solutions with a bifurcate Killing horizon, and that the canonical mass and angular momentum of solutions are well defined at infinity, it is possible to show that the first law of black hole mechanics always holds for perturbations to nearby stationary black hole solutions. In particular, the quantity playing the role of black hole entropy turns out to be 2π times the integral over the horizon bifurcation surface of the Noether charge $(D-2)$-form associated with the horizon Killing vector field, normalized so as to have unit surface gravity.

This result was first obtained by Wald in his seminal paper [gr-qc/9307038], further analysed in his second paper [gr-qc/9403028] together with Iyer where the possibility of extending such a result to the non-stationary case was also analysed in detail. More recently, in their systematic discussion of the boundary terms in covariant phase space formalism [gr-qc/1906.08616], Harlow and Wu clarified certain subtleties due to boundary terms/ambiguities in the formalism and identified a key step in Iyer-Wald analysis to be actually related to the contribution of the boundary Lagrangian which was not explicitly introduced from the very beginning in the original derivations. The presentation of this lecture will mainly follow these references to which we refer for those details omitted here due to time restrictions. In particular, as a fully satisfactory dynamical extension of Wald's entropy formula to non-stationary black holes is still subject of active research, we will focus here on the better understood stationary case and only briefly comment on the status of the extension at the end.

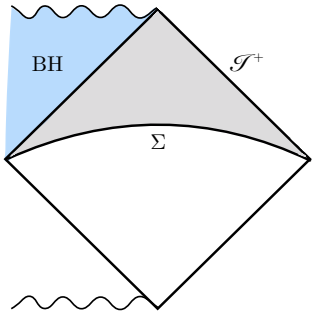
11.1 Noether charges with bifurcate Killing horizons

Let us consider generic theories of gravity defined on a D-dimensional spacetime $\mathcal{M}$ with dynamical fields consisting of a Lorentzian spacetime metric and possibly other matter fields[1] such that the equations of motion for the metric and the other fields are derived from a diffeomorphism covariant Lagrangian.

[1] For example, as discussed in lecture 10, Maxwell electromagnetic field can be included to describe electrically charged black hole solutions to Einstein-Maxwell equations.

Let us further assume that a suitable notion of asymptotic flatness is defined for our spacetimes as e.g. the coordinate-free definition of asymptotically flat spacetime given in lecture 10.

More specifically, we are interested in black hole spacetimes. From lecture 9, we recall that the black hole (BH) region of an asymptotically flat spacetime is defined to be the complement of the past of the asymptotic region. For example, in the case of the Kruskal extension of a Schwarzschild spacetime reported in the side figure, by considering a Cauchy slice Σ, the BH interior region (light blue) is the complement in the future of Σ of the past of the asymptotic region $\mathscr{I}^+$ (grey).



A stationary black hole's event horizon is a Killing horizon and we recall from lecture 9 that a Killing horizon (KH) is a null hypersurface $\mathcal{N}$ such that there exists a Killing vector field (KVF) $\xi \in \mathfrak{X}(\mathcal{M})$ such that, on $\mathcal{N}$, ξ is normal to $\mathcal{N}$ and satisfies the condition

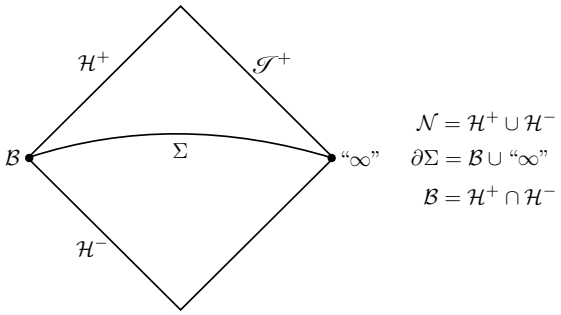
$$\xi^a \nabla_a \xi^b \big|_{\mathcal{N}} = \kappa\, \xi^b \,, \tag{11.1}$$

where κ is the surface gravity. Denoting by $\mathcal{B}$ the $(D-2)$-dimensional bifurcation surface, the KVF ξ vanishes on $\mathcal{B}$

$$\xi^a|_{\mathcal{B}} = 0 \,. \tag{11.2}$$

Bifurcate Killing horizons have constant non-vanishing surface gravity and it can be shown that any Killing horizon with constant, non-vanishing surface gravity can be locally extended (if necessary) to a bifurcate horizon. We will come back to the thermodynamic interpretation of the surface gravity in relation to Hawking temperature and the 0-th law in the next lecture.

Wald's derivation is based on applying the covariant phase space formalism to a single exterior subregion of an equilibrium wormhole solution. To this aim, let Σ be (the portion of) a Cauchy surface in the exterior region of our stationary black hole spacetime which lies between the bifurcate Killing horizon and asymptotic infinity. Looking for instance at the side figure for the Kruskal spacetime, the bifurcate Killing horizon is given by the null surface $\mathcal{N} = \mathcal{H}^+ \cup \mathcal{H}^-$ with bifurcation surface $\mathcal{B} = \mathcal{H}^+ \cap \mathcal{H}^-$. The boundary of the Cauchy slice Σ then has two components respectively at $\mathcal{B}$ and at infinity, namely $\partial\Sigma = \mathcal{B} \cup$ "∞". Here we are intentionally sloppy in our notion of infinity and colloquially denoted by "∞" the $(D-2)$-dimensional sphere at infinity without specifying whether we are referring to spatial infinity or the past of future null infinity. We will come back on this point below.



As we have discussed in lecture 8 when deriving the expression of the diffeomorphism charges, in a diffeomorphism-covariant theory the Noether charge H_ξ associated with a spacetime diffeomorphism generated by the vector field $\xi \in \mathfrak{X}(\mathcal{M})$ is a pure boundary term on-shell (i.e. on solutions of the equations of motion) given by

$$H_\xi = \int_{\partial\Sigma} Q_\xi + \iota_\xi \ell + I_{X_\xi} C \qquad , \qquad I_{X_\xi} \Omega_\Sigma = \delta H_\xi \tag{11.3}$$

where Q_ξ is a $(D-2,0)$-form such that $J_\xi = \mathrm{d}Q_\xi$, with J_ξ the Noether potential/current $(D-1,0)$-form, ℓ is the boundary Lagrangian $(D-1,0)$-form, X_ξ is the lift on field space of the vector field ξ, and C is the $(D-2,1)$-form ensuring the action to be stationary up to past/future boundary terms. In the original papers by Wald or by Iyer and Wald, C was not present ($C=0$) as in their analysis now C term was included in the definition of the (pre)symplectic potential as $\theta - \mathrm{d}C$. In the following, we will also set $C=0$ to allow the comparison with the literature and comment on its possible role later at the end of the lecture.

Specialising Eq. (11.3) for the above setup of a stationary black hole solution with a bifurcate Killing horizon, $\partial\Sigma = \mathcal{B} \cup$ "∞" so that δH_ξ can be split into two contributions

$$\delta H_\xi = \delta H_\xi^{\text{ext}} - \delta \int_{\mathcal{B}} Q_\xi \ , \tag{11.4}$$

where the "ext" superscript denotes the asymptotic contribution, while the second term in the only non-zero contribution on $\mathcal{B}$ since the latter is a bifurcation surface and the KVF ξ vanishes at $\mathcal{B}$ and also X_ξ, which for covariant Lagrangians is given by $X_\xi = \int_{\mathcal{M}} \mathrm{d}^D x (\mathcal{L}_\xi \phi) \cdot \frac{\delta}{\delta\phi}$ (cfr. Eq. (8), Lecture 8), vanishes at any point in pre-phase space where it generates a symmetry (i.e. where $\mathcal{L}_\xi \phi = 0$ for all dynamical fields ϕ). The relative sign on the r.h.s of (11.4) is due to the orientation of $\mathcal{B}$ with normal pointing towards the interior of Σ. As for the l.h.s. of Eq. (11.3), we have

$$I_{X_\xi} \Omega_\Sigma = \int_\Sigma I_{X_\xi} \omega = 0 \ , \tag{11.5}$$

as a direct consequence of the vanishing of X_ξ when ξ is a symmetry of all dynamical fields as e.g. it is the case for the metric as the KVF ξ generates the isometries of the metric ($\mathcal{L}_\xi g = 0$). For a stationary BH solution with bifurcate Killing horizon, the Killing vector field ξ vanishing on $\mathcal{B}$ can be chosen to be normalised so that

$$\xi^a = t^a + \Omega_{\mathcal{N}}^{(\mu)} \varphi^a_{(\mu)} \ , \tag{11.6}$$

where t^a is the stationary Killing field with unit norm at infinity, $\varphi^a_{(\mu)}$ denotes the axial Killing fields acting in orthogonal planes[2], and $\Omega_{\mathcal{N}}$ is the angular velocity of the horizon (we used the subscript $\mathcal{N}$ to avoid confusion with the symplectic form Ω). Now, as discussed in lecture 10 for the case of null infinity, asymptotically flat spacetimes admit well-defined notions of energy E and angular momentum $\mathcal{J}$ which are identified with the Noether charges H_ξ^{ext} associated with the generators of asymptotic time translations $\xi^a = t^a$ and asymptotic rotations $\xi^a = \Omega_{\mathcal{N}}^{(\mu)} \varphi^a_{(\mu)}$, respectively[3].

A few comments are in order at this point.

- **spatial or null infinity:** the discussion of the previous lecture mainly focused on null infinity and it was only briefly mentioned that a similar discussion can be carried out for spatial infinity. The notions of energy and angular momentum charges on null infinity are related to the so-called Bondi energy-momentum. In their analysis, Iyer and Wald instead

[2] The sum over μ is needed in dimensions higher then 4 ($D \geq 5$). In $D=4$ dimensions, we can think of φ as playing the role of the vector field $Y = \partial_\varphi$ generating rotations around φ in the case of the Kerr-Newman solution discussed at the end of the last lecture.

[3] Here, unlike lecture 10, we use calligraphic font $\mathcal{J}$ for the angular momentum and not J to avoid confusion with the generic notation for the Noether current $J_\xi = \mathrm{d}Q_\xi$ mentioned above.

considered the spatial slice Σ to have one component of its boundary extending at spatial infinity so that the corresponding charges are the so-called ADM energy-momentum charges. Here there is no particular conflict in what we may think of being a discrepancy and in fact, in lecture 10, the definition of the charges at null infinity of a black hole spacetime was obtained by taking as starting point the expression of the charges H_ξ derived in lecture 8, where the very same quantity was shown to agree with the known results of for the charges obtained via the canonical formalism (for the explicit case of GR). In general, indeed, as long as no Bondi *news* is involved (i.e. no gravitational radiation), it has been discussed in the literature (see e.g. [gr-qc/0511036]) that the ADM energy-momentum can be regarded as the past limit of Bondy energy-momentum. In presence of non-zero *news*, however, the ADM total energy is no longer the past limit of the Bondi mass. This is not a problem for our present discussion as for the case of our interest no *news* is involved (cfr. discussion of the Kerr-Newman metric in Bondi gauge in Sec. 10.6).

- **the role of boundary terms:** coming back to the expression of the charges $H^{\rm ext}_\xi$ with ξ given in Eq. (11.6), let us comment on the inclusion of the boundary terms from the very beginning of the analysis of covariant phase space with boundaries as done in these lecture notes following Harlow and Wu compared to the original derivation of Iyer and Wald. In Iyer-Wald, the boundary contribution to H_ξ coming from the boundary Lagrangian ℓ was not included in the analysis and one of the key (but also kind of "obscure") points in their work was that the relation

$$\delta H_\xi = \delta \int_\Sigma J_\xi - {\rm d}(\iota_\xi \theta) = \delta \int_{\partial\Sigma} Q_\xi - \int_\infty \iota_\xi \theta \,, \tag{11.7}$$

valid for ϕ solution of the equations of motion, ξ such that $\mathcal{L}_\xi \phi = 0$, and $\delta\phi$ a solution of the linearised equations but not necessarily $\mathcal{L}_\xi \delta\phi = 0$ (the only contribution to the integral over $\partial\Sigma = \mathcal{B} \cup \infty$ in second term is only the asymptotic one as $\xi = 0$ on $\mathcal{B}$), led them – actually Wald in his first paper – to argue that a Hamiltonian associated with the vector field ξ does exist if (and only if) one can find a $(D-1)$-form B such that

$$\delta \int_\infty \iota_\xi B := \int_\infty \iota_\xi \theta \,, \tag{11.8}$$

so that

$$H^{\rm ext}_\xi = \int_\infty Q_\xi - \iota_\xi B \,. \tag{11.9}$$

The expression for B was then found case by case by computing it via Eq. (11.8) and showing that the resulting expression (11.9) for the Hamiltonian $H^{\rm ext}_\xi$ reproduced the correct result for the energy of asymptotically flat solutions of vacuum GR (see discussion p. 15-16, Eqs. (83)-(90) of Iyer-Wald paper). However, the $(D-1)$-form B was not derived systematically from the analysis. It was only recently that in their systematic investigation of boundary terms in the covariant phase space formalism, Harlow and Wu found out that the B term of Iyer-Wald analysis is nothing but the contribution coming from the inclusion of the boundary Lagrangian in the very beginning of the analysis as we also did in our discussion of the Noether charges in Secs. 7 and 8. Eq. (11.9) is in fact nothing but the expression (11.3) for H_ξ at infinity with $\ell \to -B$ (and $C = 0$). Moreover, as it was already explained in Iyer-Wald (cfr. discussion below Eq. (90) on p. 16), such a boundary term in the expression of the energy accounts for the discrepancy (by a factor

2) in the Komar mass formula, while the angular momentum coincides exactly with the Komar angular momentum.

With this being said, let us continue with our discussion of Eqs. (11.3)-(11.6). Using Eqs. (11.4) and (11.5), the second equation in (11.3) yields

$$\delta H_\xi^{\rm ext} = \delta \int_{\mathcal{B}} Q_\xi \ , \tag{11.10}$$

and, for ξ given in Eq. (11.6), we have

$$\delta H_\xi^{\rm ext} = \delta E - \Omega_{\mathcal{N}} \delta \mathcal{J} \ , \tag{11.11}$$

$$E = \int_\infty Q_t + \iota_t \ell = \int_\infty Q_t - \iota_t B \ , \tag{11.12}$$

$$\mathcal{J} = -\int_\infty Q_\varphi \ , \tag{11.13}$$

where t^a is the vector field generating asymptotic time translations and φ^a the vector field generator of asymptotic rotations. The latter is everywhere tangent to the $(D-2)$-sphere at infinity so that the pull-back of $\iota_\varphi \theta$ to that surface vanishes and no $\iota_\xi \theta$ contribution survives in $\mathcal{J}$. The sign difference in the Q_t and Q_φ part is due to the Lorentz signature of the spacetime metric[4]. All together, we have the following result

$$\boxed{\delta E - \Omega_{\mathcal{N}} \delta \mathcal{J} = \delta \int_{\mathcal{B}} Q_\xi} \tag{11.14}$$

This is the "first law of black hole mechanics". The same result can be derived also via the canonical instantaneous Hamiltonian formalism. Wald's derivation based on covariant phase space formalism has the advantage that the surface term at the BH bifurcation surface on the r.h.s. of (11.14) comes now to be explicitly identified with the variation of the Noether charge Q_ξ on $\mathcal{B}$. Eq. (11.14) however is not yet in the desired form to be identified/related with the first law as, first of all, Q_ξ on the r.h.s. of (11.14) is locally constructed from ξ^a, its derivatives, as well as from the dynamical fields of the theory and has not been written yet in terms of local geometric quantities on $\mathcal{B}$ which depend only on the dynamical fields, and second we need to relate it to the usual black hole entropy term multiplied by the surface gravity. This turns out to be the case when we further restrict $\delta\phi$ to be a perturbation to a nearby stationary black hole as we will discuss in the next subsection.

11.2 (Towards) the first law of black hole mechanics

The δQ_ξ term on the r.h.s. of Eq. (11.14) can be recast in terms of a local geometric quantity on the bifurcation surface $\mathcal{B}$ multiplied by the surface gravity by using the properties of Killing vector fields as follows. First, for any Killing vector field ξ^a, any derivative $\nabla_{a_1} \dots \nabla_{a_n} \xi^b$ can be expressed in terms of linear combinations of ξ^a and its first derivative $\nabla_a \xi_b$ with coefficients depending on the Riemann curvature tensor and its derivatives[5]. Second, on $\mathcal{B}$, we have $\xi^a = 0$ and $\nabla_a \xi_b = \kappa\, \epsilon_{ab}$ (cfr. Eq. (11.1)), where ϵ_{ab} denotes the binormal to the bifurcation surface $\mathcal{B}$.

[4] A similar relative sign difference occurs for instance in the definitions of energy $E = -p_a t^a$ and angular momentum $\mathcal{J} = p_a \varphi^a$ of a particle in special relativity.

[5] Indeed, recalling the definition of Riemann tensor $\nabla_a \nabla_b \xi_c - \nabla_b \nabla_a \xi_c = R_{abc}{}^d \xi_d$ and using the Killing equation $\nabla_a \xi_b + \nabla_b \xi_a = 0$, we have $\nabla_a \nabla_b \xi_c + \nabla_b \nabla_c \xi_a = R_{abc}{}^d \xi_d$. Then, adding the same equation with

Using these properties, we can get rid of the dependence on higher derivatives of ξ in the $(D-2)$-form Q_ξ by expressing them in terms of ξ^a and $\nabla_a\xi_b$. Moreover, as the r.h.s. of (11.14) only involves an integral over $\mathcal{B}$, we can set $\xi^a = 0$ and re-express $\nabla_a\xi_b$ terms as $\kappa\,\epsilon_{ab}$ thus eliminating any reference to ξ^a in the resulting expression. Denoting by $\tilde{Q}$ the $(D-2)$-form so obtained, we have that $\tilde{Q}$ is locally constructed out of only the dynamical fields ϕ entering the Lagrangian of the theory. The explicit expression is rather involved and generically contains the derivatives of the Lagrangian w.r.t. the Riemann tensor and its covariant derivatives. We do not report it here as it is not necessary for our discussion and we refer to the paper by Iyer and Wald for details (cfr. Lemma 3.1, Proposition 4.1, and Theorem 6.1). The key observation is that, since we are considering Lagrangian theories which are covariant under diffeomorphisms acting on the dynamical fields, $\tilde{Q}$ is invariant under spacetime diffeomorphisms mapping $\mathcal{B}$ into itself. Therefore, the r.h.s. of Eq. (11.14) can be now rewritten in terms of a local geometric quantity $\tilde{Q}$ on $\mathcal{B}$. Such a $(D-2)$-form $\tilde{Q}$ is nothing but the Noether charge $(D-2)$-form associated with the rescaled horizon Killing vector field $\tilde{\xi}^a = \kappa\,\xi^a$ normalised to have unit surface gravity.

So far, Eq. (11.14) and the considerations above did not require us to restrict ourselves to any specific on-shell field variations $\delta\phi$. However, let us now consider $\delta\phi$ to be a perturbation to a nearby stationary black hole spacetime and let us identify the unperturbed and perturbed spacetimes in such a way that the Killing horizons of the two spacetimes coincide, and the unit surface gravity horizon KVFs $\tilde{\xi}$ coincide in a neighborhood of the horizons. Note that the near-horizon identification of the KVFs of the perturbed and unperturbed stationary black hole spacetimes refers to the unit surface gravity vector fields $\tilde{\xi}$ and not ξ. For a perturbation changing the surface gravity, in fact, it is not possible to identify the two spacetimes in such a way that the KVFs ξ given in Eq. (11.6) coincide on the horizon. Moreover, for a perturbation which changes $\Omega_{\mathcal{N}}$, since $\delta t^a = 0 = \delta\varphi^a$ near infinity, the requirement $\delta\xi^a = 0$ prevents ξ to be proportional to the horizon KVF near infinity in the perturbed spacetime. For such a nearby stationary black hole perturbations with the above-mentioned identification of the unit surface gravity horizon KVFs, the variation of the Noether charge δQ can be written as $\delta Q = \kappa\,\delta\tilde{Q}$ with κ the surface gravity of the unperturbed black hole. Eq. (11.14) thus yields

$$\boxed{\delta E - \Omega_{\mathcal{N}}\delta\mathcal{J} = \kappa\,\delta\int_{\mathcal{B}}\tilde{Q}}\,, \tag{11.15}$$

which can be interpreted as an expression of the first law of thermodynamics $\delta E = T\delta S$ for perturbation to a nearby stationary black hole

$$\boxed{\delta E - \Omega_{\mathcal{N}}\delta\mathcal{J} = \frac{\kappa}{2\pi}\,\delta S}\,, \tag{11.16}$$

with the role of the black hole entropy played by the quantity

$$\boxed{S = 2\pi\int_{\mathcal{B}}\tilde{Q}}\,. \tag{11.17}$$

permuted indices *bca* to the latter equation and subtracting the *cab* indices equation, it follows that $2\nabla_b\nabla_c\xi_a = (R_{abc}{}^d + R_{bca}{}^d - R_{cab}{}^d)\xi_d = -2R_{cab}{}^d\xi_d$, where in the second equality the property $R_{[abc]}{}^d = 0$ of the Riemann tensor has been used. Thus, one obtains that the relation $\nabla_a\nabla_b\xi_c = -2R_{cab}{}^d\xi_d$ holds for any Killing vector field ξ. Similar discussion holds for higher derivatives.

The result (11.17) is know as the *Wald's entropy formula* and tells us that, in a general theory of gravity, the entropy of a stationary black hole is a local geometrical quantity constructed out of the dynamical fields and is just given by 2π times the Noether charge of the horizon Killing vector field normalised to have unit surface gravity. It should be however kept in mind that in order for Eq. (11.16) to be really identified with the first law of black hole mechanics we still need to relate the $\kappa/2\pi$ factor in front of Wald's entropy with the temperature of the black hole. Such a thermodynamic interpretation for the surface gravity in terms of the so-called Hawking temperature will be discussed in the next lecture and the statement of the surface gravity being constant over the event horizon of stationary black holes will then amount to the zeroth law of black hole thermodynamics.

Finally, as proposed by Wald in his 1993 paper and then further investigated in his 1994 paper with Iyer, the local geometric character of the entropy formula (11.17) suggests a possible generalisation to the non-stationary case. The idea is that, in the non-stationary case, the same procedure described above for the bifurcation surface of a stationary black hole Killing horizon can be in principle used also to construct the quantity $\tilde{Q}$ for an arbitrary cross-section of the horizon of a non-stationary black hole. The resulting integral over the cross-section would then provide us with a candidate expression for a dynamical black hole entropy. However, Iyer and Wald argued that the covariant phase space ambiguity of adding an extact form to the presymplectic potential θ leads to an ambiguity in the definition of the Noether charge Q_ξ which vanishes only for the case of stationary solutions. Harlow and Wu, on the other hand, suggested that the inclusion of a $C \neq 0$ boundary contribution into the Noether charge H_ξ (11.3) could be maybe used to fix such an issue as the only ambiguity left in the formalism would be a simultaneous shift of θ and C which does not affect H_ξ. This is an interesting possibility which will allow to unambiguously define a generalisation of Wald's entropy for dynamical horizons within the covariant phase space formalism. This is however not straigthforward as it might seem as the identification of the proper C term requires a careful analysis of boundary conditions at the horizon and is strictly related to the construction of a correct phase space description for the exterior region. No definite answer has been given yet. Moreover, the first law is not expected to hold for perturbations of non-stationary configurations, for which only the second law is expected to apply. In this respect, as discussed by Wald, the black hole entropy-Noether charge relation would imply that, for a dynamical process from an initially stationary black hole configuration to a stationary final state, the net change of entropy is given by the flux through the horizon of the Noether current associated with time translations on the horizon. This in turn hints at a possible connection between the validity of the second law of BH mechanics and the positive energy properties of the theory under consideration. A covariant phase space derivation of the second law is also subject of open research.

Further reading

[1] R. M. Wald, *Black Hole Entropy is Noether Charge*, (1993), Phys. Rev. D **48**: 3427-3431, 1993. `gr-qc/9307038`

[2] V. Iyer and R. M. Wald, *Some Properties of Noether Charge and a Proposal for Dynamical Black Hole Entropy*, (1994), Phys. Rev. D **50** (1994) 846-864, `gr-qc/9403028`

[3] D. Harlow and J. Wu, *Covariant phase space with boundaries*, JHEP **10** (2020) 146, `hep-th/1906.08616`