Weak cosmic censorship and the second law of black hole thermodynamics in higher derivative gravity

Feng-Li Lin,[1,†] Bo Ning,[2,3,*] and Yanbei Chen[4,‡]

[1]*Department of Physics, National Taiwan Normal University, Taipei 116, Taiwan*
[2]*College of Physics, Sichuan University, Chengdu, Sichuan 610064, China*
[3]*Peng Huanwu Center for Fundamental Theory, Hefei, Anhui 230026, China*
[4]*Burke Institute of Theoretical Physics and Theoretical Astrophysics 350-17, California Institute of Technology, Pasadena, California 91125, USA*

(Received 19 January 2023; accepted 19 July 2023; published 15 August 2023)

Infalling matter may destroy a black hole and expose the naked singularity. Thus, Penrose proposed the weak cosmic censorship conjecture to avoid such a possibility. On the other hand, if the black hole is not destroyed by infalling matter, from the second law of black hole thermodynamics, the black hole entropy should increase due to the information carried by the infalling matter. In this work, we demonstrate by examples of perturbative near-extremal black holes in higher derivative gravity theories that the second law implies weak cosmic censorship. We also compare our proposal to the one developed by Sorce and Wald based on the first law of black hole thermodynamics and show that the latter fails to yield weak cosmic censorship in such cases. Finally, we give proof of our proposal for generic gravity theories.

DOI: 10.1103/PhysRevD.108.044025

I. INTRODUCTION

Black holes are the simplest objects predicted by general relativity—with intriguing features. Even though black holes have curvature singularities, around which tidal gravity diverges and physical laws break down, in analytic black hole solutions, these singularities are always shielded by the event horizon [1]. Penrose further proposed [2] the weak cosmic censorship conjecture (WCCC): the curvature singularity will always be hidden behind the horizon for generic black holes, i.e., no naked singularities. Another intriguing feature is that the first and second laws of thermodynamics govern black holes. Bekenstein's conjecture that a black hole's entropy must be proportional to its area [3,4] was substantiated by the theoretical discovery of Hawking radiation, and the fact that this thermal radiation has a temperature proportional to the black hole's surface gravity [5].

Naively, one shall expect the connection between WCCC and the second law. As the second law requires, the entropy of a black hole can never decrease. This prevents the appearance of a naked singularity. The proof for the second law for Einstein gravity given in [6–8] can imply WCCC, later more direct connection is discussed in [9]. However, a demonstration for modified gravities is nontrivial, since, in this case, the entropy follows Wald's entropy formula [10,11] but not the area law. Thus, the second law may not ensure the area increase, and the above connection is unclear. To ensure WCCC is a universal physical principle, in this paper, we demonstrate such a connection explicitly for modified gravities by showing that the WCCC follows as long as the second law holds.

Wald started the demonstration of WCCC by gedanken experiments that attempt to destroy the horizon by overcharging or overspinning a black hole with infalling matter [9,12]. For simplicity, we shall focus on overcharging nonspinning black holes. Assume a family of electrovacuum solutions to the gravitational and electromagnetic field equations, parametrized by mass m and charge q to describe the configurations before and after the matter "falls in". We denote the condition for the spacetime to be a black hole, i.e., with a horizon that covers the singularity, by

$$W(m,q) \geq 0. \tag{1}$$

The exact form of $W(m,q)$ depends on the underlying theory. For example, the (outer) horizon of a Reissner-Nordström black hole is $r_+ = m + \sqrt{m^2 - q^2}$, thus $W(m,q) = m^2 - q^2$ so that (1) guarantees a positive and real r_+, thus the existence of a horizon.

The demonstration of WCCC is to show $W(m + \Delta m, q + \Delta q) \geq 0$ given the initial mass m and charge q,

*Corresponding author: ningbo@scu.edu.cn
†fengli.lin@gmail.com
‡yanbei@caltech.edu

for all the respective physically allowed changes Δm and Δq due to the infalling matter. Values of Δm and Δq depend on how the matter falls in and the underlying gravity theory. Intuitively, the physical constraints on Δm and Δq should come from the laws of black hole dynamics. Indeed, in [13], we have demonstrated that the first law of black hole dynamics is a universal condition to guarantee WCCC for extremal black holes in generic gravity theories. For near-extremal black holes, Sorce and Wald [9] generalized the first-law constraint to second-order variations and showed that this could guarantee WCCC in Einstein-Maxwell theory.[1]

One will face some challenges when trying to generalize the approach of [9] to modified gravities. The main challenge is to unambiguously define the canonical energy of gravitational waves for modified gravity and the respective energy condition required to ensure WCCC. Without such energy conditions, one can only consider the spherical collapsing with no induced gravitational wave. Moreover, in [9], the canonical energy is evaluated by relating it to the black hole entropy by the first law. Still, such a substitution is unclear due to the ambiguity of canonical energy in modified gravity.

Indeed, we show that the approach of [9] fails to demonstrate WCCC for the modified gravities. To bypass the aforementioned challenges and remedy the resultant failure, we propose demonstrating the WCCC with the second law. Our proposal does not need canonical energy or conditions for matter and gravity. All we need is Wald's formula for black hole entropy. In Einstein's gravity, the first law and energy condition can guarantee the second law, but it is unclear for modified gravities. The result obtained here may also shed some light on this issue. Finally, before we proceed, we shall emphasize the demonstration is not a tautology. Although the existence of entropy is the premise of the second law, itself does not guarantee the WCCC condition (1), since a decreasing entropy would indicate naked singularity in general relativity according to [9]. Thus, our demonstration is a consistency check in the same spirit of [9].

II. WCCC CONDITION IN HIGHER DERIVATIVE GRAVITY THEORIES

We consider the general quartic-order corrections to Einstein-Maxwell theory, which is given by the following Lagrangian:

$$\begin{aligned} L = {} & \frac{1}{2\kappa}R - \frac{1}{4}F_{\mu\nu}F^{\mu\nu} + c_1R^2 + c_2R_{\mu\nu}R^{\mu\nu} + c_3R_{\mu\nu\rho\sigma}R^{\mu\nu\rho\sigma} \\ & + c_4\kappa RF_{\mu\nu}F^{\mu\nu} + c_5\kappa R_{\mu\nu}F^{\mu\rho}F^{\nu}{}_{\rho} + c_6\kappa R_{\mu\nu\rho\sigma}F^{\mu\nu}F^{\rho\sigma} \\ & + c_7\kappa^2F_{\mu\nu}F^{\mu\nu}F_{\rho\sigma}F^{\rho\sigma} + c_8\kappa^2F_{\mu\nu}F^{\nu\rho}F_{\rho\sigma}F^{\sigma\mu}, \end{aligned} \tag{2}$$

where $\kappa = 8\pi G_N$, which will be set to 2 below, and c_i's are dimensionless constants. From the point of view of effective field theory, the above higher derivative theories can arise naturally from quantum corrections. Thus, some of these theories can be the genuine description of low-energy black hole dynamics but remains experimentally elusive due to smallness of c_i's. If WCCC is a fundamental principle for protecting the predictive power of theory, it should also apply to generic effective field theories of gravity.

To study WCCC, we first generalize the perturbative method of [15] to solve the charged black hole solutions up to $\mathcal{O}(c_ic_j)$ with $i,j = 1,\cdots 8$.[2] Based on these solutions, we can find the following $W(m,q)$ for (1),

$$W(m,q) = m^2 - q^2\left(1 - \frac{4c_0}{5q^2} + \frac{128c_4^2}{21q^4} + \cdots\right)^2 \tag{3}$$

with $c_0 \equiv c_2 + 4c_3 + c_5 + c_6 + 4c_7 + 2c_8$, and $\cdots$ denotes the other $\mathcal{O}(c_ic_j)$ terms. For simplicity, below we will only show the result for the case with nonzero c_4 as a demonstration. The other cases with nonzero $c_{i\neq 4}$ can be found in Appendix A. Besides, the black hole entropy can be obtained by Wald's formula [10,16], and it yields

$$\begin{aligned} S(m,q) = -2\pi A_h\Big[& -\frac{1}{2} - 4c_1R - 4c_2R^{rv} + 8c_3R^{rvrv} \\ & + 4(2c_4 + c_5 + 2c_6)F^{rv}F^{rv}\Big], \end{aligned} \tag{4}$$

where the area of the horizon A_h, the curvatures, and field strengths are evaluated by the on shell solution.

III. CHECK WCCC IN WALD'S GEDANKEN EXPERIMENT BY SECOND-LAW CONSTRAINTS

To follow Wald's gedanken experiment, we consider charged matter falling through the black hole's horizon within a finite time interval. Then, the black hole and the infalling matter are settled to a final stationary state belonging to the same family of solutions, either a new black hole or a naked singularity. The scheme is shown in Fig. 1.

As argued by Sorce and Wald [9], for a near-extremal black hole WCCC might be violated from first-order considerations [17], but in fact is preserved at second order. Therefore, we need to consider the variations of m and q caused by the infalling matter up to the second order. Here we outline the steps of checking WCCC upon the second law of black hole (thermo)dynamics, which basically require that the entropy difference between B and B^* due to infalling matter through $\mathcal{H}$ (see Fig. 1) is nondecreasing.

[1]Christodoulou proved that naked singularity can occur in Einstein-scalar system though is unstable, hence the cosmic censorship is still preserved [14].

[2]The detailed solutions can be found in Appendix A.

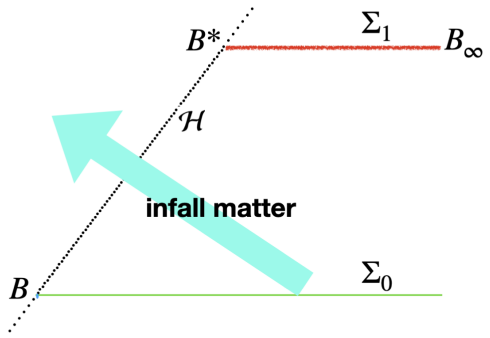

FIG. 1. Wald's gedanken experiment by throwing the charged matter into a black hole. The infalling matter crosses the horizon $\mathcal{H}$ within a finite time interval.

Let us consider an initial black hole with (m, q), with a one-parameter family of infalling matter, finally settling down to a new solution with

$$m(\lambda) = m + \lambda\delta m + \frac{\lambda^2\delta^2 m}{2}, \qquad q(\lambda) = q + \lambda\delta q + \frac{\lambda^2\delta^2 q}{2}. \tag{5}$$

Here we keep mass and charge increases up to second order in λ. We shall also restrict ourselves to nearly extremal black holes, and for the moment only consider c_4. The initial black hole is second order away from being extremal,

$$q = \sqrt{1-\epsilon^2}\left(m - \frac{128c_4^2}{21m^3}\right). \tag{6}$$

Similar to Sorce and Wald, we shall assume ϵ and λ to be of the same order of smallness, and check whether $W(m, q) > 0$ is satisfied up to second order. More specifically, we need to check whether constraints on $(\delta m, \delta q, \delta^2 m, \delta^2 q)$, arising from

$$S(m(\lambda), q(\lambda)) \geq S(m, q) \tag{7}$$

will guarantee $W(m + \Delta m, q + \Delta q) \geq 0$.

Since we will consider up to second-order variations, we assume the first-order variation due to the infalling matter to be optimally done with the second law being satisfied marginally,

$$\delta S = \frac{\partial S}{\partial m}\delta m + \frac{\partial S}{\partial q}\delta q = 0. \tag{8}$$

Solving this condition gives a relation between δm and δq. For the c_4 case that we show explicitly, it yields

$$\delta m = \left[1 - \epsilon - \frac{64(2 + 1098\epsilon)c_4^2}{7m^4}\right]\delta q + \mathcal{O}(\epsilon^2). \tag{9}$$

For extremal black holes, we have $\epsilon = 0$, up to $\mathcal{O}(c_i)$, we can truncate the terms of $\mathcal{O}(c_i^2)$ in the above to show that the first law $\delta S \geq 0$ gives

$$\delta m \geq \left(1 + \frac{4c_0}{5q^2}\right)\delta q. \tag{10}$$

This is just the WCCC condition for the extremal black holes, as demonstrated in [13] via Sorce-Wald.

Let us now consider second-order variations due to the infalling matter such that the second law holds, i.e.,

$$\delta^2 S = \frac{\partial^2 S}{\partial m^2}(\delta m)^2 + 2\frac{\partial^2 S}{\partial m \partial q}\delta m\delta q + \frac{\partial^2 S}{\partial q^2}(\delta q)^2 + \frac{\partial S}{\partial m}\delta^2 m + \frac{\partial S}{\partial q}\delta^2 q \geq 0. \tag{11}$$

For the c_4 case, combining this equation with (9), we obtain

$$\delta^2 m \geq \left[\frac{1-\epsilon}{m} + \frac{256(1655 - 17372\epsilon + 33099\epsilon^2)c_4^2}{21m^5}\right](\delta q)^2 + \left[1 - \epsilon + \frac{\epsilon^2}{2} - \frac{64(2 + 1098\epsilon - 8815\epsilon^2)c_4^2}{7m^4}\right]\delta^2 q. \tag{12}$$

This leads to

$$W(\lambda) = \left(\epsilon\left(\frac{256c_4^2}{21m^3} - m\right) + \lambda\left(1 + \frac{211072c_4^2}{21m^4}\right)\delta q\right)^2 + \mathcal{O}(c_4^3, \epsilon^3, \lambda^3), \tag{13}$$

where the $\mathcal{O}(c_4^3, \epsilon^3, \lambda^3)$ denotes the higher-order terms which will be omitted later for simplicity. Thus, we can conclude that WCCC is preserved by the second-law constraints up to $\mathcal{O}(c_4^2)$. If we consider $W(\lambda)$ only up to $\mathcal{O}(c_i)$, it takes a simple but not positive definite form

$$W(\lambda) = (\epsilon m - \lambda\delta q)^2 + \frac{8}{5m^2}(\epsilon m - \lambda\delta q) \times (c_0(\epsilon m + 3\lambda\delta q) + 10c_6\lambda\delta q) + \mathcal{O}(c_i c_j). \tag{14}$$

Completing the square of (14) requires $\mathcal{O}(c_i c_j)$ terms. This is why we need to use the near-extremal black hole solutions up to $\mathcal{O}(c_i c_j)$ to check WCCC. This is in the same spirit as invoking second-order variations in [9] to remedy the earlier mistake of [17] in checking WCCC. The check of WCCC for the other cases with $c_{i\neq 4}$ and the case of c_2 and c_4 can be found in Appendix A. All results are consistent with our proposal that the second-law constraints imply WCCC.[3]

[3]Especially, the Einstein-Maxwell-Gauss-Bonnet theory with $c_1 = c_3 = -\frac{1}{4}c_2$ gives no contribution to the $\mathcal{O}(c_i)$ term of (14), thus preserving the WCCC. We demonstrate this by the spherical thin-shell collapse in Appendix C.

Finally, due to the complication of solving the rotating charged black holes with c_i corrections, we do not check the WCCC for such cases. However, it is straightforward to check WCCC by our second-law formalism for Kerr-Newman black holes with its spin denoted by j in Einstein-Maxwell theory and the result is

$$W(\lambda)=\left(\frac{(j^2-m^4)q\delta q-2jm^2\delta j}{m(m^4+j^2)}\lambda+m\epsilon\right)^2+\mathcal{O}(\epsilon^3,\lambda^3). \tag{15}$$

The form of (15) is exactly the same as the one in [9]. This shows that our second-law proposal yields the same WCCC result as ensured by the first-law one of Sorce-Wald.

IV. COMPARISON WITH SORCE-WALD FORMALISM

For comparison, we will show that Sorce-Wald formalism fails to yield WCCC for the modified gravities considered above. Sorce-Wald formalism [9] uses the first-law constraints to check WCCC. At the second-order variations, one needs to take into account the energy contribution from the induced gravitational and electromagnetic waves, which makes the problem technically involved. For simplicity, we assume the infalling matter is spherical symmetric so that no such waves will be induced.

Our steps outlined earlier to verify WCCC are inspired by the Sorce-Wald formalism. The only difference is that we shall replace the second-law constraints by the first-law ones. The latter take the following general form [9,18]

$$\begin{aligned}&\delta^n m_{\mathrm{ADM}}-\Phi_{\mathrm{H}}(\delta^n q_{\mathrm{H}}+\delta^n q_B)-T_H\delta^n S_B\\&\quad=\delta_{n,2}\mathcal{E}_\Sigma(\phi;\delta\phi,\mathcal{L}_\xi\phi)-\int_{\mathcal{H}}\xi^a\epsilon_{ebcd}\delta^n T_a{}^e\\&\quad\geq\delta_{n,2}\mathcal{E}_\Sigma(\phi;\delta\phi,\mathcal{L}_\xi\phi).\end{aligned} \tag{16}$$

Here $n=1,2$ is the order of variation, ξ^μ is the timelike Killing vector of the background metric, and $\Phi_H=-\xi^\mu A_\mu|_{r=r_+}$ is the chemical potential on the horizon. We have used the energy condition on the stress tensor $\delta^n T_{ab}$ of the infalling matter to arrive the last inequality. Sorce-Wald assumed no matter around the bifurcation sphere B of Fig. 1 so that the variations of charge and Wald's entropy vanish, i.e., $\delta^n q_B=\delta^n S_B=0$. On the other hand, when considering the standard first law without source perturbation, we will instead set $\delta^n q_{\mathrm{H}}=\delta^n T_{ab}=0$.[4] The higher derivative corrections to Einstein-Maxwell theory cannot affect the Arnowitt-Deser-Misner (ADM) mass m_{ADM} and the charge q_{H} of the black hole due to their higher powers of $1/r$ suppression. This implies $\delta^n m_{\mathrm{ADM}}=\delta^n m$ and $\delta^n q_{\mathrm{H}}=\delta^n q$ as in Einstein-Maxwell theory. The gravitational energy $\mathcal{E}_\Sigma$ on the Cauchy surface $\Sigma=\mathcal{H}\cup\Sigma_1$ of Fig. 1 is the self-gravitating effect, thus is absent for $n=1$. Moreover, since no wave is induced around $\mathcal{H}$, $\mathcal{E}_\Sigma=\mathcal{E}_{\Sigma_1}$.

Assume the first-law constraint of $n=1$ is optimally done, i.e., $\delta m-\Phi_H\delta q=0$, for the c_4 case, it explicitly gives

$$\delta m=\left[1-\epsilon-\frac{64(2-22\epsilon)c_4^2}{7m^4}\right]\delta q+\mathcal{O}(\epsilon^2) \tag{17}$$

which is different from (9) at $\mathcal{O}(\epsilon c_4^2)$. To evaluate $\mathcal{E}_{\Sigma_1}$ when considering the $n=2$ case, Sorce and Wald assumed that the late-time perturbation $\delta\phi$ approaches a stable linear on shell configuration $\delta\phi^{\mathrm{linear}}$, and one can apply (16) of $n=2$ on Σ_1 with $\delta^2 m=\delta^2 q=0$ so that

$$\mathcal{E}_{\Sigma_1}(\phi;\delta\phi,\mathcal{L}_\xi\phi)=\mathcal{E}_{\Sigma_1}(\phi;\delta\phi^{\mathrm{linear}},\mathcal{L}_\xi\phi)=-T_H\delta^2 S^*, \tag{18}$$

where T_H is the Hawking temperature[5] of the initial black hole, but the variation of Wald's entropy $\delta^2 S^*$ is evaluated at B^* of Fig. 1 with respect to $\phi+\delta\phi^{\mathrm{linear}}$. By construction $\delta^2 S^*=\delta^2 S^*(\delta m,\delta q)$, hence the $n=2$ first-law constraint now takes a second-lawlike form

$$\delta^2 S^*(\delta m,\delta q)+\frac{1}{T_H}(\delta^2 m-\Phi_H\delta^2 q)\geq 0. \tag{19}$$

With the help of (17), for the case c_4 we show explicitly, (19) gives

$$\begin{aligned}\delta^2 m\geq&\left[\frac{1-\epsilon}{m}-\frac{256(1285-9088\epsilon+33261\epsilon^2)c_4^2}{21m^5}\right](\delta q)^2\\&+\left[1-\epsilon+\frac{\epsilon^2}{2}-\frac{64(2-22\epsilon+145\epsilon^2)c_4^2}{7m^4}\right]\delta^2 q,\end{aligned} \tag{20}$$

which is different from (12) at $\mathcal{O}(c_4^2)$. Based on (17) and (20), we can evaluate $W(m(\lambda),q(\lambda))$ for the case c_4 and the result is

$$\begin{aligned}W(\lambda)=&\left(\epsilon\left(\frac{161024c_4^2}{21m^3}+m\right)-\lambda\left(1-\frac{165248c_4^2}{21m^4}\right)\delta q\right)^2\\&-\frac{15360\epsilon^2c_4^2}{m^2}\end{aligned} \tag{21}$$

which cannot be completed the square at $\mathcal{O}(c_4^2)$ to protect WCCC. Similar results for the others cases of $c_{i\neq 4}$ and of c_2 and c_4 up to $\mathcal{O}(c_ic_j)$ can be found in Appendix A.

To conclude our work, in the following we outline a general proof of WCCC based on our second-law proposal.

[4] In Appendix B we check the first law for $n=1$ for the black holes considered in this work.

[5] Due to our convention for m and q by a scale factor $1/4\pi$, here T_H is the scaled Hawking temperature by the same factor.

V. PROOF OF WCCC IN GENERAL

Suppose we have m and q_j, and $m = m_{\rm ex}(q_j)$ is the mass of extremal black holes, with black holes given by $m \geq m_{\rm ex}(q_j)$. Let us define $\mu = m - m_{\rm ex}(q_j)$, which encodes the (deviation from the) extremality condition. Let us also denote $r_h(\mu, q_j)$ the horizon radius, with $R(q_j) = r_h(0, q_j)$ the radius of extremal black holes as a function of q_j. In Appendix D, we argue that for μ inside an open neighborhood of 0,

$$r_h(\mu, q_j) = R(q_j) + \sqrt{\mu}\rho(q_j, \sqrt{\mu}) \tag{22}$$

with ρ a smooth function of its two arguments. Suppose a quantity like the Wald entropy can still be defined in a modified theory of gravity, and that it is expressed as a smooth function of r_h, m, and q_j. Of course, we can also express it in terms of m and q_j, but that expression may not be infinitely smooth in an open neighborhood of the extremal boundary. Let us write

$$S = S(r_h, \mu, q_j) \tag{23}$$

with $\partial S/\partial r_h \neq 0$. For example, the S defined in (4), is of this form. Since $A = 4\pi r_h^2$, and the correction terms are expected to be much less than unity, $\partial S/\partial r_h$ is nonzero. For a family of solutions parametrized by λ, we require that $S(\lambda > 0) \geq S(\lambda = 0)$ still holds, as a generalized second law of black hole thermodynamics.

Let us now start from a configuration with $(\mu, q_j) = (\epsilon^2, q_{j0})$, with $\epsilon > 0$ a small quantity, and deviate away from it with

$$\mu = \epsilon^2 + \delta\mu\lambda + \delta^2\mu\frac{\lambda^2}{2}, \tag{24}$$

$$q_j = q_{j0} + \delta q_j\lambda + \delta^2 q_j\frac{\lambda^2}{2}. \tag{25}$$

Note that the deviation from the extremality is $\mathcal{O}(\epsilon^2)$. We will treat ϵ and λ as quantities with the same order of smallness, and use the fact that $dS/d\lambda$ and $d^2S/d\lambda^2$ should be finite at $\lambda = 0$, as $\epsilon \to 0$. For $dS/d\lambda$, we have a leading contribution of

$$\left.\frac{dS}{d\lambda}\right|_{\lambda=0} \sim \frac{\partial S}{\partial r_h}\frac{\rho}{2\epsilon}\delta\mu, \tag{26}$$

where we have used $\mu = \epsilon^2$ for $\lambda = 0$. Here in order for $dS/d\lambda$ to be finite, we will require $\delta\mu \sim \epsilon$. Inserting this into the second derivative, we obtain

$$\left.\frac{d^2S}{d\lambda^2}\right|_{\lambda=0} \sim \frac{\rho}{2\epsilon^3}\left(\epsilon^2\delta^2\mu - \frac{1}{2}\delta\mu^2\right)\frac{\partial S}{\partial r_h}. \tag{27}$$

From $\partial S/\partial r_h \neq 0$ and since $\delta\mu \sim \epsilon$, this term above is $\sim 1/\epsilon$ unless

$$\delta^2\mu = \frac{\delta\mu^2}{2\epsilon^2}. \tag{28}$$

Inserting (28) back into (24), we obtain

$$\mu = \epsilon^2 + \delta\mu\lambda + \frac{\lambda^2\delta\mu^2}{4\epsilon^2} = \left(\epsilon + \frac{\delta\mu\lambda}{2\epsilon}\right)^2. \tag{29}$$

This ensures that μ stays positive and WCCC holds. Due to lack of the explicit form of S, in the above proof we have only considered the marginal case of the second law, i.e., $\delta S = \delta^2 S = 0$. However, in the explicit examples considered above, we do not need to require the regularity of $\delta^2 S$, so that we can consider the nonmarginal cases, i.e., $\delta^2 S \geq 0$.

The notion of black hole entropy S is well-defined only if the event horizon exists, i.e., $\mu > 0$. This is not assuming what we want to show, as can be understood from the following perspective: for sufficiently small perturbations of a nonextremal black hole, the solution will certainly have a horizon; we can calculate the change in the entropy to the second order in this regime, and use this to show that at this order in perturbation theory μ is positive. Moreover, the second law should be manifested from the underlying dynamical theory. The validity of our proposal implies that WCCC is guaranteed dynamically. The nontrivial part of the proof is that the variation $\delta\mu$ due to the infalling matter is $\mathcal{O}(\epsilon)$ but the initial deviation from the extremality bound is $\mathcal{O}(\epsilon^2)$. It seems that the WCCC can be easily violated, but in fact it is not by requiring the second law. This is in the same spirit of the first-law approach by Sorce-Wald, in which the variation of entropy is assumed and used to evaluate the canonical energy.

VI. DISCUSSION

WCCC is important to protect a gravity theory from the pathology of naked singularity. In this work we propose and show that the second law of black hole thermodynamics ensures WCCC due to the peculiar dependence of the entropy on the extremality condition, and we explicitly demonstrate our proposal for a general class of quartic theories of gravity and electromagnetism.

Naively, we expect to arrive the second law by the first law along with the energy condition of the infalling matter in Wald's gedanken experiment, however we find that this is not the case for our near-extremal charged black hole solutions in higher derivative gravity. In Appendix B we show that the $n = 1$ first law is apparently violated at $\mathcal{O}(c_i^2)$. This might be related to the gauge issue of Wald formalism. For gravity theory with fields with internal gauge freedom, one will expect the first law to be gauge invariant, however the chemical potential Φ_H depends explicitly on the gauge choice. This ambiguity may cause

subtlety when applying the Sorce-Wald formalism straightforwardly to higher derivative gravity. The framework developed by Prabhu [19] using the principal bundle might be helpful to clarify this issue. On the other hand, since the entropy is gauge invariant, we can define the chemical potential as well as the Hawking temperature in terms of the variation of the entropy to derive a gauge invariant first law. This is just what we have done in our second-law approach.

For general nonspherical collapsing case, the construction of the canonical energy would be quite involved in higher derivative gravity, and is crucial to check the second-order first law without source perturbation. Thorough treatment on this issue is expected for future study.

ACKNOWLEDGMENTS

We thank Robert Wald for his helpful comments and discussions. We also thank the anonymous referee for the valuable insight. F. L. L. is supported by National Science and Technology Council (Taiwan), Grant No. 109-2112-M-003-007-MY3. B. N. is supported by the National Natural Science Foundation of China with Grants No. 11975158 and No. 12247103. Y. C. acknowledges the support from the Brinson Foundation, the Simons Foundation (Award No. 568762), and the National Science Foundation, Grants No. PHY-1708212 and No. PHY-1708213.

APPENDIX A: CHECKS OF WCCC FOR QUARTIC DERIVATIVE THEORIES OF GRAVITY AND ELECTROMAGNETISM

We show the explicit forms of the second-order perturbative solutions to the higher derivative theories with only one c_i is turned on, as well as the corresponding details of checking WCCC via both the second-law and the Sorce-Wald formalism.

1. c_1 case

The second-order solutions are solved by extending the procedure in [15] to $O(c_i{}^2)$. For the Lagrangian

$$L = \frac{1}{2\kappa}R - \frac{1}{4}F_{\mu\nu}F^{\mu\nu} + c_1 R^2, \tag{A1}$$

the solution is just the same as the one in Einstein-Maxwell theory, since the Ricci scalar R of the unperturbed background is vanishing hence gives no contribution to the higher-order corrections of energy-momentum tensor. The check for WCCC is also the same as in Einstein-Maxwell theory.

2. c_2 case

For the Lagrangian

$$L = \frac{1}{2\kappa}R - \frac{1}{4}F_{\mu\nu}F^{\mu\nu} + c_2 R_{\mu\nu}R^{\mu\nu}, \tag{A2}$$

the charged black hole solution turns out to be of the form

$$ds^2 = -f(r)dt^2 + \frac{dr^2}{g(r)} + r^2 d\Omega \tag{A3}$$

in which

$$f(r) = 1 - \frac{\kappa m}{r} + \frac{\kappa q^2}{2r^2} + c_2\left(-\frac{2\kappa^2 q^2}{r^4} + \frac{\kappa^3 m q^2}{r^5} - \frac{\kappa^3 q^4}{5r^6}\right) + {c_2}^2\left(\frac{48\kappa^3 q^2}{r^6} - \frac{80\kappa^4 m q^2}{r^7} + \frac{32\kappa^5 m^2 q^2}{r^8} + \frac{240\kappa^4 q^4}{7r^8} - \frac{51\kappa^5 m q^4}{2r^9} + \frac{68\kappa^5 q^6}{15r^{10}}\right), \tag{A4}$$

$$g(r) = 1 - \frac{\kappa m}{r} + \frac{\kappa q^2}{2r^2} + c_2\left(-\frac{4\kappa^2 q^2}{r^4} + \frac{3\kappa^3 m q^2}{r^5} - \frac{6\kappa^3 q^4}{5r^6}\right) + {c_2}^2\left(\frac{144\kappa^3 q^2}{r^6} - \frac{304\kappa^4 m q^2}{r^7} + \frac{160\kappa^5 m^2 q^2}{r^8} + \frac{1192\kappa^4 q^4}{7r^8} - \frac{351\kappa^5 m q^4}{2r^9} + \frac{704\kappa^5 q^6}{15r^{10}}\right) \tag{A5}$$

with the gauge potential

$$A_t = -\frac{q}{r} - c_2\frac{\kappa^2 q^3}{5r^5} + {c_2}^2\left(\frac{48\kappa^3 q^3}{7r^7} - \frac{8\kappa^4 m q^3}{r^8} + \frac{9\kappa^4 q^5}{2r^9}\right). \tag{A6}$$

For simplicity we set $\kappa = 2$ in the following. The existence of double root for either $f(r) = 0$ or $g(r) = 0$ determines the extremal condition

$$m = |q|\left(1 - \frac{4c_2}{5q^2} - \frac{8{c_2}^2}{21q^4}\right), \tag{A7}$$

or

$$|q| = m\left(1 + \frac{4c_2}{5m^2} - \frac{136{c_2}^2}{525m^4}\right). \tag{A8}$$

The location of the horizon is also modified compared to the black hole solution in Einstein-Maxwell theory,

$$\begin{aligned} r_h = {} & r_0 + \frac{4c_2q^2(q^2 - 5mr_0 + 5r_0^2)}{5r_0^3(mr_0 - q^2)} \\ & - \frac{8{c_2}^2q^2}{525r_0^7(mr_0 - q^2)^3}(4571q^8 + 5q^6r_0(3306r_0 - 6881m) + 4200m^2r_0^4(3r_0^2 - 10mr_0 + 8m^2) \\ & - 75mq^2r_0^3(294r_0^2 - 1262mr_0 + 1197m^2) + 5q^4r_0^2(1995r_0^2 - 14004mr_0 + 17269m^2)), \end{aligned} \tag{A9}$$

in which $r_0 = m + \sqrt{m^2 - q^2}$. The Hawking temperature could be obtained by the vanishing of the conical singularity for the corresponding Euclidean black hole,

$$\begin{aligned} T_H = {} & \frac{mr_0 - q^2}{2\pi r_0^3} - \frac{2c_2q^2(3q^2 - 4mr_0)(6q^2 - 10mr_0 + 5r_0^2)}{5\pi r_0^7(mr_0 - q^2)} \\ & + \frac{4{c_2}^2q^2}{525\pi r_0^{11}(mr_0 - q^2)^3}(245840q^{10} - 4q^8r_0(392462m - 84765r_0) \\ & - 8400m^3r_0^5(88m^2 - 74mr_0 + 15r_0^2) + 150m^2q^2r_0^4(20104m^2 - 14216mr_0 + 2275r_0^2) \\ & - 50mq^4r_0^3(97916m^2 - 54966mr_0 + 6195r_0^2) + q^6r_0^2(3943072m^2 - 1575720mr_0 + 93975r_0^2)). \end{aligned} \tag{A10}$$

To obtain the Wald entropy, we first recall the Wald's formula [10,11]

$$S = -2\pi A \frac{\delta L}{\delta R_{\mu\nu\rho\sigma}} \epsilon_{\mu\nu}\epsilon_{\rho\sigma}\Big|_{r_h} \tag{A11}$$

in which $A = 4\pi r_h$ is the area of the horizon. For convenience we will introduce the null coordinate, i.e., define $dv = \sqrt{f/g}dt + dr/g$, the metric (A3) then becomes

$$ds^2 = 2dvdr - g(r)dv^2 + r^2d\Omega, \tag{A12}$$

and the gauge potential $\tilde{A}_\mu$ in the null coordinates are

$$\tilde{A}_v = \sqrt{\frac{g}{f}}A_t, \qquad \tilde{A}_r = -\sqrt{\frac{1}{fg}}A_t. \tag{A13}$$

The Wald's formula then straightforwardly gives rise to

$$S = -2\pi A_h\left(-\frac{1}{\kappa} - 4c_2R^{rv}\right), \tag{A14}$$

which turns out to be

$$S = 4\pi^2 r_0^2 + \frac{32c_2\pi^2 q^2(6q^2 - 10mr_0 + 5r_0^2)}{5r_0^2(mr_0 - q^2)} - \frac{64{c_2}^2\pi^2 q^2}{525r_0^6(mr_0 - q^2)^3}(41972q^8 + 32q^6 r_0(1245r_0 - 6503m) + 4200m^2 r_0^4(3r_0^2 - 16mr_0 + 23m^2) - 75mq^2 r_0^3(301r_0^2 - 2256mr_0 + 4200m^2) + 20q^4 r_0^2(525r_0^2 - 7134mr_0 + 19243m^2)). \tag{A15}$$

For near-extremal black holes, we introduce a small parameter ϵ to characterize the solution in such a way

$$|q| = \sqrt{1-\epsilon^2}m\left(1 + \frac{4c_2}{5m^2} - \frac{136{c_2}^2}{525m^4}\right). \tag{A16}$$

Assuming that the first-order variation is optimally done, i.e., the second law is satisfied marginally

$$\delta S = 0, \tag{A17}$$

we obtain the following relation

$$\delta m = \delta q\left(1 - \epsilon + \frac{\epsilon^2}{2} + \frac{2c_2}{5m^2}(2 + 4\epsilon - 27\epsilon^2) - \frac{4{c_2}^2}{175m^4}(6 + 8936\epsilon - 88661\epsilon^2)\right). \tag{A18}$$

The second-order variation which satisfies the second law

$$\delta^2 S \geq 0 \tag{A19}$$

gives rise to the inequality

$$\delta^2 m \geq \frac{1}{m}(\delta q)^2 + \delta^2 q - \frac{4c_2}{5m^3}(8(\delta q)^2 - m\delta^2 q) + \frac{8{c_2}^2}{525m^5}(27416(\delta q)^2 - 9m\delta^2 q), \tag{A20}$$

in which we have plugged in the relation (A18). From (A7) we know the WCCC is hold if

$$W(m,q) \equiv m^2 - q^2\left(1 - \frac{4c_2}{5q^2} - \frac{8{c_2}^2}{21q^4}\right)^2 \geq 0. \tag{A21}$$

To check (A21), consider a one-parameter family of solutions with $m = m(\lambda)$, $q = q(\lambda)$. Expanding $W(m(\lambda), q(\lambda))$ to $O(\lambda^2)$ and using (A16), (A18), and (A20), we finally get

$$W(\lambda) \geq (\epsilon m - \lambda\delta q)^2 + \frac{8c_2}{5m^2}(\epsilon m - \lambda\delta q)(\epsilon m + 3\lambda\delta q) + \frac{64{c_2}^2}{525m^4}(2\epsilon^2 m^2 - 3351\epsilon\lambda m\delta q + 3433\lambda^2(\delta q)^2), \tag{A22}$$

which could be recast to a perfect square

$$W(\lambda) \geq \left[\epsilon\left(m + \frac{4c_2}{5m} - \frac{104{c_2}^2}{525m^3}\right) - \lambda\left(1 - \frac{12c_2}{5m^2} + \frac{108344{c_2}^2}{525m^4}\right)\delta q\right]^2 + O({c_2}^3), \tag{A23}$$

hence $W(\lambda) \geq 0$ and WCCC is preserved up to $O({c_2}^2)$ by the second law.

On the other hand, according to the Sorce-Wald formalism [9], the first-order variation is optimally done when

$$\delta m = \Phi_h \delta q, \tag{A24}$$

from which we obtain the following relation

$$\delta m = \delta q\left(1 - \epsilon + \frac{\epsilon^2}{2} + \frac{2c_2}{5m^2}(2 + 4\epsilon - 27\epsilon^2) - \frac{4{c_2}^2}{175m^4}(6 + 536\epsilon - 4661\epsilon^2)\right), \tag{A25}$$

which is slightly different from (A18) at $O({c_2}^2)$. The second-order variation inequality,

$$\delta^2 m - \Phi_h \delta^2 q \geq -\frac{T_H}{4\pi}\delta^2 S^*, \tag{A26}$$

combined with (A25) gives raise to

$$\delta^2 m \geq \frac{1}{m}(\delta q)^2 + \delta^2 q - \frac{4c_2}{5m^3}(8(\delta q)^2 - m\delta^2 q) - \frac{8{c_2}^2}{525m^5}(10384(\delta q)^2 + 9m\delta^2 q), \tag{A27}$$

which is also different from (A20). Then $W(\lambda)$ turns out to satisfy

$$W(\lambda) \geq (\epsilon m - \lambda\delta q)^2 + \frac{8c_2}{5m^2}(\epsilon m - \lambda\delta q)(\epsilon m + 3\lambda\delta q) + \frac{64{c_2}^2}{525m^4}(2\epsilon^2 m^2 - 201\epsilon\lambda m\delta q - 1292\lambda^2(\delta q)^2). \tag{A28}$$

The above expression could not be rewritten as a perfect square up to $O({c_2}^2)$, as could be checked by examine the discriminant of the coefficients of λ as in a quadratic equation. The best we can arrive is

$$W(\lambda) \geq \left[\epsilon\left(m + \frac{4c_2}{5m} + \frac{50296{c_2}^2}{525m^3}\right) - \lambda\left(1 - \frac{12c_2}{5m^2} - \frac{42856{c_2}^2}{525m^4}\right)\delta q\right]^2 - \frac{192\epsilon^2{c_2}^2}{m^2}, \tag{A29}$$

hence WCCC is not guaranteed by the Sorce-Wald formalism.

3. c_3 case

For the Lagrangian

$$L = \frac{1}{2\kappa}R - \frac{1}{4}F_{\mu\nu}F^{\mu\nu} + c_3 R_{\mu\nu\rho\sigma}R^{\mu\nu\rho\sigma}, \tag{A30}$$

the solution is

$$f(r) = 1 - \frac{\kappa m}{r} + \frac{\kappa q^2}{2r^2} + c_3\left(-\frac{8\kappa^2 q^2}{r^4} + \frac{4\kappa^3 m q^2}{r^5} - \frac{4\kappa^3 q^4}{5r^6}\right) + {c_3}^2\left(\frac{768\kappa^3 q^2}{r^6} - \frac{1280\kappa^4 m q^2}{r^7} + \frac{512\kappa^5 m^2 q^2}{r^8} + \frac{3840\kappa^4 q^4}{7r^8} - \frac{408\kappa^5 m q^4}{r^9} + \frac{1088\kappa^5 q^6}{15r^{10}}\right), \tag{A31}$$

$$g(r) = 1 - \frac{\kappa m}{r} + \frac{\kappa q^2}{2r^2} + c_3\left(-\frac{16\kappa^2 q^2}{r^4} + \frac{12\kappa^3 m q^2}{r^5} - \frac{24\kappa^3 q^4}{5r^6}\right) + {c_3}^2\left(\frac{2304\kappa^3 q^2}{r^6} - \frac{4864\kappa^4 m q^2}{r^7} + \frac{2560\kappa^5 m^2 q^2}{r^8} + \frac{19072\kappa^4 q^4}{7r^8} - \frac{2808\kappa^5 m q^4}{r^9} + \frac{11264\kappa^5 q^6}{15r^{10}}\right), \tag{A32}$$

$$A_t = -\frac{q}{r} - c_3\frac{4\kappa^2 q^3}{5r^5} + {c_3}^2\left(\frac{768\kappa^3 q^3}{7r^7} - \frac{128\kappa^4 m q^3}{r^8} + \frac{72\kappa^4 q^5}{r^9}\right). \tag{A33}$$

The check for WCCC condition is straightforward just like the previous case, hence we just give the final result. The second law again gives raise to a perfect square

$$W(\lambda) \geq \left[\epsilon\left(m + \frac{16c_3}{5m} - \frac{1664{c_3}^2}{525m^3}\right) - \lambda\left(1 - \frac{48c_3}{5m^2} + \frac{927104{c_3}^2}{525m^4}\right)\delta q\right]^2 + O({c_3}^3), \tag{A34}$$

while the Sorce-Wald formalism gives

$$W(\lambda) \geq \left[\epsilon\left(m + \frac{16c_3}{5m} + \frac{401536{c_3}^2}{525m^3}\right) - \lambda\left(1 - \frac{48c_3}{5m^2} - \frac{282496{c_3}^2}{525m^4}\right)\delta q\right]^2 - \frac{1536\epsilon^2{c_3}^2}{m^2} + O({c_3}^3). \tag{A35}$$

4. c_4 case

For the Lagrangian

$$L = \frac{1}{2\kappa}R - \frac{1}{4}F_{\mu\nu}F^{\mu\nu} + c_4\kappa R F_{\mu\nu}F^{\mu\nu}, \tag{A36}$$

the solution is

$$f(r) = 1 - \frac{\kappa m}{r} + \frac{\kappa q^2}{2r^2} + c_4\left(\frac{4\kappa^2q^2}{r^4} - \frac{6\kappa^3mq^2}{r^5} + \frac{4\kappa^3q^4}{r^6}\right) + {c_4}^2\left(-\frac{32\kappa^4q^4}{7r^8} - \frac{6\kappa^5mq^4}{r^9} + \frac{32\kappa^5q^6}{3r^{10}}\right), \tag{A37}$$

$$g(r) = 1 - \frac{\kappa m}{r} + \frac{\kappa q^2}{2r^2} + c_4\left(-\frac{16\kappa^2q^2}{r^4} + \frac{14\kappa^3mq^2}{r^5} - \frac{6\kappa^3q^4}{r^6}\right) + {c_4}^2\left(\frac{1088\kappa^4q^4}{7r^8} - \frac{126\kappa^5mq^4}{r^9} + \frac{152\kappa^5q^6}{3r^{10}}\right), \tag{A38}$$

$$A_t = -\frac{q}{r} - c_4\frac{2\kappa^2q^3}{r^5} + {c_4}^2\left(\frac{576\kappa^3q^3}{7r^7} - \frac{96\kappa^4mq^3}{r^8} + \frac{50\kappa^4q^5}{r^9}\right). \tag{A39}$$

The second law gives raise to a perfect square for the WCCC condition

$$W(\lambda) \geq \left[\epsilon\left(m - \frac{256{c_4}^2}{21m^3}\right) - \lambda\left(1 + \frac{211072{c_4}^2}{21m^4}\right)\delta q\right]^2 + O({c_4}^3), \tag{A40}$$

while the Sorce-Wald formalism gives

$$W(\lambda) \geq \left[\epsilon\left(m + \frac{161024{c_4}^2}{21m^3}\right) - \lambda\left(1 - \frac{165248{c_4}^2}{21m^4}\right)\delta q\right]^2 - \frac{15360\epsilon^2{c_4}^2}{m^2} + O({c_4}^3). \tag{A41}$$

5. c_5 case

For the Lagrangian

$$L = \frac{1}{2\kappa}R - \frac{1}{4}F_{\mu\nu}F^{\mu\nu} + c_5\kappa R_{\mu\nu}F^{\mu\rho}F^{\nu}{}_{\rho}, \tag{A42}$$

the solution is

$$f(r) = 1 - \frac{\kappa m}{r} + \frac{\kappa q^2}{2r^2} + c_5\left(-\frac{\kappa^3mq^2}{r^5} + \frac{4\kappa^3q^4}{5r^6}\right) + {c_5}^2\left(-\frac{12\kappa^4q^4}{7r^8} + \frac{9\kappa^5mq^4}{2r^9} - \frac{164\kappa^5q^6}{45r^{10}}\right), \tag{A43}$$

$$g(r) = 1 - \frac{\kappa m}{r} + \frac{\kappa q^2}{2r^2} + c_5\left(-\frac{6\kappa^2q^2}{r^4} + \frac{5\kappa^3mq^2}{r^5} - \frac{11\kappa^3q^4}{5r^6}\right) + {c_5}^2\left(\frac{548\kappa^4q^4}{7r^8} - \frac{139\kappa^5mq^4}{2r^9} + \frac{284\kappa^5q^6}{9r^{10}}\right), \tag{A44}$$

$$A_t = -\frac{q}{r} - c_5\frac{\kappa^2q^3}{5r^5} + {c_5}^2\left(\frac{48\kappa^3q^3}{7r^7} - \frac{8\kappa^4mq^3}{r^8} + \frac{43\kappa^4q^5}{6r^9}\right). \tag{A45}$$

The second law gives raise to a perfect square

$$W(\lambda) \geq \left[\epsilon\left(m + \frac{4c_5}{5m} - \frac{11512{c_5}^2}{1575m^3}\right) - \lambda\left(1 - \frac{12c_5}{5m^2} + \frac{2469832{c_5}^2}{1575m^4}\right)\delta q\right]^2 + O({c_5}^3), \tag{A46}$$

while the Sorce-Wald formalism gives

$$W(\lambda) \geq \left[\epsilon\left(m + \frac{4c_5}{5m} + \frac{1802888{c_5}^2}{1575m^3}\right) - \lambda\left(1 - \frac{12c_5}{5m^2} - \frac{1763768{c_5}^2}{1575m^4}\right)\delta q\right]^2 - \frac{2304\epsilon^2{c_5}^2}{m^2} + O({c_5}^3). \tag{A47}$$

6. c_6 case

For the Lagrangian

$$L = \frac{1}{2\kappa}R - \frac{1}{4}F_{\mu\nu}F^{\mu\nu} + c_6\kappa R_{\mu\nu\rho\sigma}F^{\mu\rho}F^{\rho\sigma}, \tag{A48}$$

the solution is

$$\begin{aligned} f(r) = 1 &- \frac{\kappa m}{r} + \frac{\kappa q^2}{2r^2} + c_6\left(-\frac{2\kappa^2q^2}{r^4} + \frac{\kappa^3mq^2}{r^5} - \frac{\kappa^3q^4}{5r^6}\right) \\ &+ {c_6}^2\left(-\frac{320\kappa^4mq^2}{7r^7} + \frac{128\kappa^5m^2q^2}{7r^8} + \frac{530\kappa^4q^4}{7r^8} - \frac{411\kappa^5mq^4}{14r^9} - \frac{47\kappa^5q^6}{15r^{10}}\right), \end{aligned} \tag{A49}$$

$$\begin{aligned} g(r) = 1 &- \frac{\kappa m}{r} + \frac{\kappa q^2}{2r^2} + c_6\left(-\frac{8\kappa^2q^2}{r^4} + \frac{7\kappa^3mq^2}{r^5} - \frac{16\kappa^3q^4}{5r^6}\right) \\ &+ {c_6}^2\left(-\frac{320\kappa^4mq^2}{r^7} + \frac{2048\kappa^5m^2q^2}{7r^8} + \frac{4352\kappa^4q^4}{7r^8} - \frac{1413\kappa^5mq^4}{2r^9} + \frac{3976\kappa^5q^6}{15r^{10}}\right), \end{aligned} \tag{A50}$$

$$A_t = -\frac{q}{r} + c_6\left(-\frac{2\kappa^2mq}{r^4} + \frac{9\kappa^2q^3}{5r^5}\right) + {c_6}^2\left(-\frac{64\kappa^4m^2q}{7r^7} - \frac{160\kappa^3q^3}{7r^7} + \frac{216\kappa^4mq^3}{7r^8} - \frac{9\kappa^4q^5}{10r^9}\right). \tag{A51}$$

The second law gives raise to a perfect square

$$W(\lambda) \geq \left[\epsilon\left(m + \frac{4c_6}{5m} - \frac{10504{c_6}^2}{525m^3}\right) - \lambda\left(1 - \frac{52c_6}{5m^2} + \frac{1850344{c_6}^2}{525m^4}\right)\delta q\right]^2 + O({c_6}^3), \tag{A52}$$

while the Sorce-Wald formalism gives

$$W(\lambda) \geq \left[\epsilon\left(m + \frac{4c_6}{5m} + \frac{1199096{c_6}^2}{525m^3}\right) - \lambda\left(1 - \frac{52c_6}{5m^2} - \frac{972056{c_6}^2}{525m^4}\right)\delta q\right]^2 - \frac{4608\epsilon^2{c_6}^2}{m^2} + O({c_6}^3). \tag{A53}$$

7. c_7 case

For the Lagrangian

$$L = \frac{1}{2\kappa}R - \frac{1}{4}F_{\mu\nu}F^{\mu\nu} + c_7\kappa^2F_{\mu\nu}F^{\mu\nu}F_{\rho\sigma}F^{\rho\sigma}, \tag{A54}$$

the solution is

$$f(r) = g(r) = 1 - \frac{\kappa m}{r} + \frac{\kappa q^2}{2r^2} - \frac{4c_7\kappa^3q^4}{5r^6} + \frac{128{c_7}^2\kappa^5q^6}{9r^{10}}, \tag{A55}$$

$$A_t = -\frac{q}{r} + \frac{16c_7\kappa^2 q^3}{5r^5} - \frac{256{c_7}^2\kappa^4 q^5}{3r^9}. \tag{A56}$$

In this case, the second-law and the Sorce-Wald formalism give rise to the same perfect square

$$W(\lambda) \geq \left[\epsilon\left(m + \frac{16c_7}{5m} - \frac{48256{c_7}^2}{225m^3}\right) - \lambda\left(1 - \frac{48c_7}{5m^2} + \frac{236416{c_7}^2}{225m^4}\right)\delta q\right]^2 + O({c_7}^3). \tag{A57}$$

8. c_8 case

For the Lagrangian

$$L = \frac{1}{2\kappa}R - \frac{1}{4}F_{\mu\nu}F^{\mu\nu} + c_8\kappa^2 F_{\mu\nu}F^{\nu\rho}F_{\rho\sigma}F^{\sigma\mu}, \tag{A58}$$

the solution is

$$f(r) = g(r) = 1 - \frac{\kappa m}{r} + \frac{\kappa q^2}{2r^2} - \frac{2c_8\kappa^3 q^4}{5r^6} + \frac{32{c_8}^2\kappa^5 q^6}{9r^{10}}, \tag{A59}$$

$$A_t = -\frac{q}{r} + \frac{8c_8\kappa^2 q^3}{5r^5} - \frac{64{c_8}^2\kappa^4 q^5}{3r^9}. \tag{A60}$$

In this case, the second-law and the Sorce-Wald formalism again give rise to the same perfect square

$$W(\lambda) \geq \left[\epsilon\left(m + \frac{8c_8}{5m} - \frac{12064{c_8}^2}{225m^3}\right) - \lambda\left(1 - \frac{24c_8}{5m^2} + \frac{59104{c_8}^2}{225m^4}\right)\delta q\right]^2 + O({c_8}^3). \tag{A61}$$

9. $c_2 + c_4$ case

We also consider the case with both c_2 and c_4 terms are turned on, i.e.,

$$L = \frac{1}{2\kappa}R - \frac{1}{4}F_{\mu\nu}F^{\mu\nu} + c_2 R_{\mu\nu}R^{\mu\nu} + c_4\kappa R F_{\mu\nu}F^{\mu\nu}, \tag{A62}$$

the solution for which is a bit tedious, hence we just show the final result for WCCC. The second law still gives a perfect square

$$W(\lambda) \geq \left[\epsilon\left(m + \frac{4c_2}{5m} - \frac{8\times(13{c_2}^2 + 800{c_4}^2)}{525m^3}\right) - \lambda\left(1 - \frac{12c_2}{5m^2} + \frac{8\times(13543{c_2}^2 + 193200c_2c_4 + 659600{c_4}^2)}{525m^4}\right)\delta q\right]^2 + O({c_i}^3), \tag{A63}$$

while the Sorce-Wald formalism gives

$$W(\lambda) \geq \left[\epsilon\left(m + \frac{4c_2}{5m} - \frac{8\times(6287{c_2}^2 + 113400c_2c_4 + 503200{c_4}^2)}{525m^3}\right) - \lambda\left(1 - \frac{12c_2}{5m^2} + \frac{8\times(5357{c_2}^2 + 113400c_2c_4 + 516400{c_4}^2)}{525m^4}\right)\delta q\right]^2 - \frac{192\epsilon^2({c_2}^2 + 18c_2c_4 + 80{c_4}^2)}{m^2} + O({c_i}^3). \tag{A64}$$

APPENDIX B: CHECK OF FIRST LAW UP TO $O(c_i)$

The standard first law of black hole thermodynamics without source perturbation is to set $\delta^n q_{\rm H} = \delta^n T_{ab} = 0$ in Eq. (16) of the main text. That is,

$$\delta^n m_{\rm ADM} - \Phi_{\rm H}\delta^n q_B - T_H\delta^n S_B = \delta_{n,2}\mathcal{E}_\Sigma(\phi;\delta\phi,\mathcal{L}_\xi\phi). \tag{B1}$$

Since we do not have the explicit form of the canonical energy $\mathcal{E}_\Sigma(\phi;\delta\phi,\mathcal{L}_\xi\phi)$, we will check only the $n=1$ first law for the black hole solutions considered in this work. As argued in the main text, the higher derivative corrections will not affect the relation $\delta^m_{\rm ADM} = \delta^n m$ due to the higher powers of $1/r$ suppression at infinity. For $\delta^n q_B = \delta^n(\int_B \epsilon_{abcd}S^{cd})$ with $S^{ab} = F^{ab} + \mathcal{O}(c_i)$, we can either evaluate q_B at the bifurcation sphere B, i.e., $q_B = \sqrt{g^{tt}g^{rr}}r^2 S_{rt}|_{r=r_+}$, or at spatial infinity by adopting Gauss theorem and the no-source assumption, it is then straightforward to see $\delta^n q_B = \delta^n q$. Therefore, the $n=1$ first law implies

$$\Delta_{\rm I} := \frac{\partial S_B}{\partial q} + \Phi_{\rm H}\frac{\partial S_B}{\partial m} = 0. \tag{B2}$$

For all cases of black hole solutions up to $\mathcal{O}(c_i^2)$ considered in the main text, we find that

$$\Delta_{\rm I} = 0 + \mathcal{O}(c_i^2). \tag{B3}$$

The $\mathcal{O}(c_i^2)$ terms are finite in extremal limit. For example, for the c_2 case,

$$\Delta_{\rm I}^{(c_2)} = \frac{1536\pi^2(1-8\epsilon+\mathcal{O}(\epsilon^2))}{M^3} c_2^2 + \mathcal{O}(c_i^3), \tag{B4}$$

where ϵ is the small nonextremility parameter.

From the above, we can conclude that all the black hole solutions considered in this work satisfy the $n=1$ first law up to $\mathcal{O}(c_i)$. This fact is crucial in [13] to prove the WCCC for the extremal black holes of generic gravity theories by the Sorce-Wald formalism. The $\mathcal{O}(c_i^2)$ violation for the $n=1$ first law should be responsible for the failure of the Sorce-Wald formalism to yield the WCCC for near-extremal black holes as discussed in the main text. The check of $n \geq 2$ sourceless first law for modified gravity theories is beyond the scope of this work because it needs the explicit construction of the canonical energy.

APPENDIX C: IMPLICATION ON WCCC FROM AFLOAT SPHERICAL THIN SHELL

Since we are considering the spherical collapsing to avoid the complications due to the electromagnetic and gravitational radiations, the simplest example is the spherical collapsing shell. To find the implication on WCCC condition, we will consider the spherical thin-shell matter afloat in the spacetime described by the metric (A3) with f and g given in Sec. A up to $\mathcal{O}(c_i)$.

The motion of the thin shell around a black hole obeys the generalized Israel junction conditions [20], which can be obtained from the Gauss-Codazzi equations. However, the junction conditions for thin shell are in general highly singular in the higher derivative gravity theories except for the Gauss-Bonnet higher derivative term, see for example [21–24] for discussions. To have the regular junction conditions to yield sensible motion of thin shell, we need to impose regularity conditions on the metric around the thin shell. For more singular junction conditions, it means more regularity conditions on the metric should be imposed so that mostly it will yield only trivial solutions, i.e., no thin shell. See Sec. C 1 for some discussion. Below we will only consider the thin shell in Einstein and Einstein-Gauss-Bonnet gravities.

The junction conditions for Einstein gravity are given by [20] (set $\kappa = 2$)

$$[K_{\mu\nu} - h_{\mu\nu}K]_J = -S_{\mu\nu}, \tag{C1}$$

where $K_{\mu\nu}$ is the extrinsic curvature, $K \equiv K^\mu_\nu$, $h_{\mu\nu}$ is the induced metric on the thin shell, and S_ab is the stress tensor of the thin-shell matter. Here $[A]_J$ denotes taking difference of the quantity A on the both sides of the thin shell. Assume the spherical thin shell is located at $r = r_s$ with stress tensor $S^\mu_\nu = \mathrm{diag}(\rho, 0, p, p)$, and evaluate the extrinsic curvature with the metric (A3), the junction conditions give

$$[g]_J = -\frac{r_s}{2}\rho, \tag{C2}$$

and

$$[\sqrt{g}(2 + rf'/f)]_J = 2r_s p, \tag{C3}$$

where $f' \equiv \partial_r f$. It turns out that the above junction conditions involve only g, f, and f'. To have a finite jump on the left sides of the above junction conditions, we only need to impose the piecewise continuity of f at $r = r_s$ to yield sensible and nontrivial junction conditions. For this purpose, we choose to rescale the coordinate time so that the metrics on both sides of the thin shell are given by

$$f_+(r) = g_+(r) = 1 - 2\frac{m_+}{r} + \frac{q_+^2}{r^2}, \tag{C4}$$

but

$$\frac{1 - 2\frac{m_-}{r_s} + \frac{q_-^2}{r_s^2}}{1 - 2\frac{m_+}{r_s} + \frac{q_+^2}{r_s^2}} f_-(r) = g_-(r) = 1 - 2\frac{m_-}{r} + \frac{q_-^2}{r^2}. \tag{C5}$$

Note that $f_+(r_s) = f_-(r_s)$. We assume the matter shell is pressure-less, i.e., $p = 0$, then the junction condition (C3) can be turned into the following condition

$$m_+^2 - q_+^2 = \left(\frac{r_s - m_+}{r_s - m_-}\right)^2 (m_-^2 - q_-^2). \tag{C6}$$

This condition implies that a sub-extremal black hole with $m_-^2 > q_-^2$ remains subextremal, i.e., $m_+^2 > q_+^2$ even after throwing a pressureless spherical thin shell. This is consistent with WCCC.

Next we will show that the same condition also holds for the Einstein-Gauss-Bonnet (EGB) gravity. Note that for EGB gravity, we shall introduce the coupling $c_{\rm EGB}$ of the Gauss-Bonnet term, which is nothing but $c_{\rm EGB} = c_1 = c_2 = -\frac{1}{4}c_2$. The junction condition for EGB gravity are different from the one for Einstein gravity and are given by [21]

$$[K_{\mu\nu} - h_{\mu\nu}K + 2c_{\rm EGB}(3J_{\mu\nu} - h_{\mu\nu}J + 2\hat{P}_{\mu\rho\lambda\nu}K^{\rho\lambda})]_J = -S_{\mu\nu}, \tag{C7}$$

where

$$J_{\mu\nu} = \frac{1}{3}(2KK_{\mu\rho}K^\rho_\nu + K_{\rho\lambda}K^{\rho\lambda}K_{\mu\nu} - 2K_{\mu\rho}K^{\rho\lambda}K_{\lambda\nu} - K^2K_{\mu\nu}). \tag{C8}$$

$J \equiv J^\mu_\nu$ and

$$\hat{P}_{\mu\nu\rho\lambda} = \hat{R}_{\mu\nu\rho\lambda} + 2\hat{R}_{\nu[\rho}h_{\lambda]\mu} - 2\hat{R}_{\mu[\rho}h_{\lambda]\nu} + \hat{R}_{\nu[\rho}h_{\lambda]\mu} + h_{\mu[\rho}h_{\lambda]\nu}\hat{R}, \tag{C9}$$

where the hatted quantities are the associated unhatted quantities evaluated with respect to the induced metric $h_{\mu\nu}$. The novelty of the junction condition (C7) is only the first derivatives of the metric are involved. Using the metric (A3) and the induced metric for the spherical thin shell, we can find that $\hat{P}_{\mu\nu\rho\lambda} = 0$ and $3J_{\mu\nu} - h_{\mu\nu}J = 0$ even though $J_{\mu\nu}$ and J are nonzero. Based on the above, the junction condition (C7) for EGB gravity is indeed reduced to (C1) for Einstein gravity. Moreover, the Gauss-Bonnet term is a total derivative term so that it will not affect the field equation, and the black hole solutions are the same as the one for Einstein-Maxwell theory. In total, the junction condition of the EGB gravity will still yield the same constraint (C6) for the spherical thin shell. That is, the thin shell will not turn a subextremal black hole into a naked singularity in the EGB gravity. This is consistent with our result for EGB gravity as discussed in the main text.

1. No thin shell from a third-order junction condition

In [23] a set of third-order junction conditions for the higher derivative gravity theories have been proposed. This junction condition is obtained by collecting the singular terms in the Gauss-Codazzi equations. For the quartic action of gravity considered in the main text, the junction conditions take the following form:

$$[W^{\mu}_{\nu\rho}]_J n^{\rho} = 2S^{\mu}_{\nu}, \tag{C10}$$

where n^{μ} is the normal vector of the thin shell, and

$$W^{\mu}_{\nu\rho} = c_1 R_{;\lambda}(2\delta^{\mu}_{\nu}\delta^{\lambda}_{\rho} - g^{\mu\lambda}g_{\nu\rho} - \delta^{\lambda}_{\nu}\delta^{\mu}_{\rho}) + c_2(R^{\mu}_{\nu;\rho} + \frac{1}{2}g^{\mu}_{\nu}\hat{R}_{;\rho} - R^{\mu}_{\rho;\nu} - R^{;\mu}_{\rho\nu}) - 4c_3 R^{\mu\lambda}_{\rho\nu;\lambda}. \tag{C11}$$

After explicitly evaluating the left side of (C10) with respect to the metric (A3), the result involves the terms with g'' and f'''. This implies that we need to impose the continuity conditions $[f]_J = [g]_J = [f']_J = [g']_J = [f'']_J = 0$ to yield a finite left side of (C10), thus a sensible junction condition with finite $S^{\mu\nu}$. Since the junction condition (C10) is already $\mathcal{O}(c_i)$, thus the metric used to evaluate the junction condition should be kept up to the leading order only. Thus, the metrics on both sides of the thin shell contain in total only four integration constants, i.e., $m_{\pm}$ and $q_{\pm}$. The above five continuity conditions are overconstrained on these integration constants and can be shown only to yield trivial solutions, namely, $m_+ = m_-$ and $q_+ = q_-$. This implies no sensible thin shell for generic quartic gravities.

APPENDIX D: DEPENDENCE OF r_h ON m AND q_j NEAR THE EXTREMAL BOUNDARY

Suppose we have m and q_j, and $m = m_{\rm ex}(q_j)$ is the mass of extremal black holes, with black holes given by $m \geq m_{\rm ex}(q_j)$. Let us define $\mu = m - m_{\rm ex}(q_j)$. Let us also denote $r_h(\mu, q_j)$ the horizon radius, with $R(q_j) = r_h(0, q_j)$ the radius of extremal black holes as a function of q_j. Let us first show that for μ inside an open neighborhood of 0,

$$r_h(\mu, q_j) = R(q_j) + \sqrt{\mu}\rho(q_j, \sqrt{\mu}) \tag{D1}$$

with ρ a smooth function of its arguments. In GR, the horizon radius is the greater real root of the quadratic polynomial $r_h^2 - 2Mr_h + Q^2 = 0$,

$$r_{\pm} = M \pm \sqrt{M^2 - Q^2} \tag{D2}$$

with extremal boundary given by the location of the double root.

In modified GR, we can view the (r_h, μ, q_j) relation in two ways. We can directly write a condition of

$$\Delta(r_h, \mu, q_j) = 0 \tag{D3}$$

with extremal condition given by

$$\partial_r \Delta|_{R(q_j),0,q_j} = 0. \tag{D4}$$

As we expand around $\mu = 0$, and write r_h as an expansion,

$$r_h = R(q_j) + \delta r_h, \tag{D5}$$

we have

$$0 = \Delta(R(q_j) + \delta r_h, \mu, q_j) = \frac{1}{2}\frac{\partial^2\Delta}{\partial r^2}\bigg|_{R(q_j),0,q_j}(\delta r_h)^2 + \frac{\partial\Delta}{\partial\mu}\bigg|_{R(q_j),0,q_j}\mu + \ldots. \tag{D6}$$

From this, we can reexpand δr_h in terms of $\sqrt{\mu}$,

$$\delta r_h = \sqrt{\mu}\sum_{n=0}^{+\infty}\alpha_n(q_j)\mu^{n/2} = \sqrt{\mu}\rho(q_j, \sqrt{\mu}). \tag{D7}$$

Another way to write this, is to view Eq. (D3) as a definition of μ in terms of r_h and q_j. In the GR case, we have

$$M(r_h, Q) = \frac{r_h^2 + Q^2}{2r_h}. \tag{D8}$$

For each Q, we generically have two values of r_h that gives rise to M—only the larger value correspond to the outer horizon. Extremal black holes are when M takes a minimum at $r_h = Q$. In the modified gravity case, expecting the structure of the problem to remain unchanged, namely the fact that m is uniquely determined by r_h and q_j and the fact

that m is a minimum when $r_h(q_j) = R(q_j)$. We need to assume that when deviating away from the minimum, the value of m depends quadratically on $r - R(q_j)$,

$$\mu(r, q_j) = m(r, q_j) - m(R(q_j), q_j) = [r - R(q_j)]^2 F(r, q_j) \tag{D9}$$

with $F(r, q_j)$ a smooth, nonzero function in an open neighborhood of $(R(q_j), q_j)$. We can then write

$$r = R(q_j) + \frac{\sqrt{\mu}}{\sqrt{F(r, q_j)}}. \tag{D10}$$

This can be solved iteratively to yield Eq. (D1).

[1] R. Penrose, Gravitational Collapse and Space-Time Singularities, Phys. Rev. Lett. **14**, 57 (1965).
[2] R. Penrose, Gravitational collapse: The role of general relativity, Riv. Nuovo Cimento **1**, 252 (1969); Gen. Relativ. Gravit. **34**, 1141 (2002).
[3] J. D. Bekenstein, Black holes and entropy, Phys. Rev. D **7**, 2333 (1973).
[4] J. D. Bekenstein, Generalized second law of thermodynamics in black hole physics, Phys. Rev. D **9**, 3292 (1974).
[5] S. W. Hawking, Particle creation by black holes, Commun. Math. Phys. **43**, 199 (1975); **46**, 206(E) (1976).
[6] D. Christodoulou, Reversible and Irreversible Transforations in Black Hole Physics, Phys. Rev. Lett. **25**, 1596 (1970).
[7] S. W. Hawking, Gravitational Radiation from Colliding Black Holes, Phys. Rev. Lett. **26**, 1344 (1971).
[8] S. W. Hawking, Black holes in general relativity, Commun. Math. Phys. **25**, 152 (1972).
[9] J. Sorce and R. M. Wald, Gedanken experiments to destroy a black hole. II. Kerr-Newman black holes cannot be overcharged or overspun, Phys. Rev. D **96**, 104014 (2017).
[10] R. M. Wald, Black hole entropy is the Noether charge, Phys. Rev. D **48**, R3427 (1993).
[11] V. Iyer and R. M. Wald, Some properties of Noether charge and a proposal for dynamical black hole entropy, Phys. Rev. D **50**, 846 (1994).
[12] R. M. Wald, Gedanken experiments to destroy a black hole, Ann. Phys. (N.Y.) **82**, 548 (1974).
[13] B. Chen, F. L. Lin, B. Ning, and Y. Chen, Constraints on Low-Energy Effective Theories from Weak Cosmic Censorship, Phys. Rev. Lett. **126**, 031102 (2021); **126**, 119903(E) (2021).
[14] D. Christodoulou, The instability of naked singularities in the gravitational collapse of a scalar field, Ann. Math. **149**, 183 (1999).
[15] Y. Kats, L. Motl, and M. Padi, Higher-order corrections to mass-charge relation of extremal black holes, J. High Energy Phys. 12 (2007) 068.
[16] 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.
[17] V. E. Hubeny, Overcharging a black hole and cosmic censorship, Phys. Rev. D **59**, 064013 (1999).
[18] S. Hollands and R. M. Wald, Stability of black holes and black branes, Commun. Math. Phys. **321**, 629 (2013).
[19] K. Prabhu, The first law of black hole mechanics for fields with internal gauge freedom, Classical Quantum Gravity **34**, 035011 (2017).
[20] W. Israel, Singular hypersurfaces and thin shells in general relativity, Nuovo Cimento B **44S10**, 1 (1966); **48**, 463(E) (1967).
[21] S. C. Davis, Generalized Israel junction conditions for a Gauss-Bonnet brane world, Phys. Rev. D **67**, 024030 (2003).
[22] N. Deruelle and T. Dolezel, Brane versus shell cosmologies in Einstein and Einstein-Gauss-Bonnet theories, Phys. Rev. D **62**, 103502 (2000).
[23] A. Balcerzak and M. P. Dabrowski, Generalized Israel junction conditions for a fourth-order brane world, Phys. Rev. D **77**, 023524 (2008).
[24] C. S. Chu and H. S. Tan, Generalized Darmois–Israel junction conditions, Universe **8**, 250 (2022).