Revisiting loop corrections in single field ultraslow-roll inflation

Hassan Firouzjahi*

School of Astronomy, Institute for Research in Fundamental Sciences (IPM), P. O. Box 19395-5531, Tehran, Iran

(Received 9 November 2023; accepted 23 January 2024; published 9 February 2024)

We revisit the one-loop corrections on cosmic microwave background scale perturbations induced from small-scale modes in single field models which undergo a phase of ultraslow-roll inflation. There were concerns that large loop corrections are against the notion of the decoupling of scales and that they are canceled out once the boundary terms are included in the Hamiltonian. We highlight that the nonlinear coupling between the long and short modes and the modulation of the short mode power spectrum by the long mode are the key physical reasons behind the large loop corrections. In particular, in order for the modulation by the long mode to be significant there should be a strong scale-dependent enhancement in the power spectrum of the short mode, which is the hallmark of the ultraslow-roll inflation. We highlight the important roles played by the would-be decaying mode which were not taken into account properly in recent works claiming loop cancellation. We confirm the original conclusion that loop corrections are genuine and that they can be dangerous for primordial black hole formation unless the transition to the final attractor phase is mild.

DOI: 10.1103/PhysRevD.109.043514

I. INTRODUCTION

The question involving one-loop corrections in the power spectrum in models of single field inflation containing an intermediate phase of ultraslow-roll (USR) inflation has attracted considerable interest recently [1–20]. The models incorporating a phase of USR have been employed as a mechanism to enhance the power spectrum on small scales to source the primordial black holes (PBHs) as a candidate for dark matter [21–23]; for a review see [24–26].

It was argued in [1,2] that the one-loop corrections induced from small USR modes can significantly affect the observed cosmic microwave background (CMB) scale perturbations. Therefore, in order to keep these loop corrections under perturbative control, it was argued that the model is not trusted to generate the desired PBHs abundance. On the other hand, this conclusion was criticized in [3,4], where it was argued that this conclusion cannot be viewed as a no-go theorem and that the dangerous one-loop corrections can be harmless in a smooth transition. This question was studied in further detail in [9], in which the effects of both cubic and quartic Hamiltonians were included. In addition, the effects of the sharpness of the transition from the intermediate USR phase to the final attractor phase were highlighted as well. The analysis in [9] supports the conclusion of [1] when the transition from the USR phase to the final attractor phase is sharp. However, it was argued in [9] that the dangerous one-loop corrections can be washed out in a mild transition. This question was also studied in [11], where the one-loop corrections was calculated using the δN formalism. It was shown in [11] that for a mild transition the one-loop corrections are suppressed by the slow-roll parameters, so the setup is still reliable for PBH formations.

On the other hand, the question of loop corrections was revisited in [27,28], in which it was claimed that the one-loop correction cancels in the setup of interest. Specifically, in [27] the roles of boundary terms which were not incorporated into [1,2] and the following works were highlighted. On the other hand, in [28], relying on the Maldacena consistency condition [29], it was argued that the large loop corrections are canceled once the UV limit of the momentum is taken care of by an appropriate $i\epsilon$ prescription. Note that in both [27,28] only the cubic interactions were considered and, as in many other previous works, the contributions of the quartic interactions were not considered.

In this work first we study the physical origins of the large loop corrections in this setup. This is more important, as the existence of large loop corrections on long CMB scales induced by small scales is somewhat counterintuitive. One may argue that the existence of large loop corrections is against the common sense of "naturalness" and the concept of "decoupling of scales." In addition, we revisit the claims in [27,28] that the loop corrections cancel and highlight some conceptual and technical points with which we disagree. We conclude that the large loop corrections on CMB scales are genuine and can be dangerous if the transition to the final attractor phase is sharp.

*firouz@ipm.ir

II. USR INFLATION SETUP

The USR setup is a model of inflation in which the potential is flat [30–32]. Originally, the USR setup attracted interest as a nontrivial example for the violation of the Maldacena non-Gaussianity consistency condition [29,33]. Since the potential is flat in the USR setup, the inflaton velocity falls off exponentially and the curvature perturbations grow on superhorizon scales [34]. The enhancement of curvature perturbations on superhorizon scales is the key behind the violation of the Maldacena consistency condition in the USR setup [34–45]. The amplitude of the local-type non-Gaussianity in the USR model is calculated in [34] to be $f_{NL} = \frac{5}{2}$. This question was further investigated in [46], where it was shown that the final amplitude of f_{NL} depends on the sharpness of the transition from the USR phase to the final slow-roll (SR) phase. In particular, in the example of an extremely sharp transition from the USR phase to the SR phase, as considered in [34], f_{NL} acquires its maximum value $\frac{5}{2}$. However, for a mild transition the curvature perturbations evolve after the USR phase until it reaches its final attractor value. As a result, much of the amplitude of f_{NL} is washed out toward the end of inflation. The important lesson is that the sharpness of the transition from the USR phase to the final SR phase plays an important role when looking at the final amplitude of the cosmological observables.

The setup we study here, as in [1,2], comprises three phases of inflation, $SR \to USR \to SR$, with a single field inflation driven by the scalar field ϕ with the potential $V(\phi)$. The first stage of inflation is in the SR phase during which the large CMB scales leave the horizon. This period may take 20–30 e-folds, depending on the mass and abundance of the PBHs. The curvature perturbation is nearly scale invariant and Gaussian with an amplitude fixed by the Cosmic Background Explorer (COBE) normalization. The second stage is the USR phase in which the potential becomes exactly flat and the curvature perturbation grows exponentially to seed the PBH formation on small scales. Typically, the duration of the USR phase is assumed to be about a few e-folds to obtain a sizable fraction of the dark matter from the PBH formation. The USR phase is glued to a second SR phase, which is the final stage of inflation. Depending on the model parameters, the transition to the final attractor phase can be either mild or sharp, which plays an important role in the amplitude of the loop corrections.

Starting with the Friedmann-Lemaître-Robertson-Walker metric

$$ds^2 = -dt^2 + a(t)^2 d\mathbf{x}^2, \tag{1}$$

the inflaton field equation in the USR phase is given by

$$\ddot{\phi}(t) + 3H\dot{\phi}(t) = 0, \qquad 3M_P^2 H^2 \simeq V_0, \tag{2}$$

in which M_P is the reduced Planck mass, H is the Hubble rate during inflation, and V_0 is the value of the potential during the USR phase. Since V_0 is constant, H is nearly constant while $\dot{\phi} \propto a^{-3}$ during the USR phase. The slow-roll parameters related to H are defined as follows:

$$\epsilon \equiv -\frac{\dot{H}}{H^2} = \frac{\dot{\phi}^2}{2M_P^2 H^2}, \qquad \eta \equiv \frac{\dot{\epsilon}}{H\epsilon}. \tag{3}$$

During the SR phases both ϵ and η are nearly constant and small. However, during the USR phase, ϵ falls off like a^{-6} while $\eta \simeq -6$, which is the hallmark of USR inflation [30]. Going to conformal time $d\tau = dt/a(t)$ with $aH\tau \simeq -1$, $\epsilon(\tau)$ is given as

$$\epsilon(\tau) = \epsilon_i \left(\frac{\tau}{\tau_s}\right)^6, \tag{4}$$

in which ϵ_i is the value of ϵ prior to the USR phase. We assume that the USR phase is extended during the period $\tau_s < \tau < \tau_e$, so ϵ at the end of the USR phase is $\epsilon_e = \epsilon_i (\frac{\tau_e}{\tau_s})^6$. When the number of e-folds is defined as $dN = Hdt$, the duration of the USR phase is given by $\Delta N \equiv N(\tau_e) - N(\tau_s)$, so $\epsilon_e = \epsilon_i e^{-6\Delta N}$.

As in [46], suppose the potential after the USR phase supports a period of SR inflation such that

$$V(\phi) = V(\phi_e) + \sqrt{2\epsilon_V} V(\phi_e)(\phi - \phi_e) + \frac{\eta_V}{2} V(\phi_e)(\phi - \phi_e)^2 + \cdots. \tag{5}$$

Here $2\epsilon_V \equiv M_P^2 (V'(\phi_e)/V(\phi_e))^2$ and $\eta_V \equiv M_P^2 V''(\phi_e)/V(\phi_e)$ are the usual slow-roll parameters defined in terms of the first and second derivatives of the potential. We assume that the potential is continuous at $\phi = \phi_e$. If we further require the derivative of the potential to be continuous as well, then $\epsilon_V = 0$ and the transition becomes smooth. However, if $\epsilon_V \neq 0$, then the derivative of the potential is not continuous and there is a kink in the potential. Depending on the value of $\frac{\epsilon_V}{\eta_V}$ the transition can be either mild or sharp. As we are interested mostly in a sharp transition, below we consider $\eta_V = 0$. However, this is not a restrictive assumption and most of our analysis will be carried out to the case where $\eta_V \neq 0$.

The background field equation in the final SR phase is given by [46] (see also [47])

$$\frac{d^2\phi}{dN^2} + 3\frac{d\phi}{dN} + 3M_P\sqrt{2\epsilon_V} \simeq 0, \qquad 3M_P^2 H^2 \simeq V(\phi_e). \tag{6}$$

Without loss of generality, assume the time of the transition to the final SR phase to be at $N = 0$. Imposing the continuity of ϕ and $\frac{d\phi}{dN}$ at $N = 0$, we obtain

$$M_P^{-1}\phi(N) = \frac{C_1}{3}e^{-3N} + \frac{h}{6}\sqrt{2\epsilon_V}N + C_2, \tag{7}$$

where the constants of integration C_1 and C_2 are given by

$$C_1 = \sqrt{2\epsilon_e}\left(1+\frac{h}{6}\right), \qquad C_2 = M_P^{-1}\phi_e - \frac{\sqrt{2\epsilon_e}}{3}\left(1+\frac{h}{6}\right). \tag{8}$$

Following [46] we have defined the parameter h as

$$h \equiv \frac{6\sqrt{2\epsilon_V}}{\dot\phi(t_e)}M_P = -6\sqrt{\frac{\epsilon_V}{\epsilon_e}}. \tag{9}$$

Since we assume that ϕ is decreasing monotonically during inflation, $\dot\phi < 0$, so $h < 0$. As emphasized in [9,46], h is the key parameter of the setup, controlling the sharpness of the transition from the USR phase to the final attractor phase.

The slow-roll parameters, as defined in Eq. (3), in the final SR phase ($N > 0$) are given by

$$\epsilon(\tau) = \epsilon_e\left(\frac{h}{6} - \left(1+\frac{h}{6}\right)\left(\frac{\tau}{\tau_e}\right)^3\right)^2 \tag{10}$$

and

$$\eta(\tau) = -\frac{6(6+h)}{(6+h) - h(\frac{\tau_e}{\tau})^3}. \tag{11}$$

Toward the final stage of inflation, $\tau \to \tau_0 \to 0$, we see that $\epsilon \to \epsilon_e(\frac{h}{6})^2$, while η vanishes like τ^3. While ϵ is smooth at the transition point, it is important to note that η has a discontinuity at $\tau = \tau_e$. Just prior to the transition (i.e., during the USR phase) $\eta = -6$, while right after the transition $\eta = -6 - h$. As a result, near the transition point we can approximate η as follows [46]:

$$\eta = -6 - h\theta(\tau - \tau_e), \qquad \tau_e^- < \tau < \tau_e^+. \tag{12}$$

Correspondingly, the above approximation yields

$$\frac{d\eta}{d\tau} = -h\delta(\tau - \tau_e), \qquad \tau_e^- < \tau < \tau_e^+. \tag{13}$$

For an infinitely sharp transition $h \to -\infty$. In this case, ϵ after the transition evolves rapidly to a larger value, so at the end of inflation the final value of ϵ is given by $\epsilon(\tau_0) \simeq \epsilon_V = \epsilon_e(\frac{h}{6})^2$. For an "instant" sharp transition studied in [1,2], $h = -6$. In this case ϵ in the final SR phase is frozen to its value at the end of USR, ϵ_e.

We comment that in order to perform the analysis, as in [1,2], we have considered the idealized case where the transitions are instantaneous with delta functions localized at $\tau = \tau_s$ and $\tau = \tau_e$. However, in a realistic case one should consider a smooth transition. This in turn makes the theoretical analysis intractable, and a full numerical analysis is required as in [13].

III. ORIGINS OF LOOP CORRECTIONS

The fact that small scales can induce large loop corrections on long CMB scales is somewhat counterintuitive. Intuitively speaking, based on the concept of decoupling of scales, one expects the effects of small scales to be negligible and under perturbative control. Therefore, it is important to ask what the physical origins of the large loop corrections on long CMB scales are. Here we try to answer this question.

A. Nonlinear long and short mode coupling

There are two physical effects as the origins of the loop corrections on large scales. The first effect is that there are nonlinear couplings between the long and short modes which are inherited from the nonlinearity of general relativity. These nonlinear couplings induce source terms for the evolution of the long mode. Second, the long mode which leaves the horizon in the early stage of inflation rescales the background expansion so that it modulates the power spectrum of the short modes. The combination of these two effects induce a backreaction on the long mode itself which is the origin of the loop corrections. This method was nicely employed in [3] to calculate the loop corrections, which we also follow in this subsection with some modifications.

The cubic action for the curvature perturbation ζ is given by [2,29]

$$S = M_P^2\int d\tau d^3x a^2\epsilon\left(\zeta'^2 - (\partial_i\zeta)^2 + \frac{\eta'}{2}\zeta'\zeta^2\right), \tag{14}$$

where here and below a prime indicates the derivative with respect to the conformal time. Technically speaking, the above action is for ζ_n as defined in [29], which is nonlinearly related to ζ via $\zeta = \zeta_n + \mathcal{O}(\zeta_n^2)$. The variable ζ_n is employed to eliminate the boundary terms [48] with the expense of inducing quartic order Hamiltonians, which should be taken care of. Since we look at the cubic interaction at this stage to understand the nature of loop corrections, the difference is not important, and we keep using ζ instead of ζ_n in this section.

The evolution of the Fourier space mode function $\zeta_{\mathbf{p}}(\tau)$ to second order in perturbation theory from the above action is given by

$$\zeta_{\mathbf{p}}'' + \frac{(a^2\epsilon)'}{a^2\epsilon}\zeta_{\mathbf{p}}' + \frac{(a^2\epsilon\eta')'}{4a^2\epsilon}\int\frac{d^3\mathbf{q}}{(2\pi)^3}\zeta_{\mathbf{q}}\zeta_{\mathbf{p}-\mathbf{q}} = 0. \tag{15}$$

From the above equation we see a nonlinear source term for the evolution of the long mode $\zeta_{\mathbf{p}}$ from the small-scale modes, which plays a crucial role in our discussions below.

To handle the analysis analytically, we consider a setup with an instant transition from the USR phase to the final attractor phase at $\tau = \tau_e$. In addition, the transition to the final attractor phase is sharp with $|h| \gg 1$. In this limit, there is a delta source in η' as given in Eq. (13), so Eq. (15) can be solved easily, yielding [2]

$$\zeta_{L,\mathbf{p}}(\tau_0) = \zeta^{(0)}_{L,\mathbf{p}} + c \int \frac{d^3\mathbf{q}}{(2\pi)^3} \left[\zeta^S_{\mathbf{q}}(\tau_e)\zeta^S_{\mathbf{p}-\mathbf{q}}(\tau_e) - \frac{2}{3q_e} \zeta'^S_{\mathbf{q}}(\tau_e)\zeta^S_{\mathbf{p}-\mathbf{q}}(\tau_e) \right], \tag{16}$$

in which ζ_L and ζ^S represent the long and short modes, $\zeta^{(0)}_L$ represents the linear solution of Eq. (15) in the absence of mode couplings, and q_e represents the scale which leaves the horizon at the end of USR, $q_e = -1/\tau_e$. The parameter c is a constant which depends on the details of the transition to the final attractor phase, which from Eq. (13) is given by [2] $c = -\Delta\eta/4 = -h/4$. There can be other terms involving either $\zeta_{\mathbf{q}}$ or $\zeta'_{\mathbf{q}}$ at higher orders in addition to the quadratic source terms given in Eq. (16). Our analysis can be extended to these general higher order sources as well without limitations. Finally, we comment that in the analysis of [3] the last term in Eq. (16) is further simplified, noting that $\zeta' = -(3/\tau)\zeta$, which is valid for the modes which become superhorizon during the USR phase. Here we keep ζ' since we consider subhorizon modes as well.

It is important to note that the left-hand side of Eq. (16) is calculated at the end of inflation $\tau = \tau_0$, while the source terms are calculated at the end of the USR phase $\tau = \tau_e$. This is because of our technical assumption that η' has a delta source at $\tau = \tau_e$, as given in Eq. (13). However, if the transition is not instantaneous and the evolution of $\eta(\tau)$ is continuous, then there are additional time integrals when solving for $\zeta_{L,\mathbf{p}}(\tau)$, so Eq. (16) will have a more complicated form [2]. Finally, note that there will be additional source terms at the start of USR phase $\tau = \tau_s$ in Eq. (16), but since the mode function grows only toward the end of the USR phase, we can safely ignore the contribution of the source term at $\tau = \tau_s$.

We are interested in two-point functions of the long mode $\zeta_{L,\mathbf{p}}$, with $\mathbf{p} \to 0$ representing the CMB scale modes. The short modes are denoted by the momentum $\mathbf{q}$ which run inside the loop integral with the hierarchy $p \ll q$. The power spectrum of the long mode is given by

$$\langle \zeta_{L,\mathbf{p}_1}\zeta_{L,\mathbf{p}_2}\rangle = \langle \zeta^{(0)}_{L,\mathbf{p}_1}\zeta^{(0)}_{L,\mathbf{p}_2}\rangle + 2c \int \frac{d^3\mathbf{q}}{(2\pi)^3} \left[\langle \zeta^{(0)}_{L,\mathbf{p}_1}\zeta^S_{\mathbf{q}}(\tau_e)\zeta^S_{\mathbf{p}_2-\mathbf{q}}(\tau_e)\rangle - \frac{2}{3q_e} \langle \zeta^{(0)}_{L,\mathbf{p}_1}\zeta'^S_{\mathbf{q}}(\tau_e)\zeta^S_{\mathbf{p}_2-\mathbf{q}}(\tau_e)\rangle \right] + \mathcal{O}(c^2). \tag{17}$$

It is understood that ζ^S are calculated at $\tau = \tau_e$, while ζ_L are calculated at $\tau = \tau_0$. The long mode leaves the horizon during the early stage of inflation, long before the USR phase, so ζ_L is nearly constant. More specifically, the decaying mode actually grows during the USR phase. However, it was suppressed for a long time before the start of USR phase, so its enhancement during the short USR phase is not significant enough to compete with the constant mode. As a result, we can take ζ_L to be constant and simply set $\zeta_L(\tau_0) \simeq \zeta_L(\tau_e)$ so that all modes in Eq. (17) are calculated at $\tau = \tau_e$. As we will see, it is important to realize that there is no time integral in the two-point function (17) while the integration is purely over the momentum space.

Finally, with a bit of calculation one can check that the terms in Eq. (17) containing c^2 are subleading. A representative contribution of these terms is given by

$$c^2 \int \frac{d^3\mathbf{q}_1}{(2\pi)^3} \int \frac{d^3\mathbf{q}_2}{(2\pi)^3} \langle \zeta^S_{\mathbf{q}_1}\zeta^S_{\mathbf{p}_1-\mathbf{q}_1}\zeta^S_{\mathbf{q}_2}\zeta^S_{\mathbf{p}_2-\mathbf{q}_2}\rangle, \tag{18}$$

but as they do not contain the extra factor $1/p^3$ we discard these terms.

Now our job is to calculate the three-point correlations $\langle \zeta^{(0)}_{L,\mathbf{p}_1}\zeta^S_{\mathbf{q}}\zeta^S_{\mathbf{p}_2-\mathbf{q}}\rangle$ and $\langle \zeta^{(0)}_{L,\mathbf{p}_1}\zeta'^S_{\mathbf{q}}\zeta^S_{\mathbf{p}_2-\mathbf{q}}\rangle$ between one long mode and two short modes. To calculate this, note that the effects of the long mode is only to rescale the background [29,33]. More specifically, going to comoving gauge, the metric is given by

$$ds^2 = -dt^2 + a(t)^2 e^{2\zeta_L} d\mathbf{x}^2. \tag{19}$$

As we discussed above, the long mode leaves the horizon long before the USR phase, so ζ_L is nearly constant. As a result it can be absorbed into the spacelike coordinate via $x_i \to e^{\zeta_L} x_i$ so that in momentum space $q \to e^{-\zeta_L} q$. Consequently, the effects of the long mode can be viewed as a modulation for the short modes. More specifically, for the three-point correlation $\langle \zeta^{(0)}_L \zeta^S \zeta^S\rangle$ we can write

$$\langle \zeta^{(0)}_L \zeta^S\zeta^S\rangle \simeq \langle \zeta^{(0)}_L \langle \zeta^S\zeta^S\rangle_{\zeta_L}\rangle \simeq \langle \zeta^{(0)}_L\rangle \langle \zeta^S_L\zeta^S_L\rangle + \langle \zeta^{(0)}_L\zeta^{(0)}_L\rangle \frac{\partial}{\partial\zeta_L}\langle \zeta^S\zeta^S\rangle. \tag{20}$$

As $\zeta^{(0)}_L$ is statistically incoherent, $\langle \zeta^{(0)}_L\rangle = 0$, and correspondingly the three-point function can be given in terms of the power spectrum P_ζ as follows:

$$\langle \zeta^{(0)}_L \zeta^S\zeta^S\rangle \simeq P^{(0)}_{\zeta_L} \frac{\partial P_{\zeta_S}}{\partial \zeta_L}. \tag{21}$$

To calculate the other correlations $\langle \zeta^{(0)}_{L,\mathbf{p}_1}\zeta'^S_{\mathbf{q}}\zeta^S_{\mathbf{p}_2-\mathbf{q}}\rangle$, we first symmetrize the noncommutative quantum operators ζ and ζ' so $2\zeta'\zeta \to \zeta'\zeta + \zeta\zeta'$, yielding

$$2\langle\zeta_L^{(0)}\zeta'^S\zeta^S\rangle = \langle\zeta_L^{(0)}\zeta'^S\zeta^S\rangle + \langle\zeta_L^{(0)}\zeta^S\zeta'^S\rangle. \qquad (22)$$

Following the same logic as above, we obtain

$$2\langle\zeta_L^{(0)}\zeta'^S\zeta^S\rangle = P_{\zeta_L}^{(0)}\frac{\partial}{\partial\zeta_L}\frac{dP_{\zeta_S}}{d\tau}. \qquad (23)$$

Plugging the relations (21) and (23) in our starting Eq. (17) with the understanding that $\mathbf{p}_1 = -\mathbf{p}_2 \to 0$, we obtain

$$P_{\zeta_L}(\mathbf{p}) = P_{\zeta_L}^{(0)}(\mathbf{p})\left[1 + 2c\int\frac{d^3\mathbf{q}}{(2\pi)^3}\left(\frac{\partial P_{\zeta_S}(\tau_e)}{\partial\zeta_L} - \frac{1}{3q_e}\frac{\partial}{\partial\zeta_L}\frac{dP_{\zeta_S}(\tau_e)}{d\tau}\right)\right]. \qquad (24)$$

As we discussed before, the role of the long mode is to rescale the background quantity such that

$$\frac{\partial P_{\zeta_S}}{\partial\zeta_L} = -\frac{\partial P_{\zeta_S}}{\partial\ln q} = (1-n_\zeta)P_{\zeta_S}, \qquad (25)$$

in which n_ζ represents the scale dependence of the short modes.

Plugging Eq. (25) into Eq. (24) and defining the dimensionless power spectrum $\mathcal{P}_\zeta$ related to P_ζ as

$$\mathcal{P}_\zeta(q) \equiv \frac{q^3}{2\pi^2}P_\zeta(q), \qquad (26)$$

we obtain

$$\mathcal{P}_{\zeta_L}(\mathbf{p}) = \mathcal{P}_{\zeta_L}^{(0)}(\mathbf{p})\left[1 - 2c\int d\ln q\left(\frac{\partial\mathcal{P}_{\zeta_S}}{\partial\ln q} - \frac{1}{3q_e}\frac{\partial}{\partial\ln q}\frac{d\mathcal{P}_{\zeta_S}}{d\tau}\right)\bigg|_{\tau=\tau_e}\right]. \qquad (27)$$

The integral above is in the form of a total derivative, yielding the following fractional loop correction in the long mode power spectrum:

$$\frac{\Delta\mathcal{P}_{\zeta_L}}{\mathcal{P}_{\zeta_L}} = -2c\int\left(d\mathcal{P}_{\zeta_S} - \frac{1}{3q_e}d\mathcal{P}'_{\zeta_S}\right)\bigg|_{\tau=\tau_e}. \qquad (28)$$

Defining the "modified" power spectrum $\bar{\mathcal{P}}$ via

$$\bar{\mathcal{P}}_{\zeta_S}(q,\tau_e) \equiv \mathcal{P}_{\zeta_S}(q,\tau_e) - \frac{1}{3q_e}\mathcal{P}'_{\zeta_S}(q,\tau_e), \qquad (29)$$

the loop correction in the long mode power spectrum is given by

$$\frac{\Delta\mathcal{P}_{\zeta_L}}{\mathcal{P}_{\zeta_L}} = -2c\left[\bar{\mathcal{P}}_{\zeta_S}(q_{\max},\tau_e) - \bar{\mathcal{P}}_{\zeta_S}(q_{\min},\tau_e)\right], \qquad (30)$$

in which $q_{\max}$ and $q_{\min}$ represent the higher UV and the lower IR regimes of the integration over the short modes. A similar result was originally obtained by [3], who simplified the second term in Eq. (16) via $\zeta' = -(3/\tau)\zeta$, which is valid for the modes which become superhorizon during the USR phase. Here, since we need to consider the subhorizon modes as well, we keep the contribution of $\mathcal{P}'_{\zeta_S}(q,\tau_e)$ in $\bar{\mathcal{P}}_{\zeta_S}(q,\tau_e)$ in its general form. Having said this, we note that for practical purposes $\bar{\mathcal{P}}_{\zeta_S}(q,\tau_e) \sim \mathcal{P}_{\zeta_S}(q,\tau_e)$. The above result is also in line with the result obtained by [28], who assumed Maldacena's consistency condition for the original field ζ. From Eq. (27) we see that in order for the loop correction to be significant, we require a strong scale dependence for the short modes. In other words, only the small scales which show scale dependence will contribute to the integral in Eq. (27).

In estimating the loop corrections, as in [1,2], a good prescription is to consider the modes which become superhorizon during the USR phase corresponding to $q_{\max} = q_e = -1/\tau_e$ and $q_{\min} = q_s = -1/\tau_s$. If we use this prescription, and noting that the power increases exponentially during the USR phase such that $\bar{\mathcal{P}}_\zeta(q_e,\tau_e) \sim \mathcal{P}(q_e,\tau_e) \sim e^{6\Delta N}\mathcal{P}_\zeta(q_s)$, then one can safely ignore the contribution from the lower bound of the integral ($q_{\min}$) and

$$\frac{\Delta\mathcal{P}_{\zeta_L}}{\mathcal{P}_{\zeta_L}} \sim -2c\mathcal{P}_\zeta(q_e,\tau_e) \sim e^{6\Delta N}\mathcal{P}_{\mathrm{CMB}}, \qquad (31)$$

in which $\mathcal{P}_{\mathrm{CMB}} \sim 2\times10^{-9}$ is the power spectrum on the CMB scales. The above result is qualitatively in agreement with the results of [1,2], highlighting the dangerous one-loop correction if one considers a large enough value of ΔN, i.e., a long enough period of USR inflation, so the factor $e^{6\Delta N}\mathcal{P}_{\mathrm{CMB}}$ can become order unity.

The authors of [28] argued that the contribution of the UV part in Eq. (28) is negligible after implementing the usual $i\varepsilon$ prescription $\tau \to (1+i\varepsilon\tau)$ such that $e^{-iq\tau} \to e^{-iq\tau+\varepsilon q\tau}$, so the UV contribution becomes negligible. However, by looking at our derivation of the loop correction in Eq. (30), this prescription is unjustified, as we have no integration over τ. More specifically, the $i\varepsilon$ prescription $\tau \to (1+i\varepsilon\tau)$ is usually performed to kill the rapid oscillations in the UV region when one is dealing with an integral over τ. However, in our analysis, there is no integration over τ. This is because the source term in Eq. (15) receives a delta source at $\tau = \tau_e$, so all the mode functions are calculated at a fixed time $\tau = \tau_e$. The rapid oscillations occur only in q space since modes which are deep inside the horizon at the time τ_e still experience the Minkowski background, and naturally they oscillate rapidly. As we demonstrate shortly, there will be a quadratic divergence in the momentum space which should be regularized as in a standard quantum field theory (QFT) analysis. In essence this is similar to regularizing the quartic divergence associated with short modes when dealing with the vacuum zero point energy and the cosmological constant problem.

Motivated by discussions in [28], now suppose that we do not follow the prescription of [1,2] and take $q_{\max}$ to the extreme UV value allowed. In the dimensional regularization approach, $q_{\max}$ can go to infinity. However, in a simple regularization employing the UV momentum cutoff approach, the largest allowed value of $q_{\max}$ is q_f, the mode which leaves the horizon just at the end of inflation, $\tau = \tau_0 \to 0$. Then, the question is, what is the power spectrum for that scale at $\tau = \tau_e$, i.e., $\bar{\mathcal{P}}_\zeta(q_f, \tau_e)$? To answer this question, we have to calculate the mode function $\zeta_q(\tau)$ for the small-scale modes, i.e., modes which are subhorizon during the USR phase but become superhorizon after the USR phase.

B. Mode function after the USR phase

To obtain the outgoing mode function, we have to impose the matching conditions at the start and at the end of the USR phase for an arbitrary mode q.

Starting with the Bunch-Davies initial condition during the first phase of inflation, the mode function in the USR phase is given by [9]

$$\zeta_q^{(2)} = \frac{H}{M_P\sqrt{4\epsilon_i q^3}}\left(\frac{\tau_s}{\tau}\right)^3\left[\alpha_q^{(2)}(1+iq\tau)e^{-iq\tau} + \beta_q^{(2)}(1-iq\tau)e^{iq\tau}\right], \tag{32}$$

with the coefficients $\alpha_q^{(2)}$ and $\beta_q^{(2)}$ given by

$$\alpha_q^{(2)} = 1 + \frac{3i}{2q^3\tau_s^3}(1+q^2\tau_s^2),$$
$$\beta_q^{(2)} = -\frac{3i}{2q^3\tau_s^3}(1+iq\tau_s)^2 e^{-2iq\tau_s}. \tag{33}$$

The mode function after the USR phase is given by

$$\zeta_q^{(3)} = \frac{H}{M_P\sqrt{4\epsilon(\tau)q^3}}\left[\alpha_q^{(3)}(1+iq\tau)e^{-iq\tau} + \beta_q^{(3)}(1-ik\tau)e^{iq\tau}\right], \tag{34}$$

in which $\epsilon(\tau)$ is given by Eq. (10), while $\alpha_q^{(3)}$ and $\beta_q^{(3)}$ are given by

$$\alpha_q^{(3)} = \frac{1}{8q^6\tau_s^3\tau_e^3}\left[3h(1-iq\tau_e)^2(1+iq\tau_s)^2 e^{2iq(\tau_e-\tau_s)} - i(2q^3\tau_s^3 + 3iq^2\tau_s^2 + 3i)(4iq^3\tau_e^3 - hq^2\tau_e^2 - h)\right]$$

and

$$\beta_q^{(3)} = \frac{-1}{8q^6\tau_s^3\tau_e^3}\left[3(1+iq\tau_s)^2(h + hq^2\tau_e^2 + 4iq^3\tau_e^3)e^{-2iq\tau_s} + ih(1+iq\tau_e)^2(3i + 3iq^2\tau_s^2 + 2q^3\tau_s^3)e^{-2iq\tau_e}\right].$$

To calculate the loop corrections in the long mode power spectrum in Eq. (30), we only need $\zeta_q^{(2)}(\tau_e)$ to calculate $\bar{\mathcal{P}}_\zeta(q,\tau_e)$. However, for later purposes, we also calculate the outgoing power spectrum at the end of inflation $\tau = \tau_0 \to 0$, which is given by

$$\mathcal{P}_\zeta(q,\tau_0) = \frac{H^2}{8M_P^2\pi^2\epsilon_V}\left|\alpha_q^{(3)} + \beta_q^{(3)}\right|^2, \tag{35}$$

in which ϵ_V is the value of the slow-roll parameter at the end of inflation.

With the mode function $\zeta_q^{(2)}(\tau_e)$ at hand, we can calculate the one-loop corrections from Eq. (30). A schematic plot of $\bar{\mathcal{P}}(q,\tau_e)$ is presented in Fig. 1. For the modes which leave the horizon during the first stage of inflation $q \ll q_s$, the power spectrum has a plateau given by the COBE normalization $\mathcal{P}_{\rm CMB}$. There is a dip prior to the USR phase and a sharp rise in the power spectrum in the intermediate USR phase, followed by a peak with oscillations superimposed. All of these properties are well understood; see, for example, [49–53]. In particular, the power spectrum grows as $\mathcal{P}_\zeta \propto q^4$ just prior to the peak.

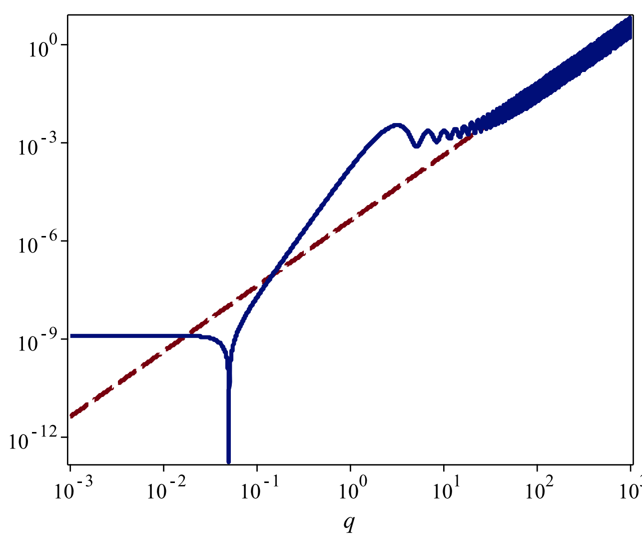

FIG. 1. The (log-log) plot of the modified power spectrum $\bar{\mathcal{P}}_\zeta(q,\tau_e)$ calculated at τ_e for $h = -6$ and $\Delta N = \ln(10) \simeq 2.3$. The USR phase starts at $q = 1$, while the small-scale modes which leave the horizon long after the USR phase correspond to $q \gg 1$. The red dashed line represents the overall factor in Eq. (36) indicating the q^2 divergence for the UV modes, while the rapid oscillations superimposed on top of this scaling can be seen as well. The position of the dip is prior to the USR phase, and the rapid rise of the power spectrum $\mathcal{P}_\zeta \propto q^4$ prior to the peak is the hallmark of the USR setup.

This is essential for the loop corrections given in Eq. (25). On the other hand, the modes with $q > q_e$ are subhorizon during the USR phase and the power spectrum grows as q^2 with rapid oscillations superimposed on top of it. More specifically, for $q \gg q_e$, we have

$$|\bar{\mathcal{P}}_\zeta(q,\tau_e)| \simeq \frac{1}{3}\left(\frac{q}{q_e}\right)^2 e^{6\Delta N}\mathcal{P}_{\rm CMB}\left[1 - 6\cos\left(\frac{2q}{q_s}\right)e^{-3\Delta N}\right], \tag{36}$$

so we see a quadratic divergence in loop corrections for $q \to \infty$, while the rapid oscillations have subleading amplitudes. These behaviors can be seen in Fig. 1 as well.

The quadratic divergence of the power spectrum for the UV scale is expected, which is the hallmark of the QFT corrections. To find a finite physical result, we have to renormalize the divergent loop corrections order by order. However, in order for the renormalization procedure to work at each order, we have to make sure that the starting one-loop corrections are under control. In order for the one-loop corrections to be small, we require $e^{6\Delta N}\mathcal{P}_{\rm CMB} \ll 1$ since this quantity controls the common amplitude of the loop corrections, as can be seen in both Eqs. (31) and (36). To perform the renormalization, one may set $q_{\rm max} = q_f$, the modes which leave the horizon at the end of inflation. In this way, one counts the contribution of all modes which become superhorizon by the end of inflation.

It is important to note that the final renormalized loop corrections are not necessarily zero. On the other hand, it was argued in [28] that the leading loop corrections vanish after one kills the rapid oscillations using an $i\epsilon$ prescription on τ. As we argued previously, this is unjustified. First, we have no integration over the time coordinate, as the mode functions in Eq. (30) are calculated at a fixed time $\tau = \tau_e$. Second, the rapid oscillations in q are subleading compared to the dominant quadratic divergence, so the renormalized power spectrum is not zero. Finally, one has to employ standard QFT methods, such as the dimensional regularization scheme, to regularize and renormalize the quadratic divergence. For earlier works concerning the loop corrections and renormalizations in slow-roll setups, see [54–56].

While the power spectrum at $\tau = \tau_e$ has the behavior shown in Fig. 1, it is also instructive to look at the final power spectrum $\mathcal{P}_\zeta(q,\tau_0)$ measured at the time of end of inflation, $\tau = \tau_0$. A schematic view of the power spectrum is presented in Fig. 2. We see that for modes which leave the horizon by the end of the USR phase with $q \lesssim q_e$ the power spectrum is similar to that in Fig. 1. However, for the modes which become superhorizon after the USR phase the power spectrum shows a significant difference in which it reaches a plateau instead of growing quadratically. When we define $x \equiv -q\tau_s$, the final power spectrum for $x \gg 1$ is given by

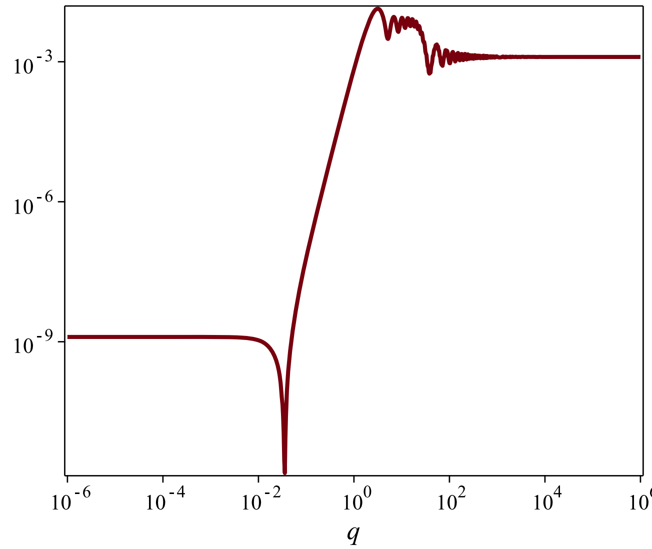

FIG. 2. Power spectrum $\mathcal{P}_\zeta(q,\tau_0)$ measured at the time of end of inflation $\tau = \tau_0$ for $h = -6$ and $\Delta N = \ln(10) \simeq 2.3$ with a USR phase starting at $q = 1$. For modes leaving the horizon by the end of a USR phase with $q \lesssim 1$, the behavior is the same as in Fig. 1, while for small scales with $q \gg 1$ there is a significant difference when the power spectrum reaches an asymptotic value given by Eq. (37).

$$\mathcal{P}_\zeta(q,\tau_0) \simeq e^{6\Delta N}\mathcal{P}_{\rm CMB}\left(\frac{h-6}{h}\right)^2\left(1 + 3\frac{\sin(2x)}{x}\right) \simeq \frac{H^2}{8\pi^2\epsilon_V M_P^2}\left(\frac{h-6}{6}\right)^2. \tag{37}$$

The power spectrum reaches a plateau given by Eq. (37). In addition, the sharper the transition, the larger the final value of the power spectrum [46].

In conclusion, we have reproduced the results in [1,2], indicating that the loop correction is a genuine phenomena. The loop corrections are under perturbative control if $e^{6\Delta N}\mathcal{P}_{\rm CMB} \ll 1$.

After performing the technical analysis, here we summarize the physical reasons behind the nontrivial loop correction. There are two important effects which should be taken into account: first, the nonlinear coupling between the long and short modes which provide the source term for the evolution of the long mode. Second, the long mode provides a modulation to the spectrum of the short mode. This modulation becomes significant if the power spectrum of the short mode experiences a significant scale-dependent enhancement. In our case at hand, this corresponds to a maximum scale-dependent $\mathcal{P}_\zeta \propto q^4$ just prior to the peak of the power spectrum. Finally, the combination of the nonlinear coupling between the long and short modes and the modulation of the short modes by the long mode backreacts on the long mode itself and induces the one-loop correction. This picture was first put forward in [3]; see also [13].

IV. ONE-LOOP CORRECTION FROM A CUBIC HAMILTONIAN

In this section we revisit the analysis of [27], who calculated the one-loop corrections from the cubic interaction Hamiltonian and concluded that the loop corrections cancel out.

A. In-in analysis

To calculate the loop corrections, we employ the standard in-in formalism [54] in which the expectation value of the operator $\hat{O}$ at the end of inflation τ_0 is given by the following perturbative series:

$$\langle \hat{O}(\tau_0)\rangle = \left\langle \left[\bar{\mathrm{T}}\exp\left(i\int_{-\infty}^{\tau_0} d\tau' H_{\rm in}(\tau')\right)\right] \times \hat{O}(\tau_0)\left[\mathrm{T}\exp\left(-i\int_{-\infty}^{\tau_0} d\tau' H_{\rm in}(\tau')\right)\right]\right\rangle, \quad (38)$$

in which T and $\bar{\mathrm{T}}$ represent the time ordering and antitime ordering respectively, while $H_{\rm in}(\tau)$ represents the interaction Hamiltonian, which in our case is $H_{\rm in}(\tau) = \mathbf{H}_3$.

The one-loop correction from the cubic self-interaction is calculated in [27] in two different methods. The first method incorporates the boundary terms directly into the cubic Hamiltonian, yielding

$$\mathbf{H}_3 = H_a + H_b, \quad (39)$$

in which H_a is a bulk term given by

$$H_a = -M_P^2 \int d^3x \left(\frac{a^2\epsilon}{2}\eta'\zeta^2\zeta'\right), \quad (40)$$

while H_b is a boundary term [48],

$$H_b = M_P^2 \int d^3x \frac{d}{d\tau}\left(\frac{a^2\epsilon}{2}\eta\zeta^2\zeta'\right). \quad (41)$$

Note that here, following [27], we work with ζ itself, while in previous sections we were working with ζ_n. The effects of the boundary term H_b were not considered in [1,2]. It was argued in [27] that once its contributions are added along with the bulk term H_a, the one-loop corrections in the power spectrum cancel each other to order $p^3/q^3 \ll 1$.

In the second method used in [27], the term containing η' in H_a is traded via a boundary term. The new boundary term cancels exactly the boundary term H_b. After using the linear field equation, one obtains the following equivalent Hamiltonian:

$$\mathbf{H}_3 = H_c + H_d = M_P^2 \int d^3x a^2\epsilon\eta\left(\zeta'^2\zeta + \frac{1}{2}\zeta^2\partial^2\zeta\right), \quad (42)$$

in which H_c and H_d are both bulk terms, given by

$$H_c = M_P^2 \int d^3x a^2\epsilon\eta\zeta'^2\zeta \quad (43)$$

and

$$H_d = \frac{M_P^2}{2}\int d^3x a^2\epsilon\eta\zeta^2\partial^2\zeta. \quad (44)$$

Interestingly, we see that the Hamiltonian (42) is exactly the same as the cubic Hamiltonian obtained in [9,38] in which $\zeta = -H\pi + \mathcal{O}(\pi^2)$, where π represents the Goldstone boson associated with the fluctuations of the inflaton field. On the other hand, no cancellation of the one-loop corrections was observed in [9] at the cubic order (even in the presence of quartic interactions). This indicates that something is going wrong in the analysis of either [9] or [27]. Here, we pay careful attention to find the source of disagreement between [9,27] and whether or not the one-loop corrections cancel out, as claimed in [27].

To perform the in-in analysis, the following relations for $\tau_s \le \tau \le \tau_e$ have been used in [27]:

$$[\zeta_{\mathbf{q}}(\tau), \zeta'_{\mathbf{p}}(\tau_0)] = (2\pi)^3\delta^3(\mathbf{q}+\mathbf{p})\frac{i}{2a^2M_P^2\epsilon(\tau)} \quad (45)$$

and

$$[\zeta_{\mathbf{q}}(\tau), \zeta_{\mathbf{p}}(\tau_0)] \simeq 0. \quad (46)$$

A careful investigation shows that Eq. (45) is correct but Eq. (46) is incorrect. Indeed, it was argued in [27] that since for the long mode $\zeta_{\mathbf{p}}$ is nearly conserved on superhorizon scales, $\zeta_{\mathbf{p}}(\tau_0) \simeq \zeta_{\mathbf{p}}(\tau)$ for $\tau_s \le \tau \le \tau_e$, and since the equal time commutator of the field vanishes, one obtains Eq. (46). However, there is a subtle flaw in this argument in which the role of the decaying mode is neglected during the USR phase. More specifically, the decaying mode will grow during the USR phase, so the approximation $\zeta_{\mathbf{p}}(\tau_0) \simeq \zeta_{\mathbf{p}}(\tau)$ may not be consistent when one is dealing with a nested integral. As shown in [9] [see Eq. (A.23) in [9]], for $\tau_s \le \tau \le \tau_e$ one instead has

$$[\zeta_{\mathbf{q}}(\tau), \zeta_{\mathbf{p}}(\tau_0)] \simeq (2\pi)^3\delta^3(\mathbf{q}+\mathbf{p})\frac{i\tau}{6a^2M_P^2\epsilon(\tau)}\left(1 + \frac{6-h}{h}\frac{\tau^3}{\tau_e^3}\right). \quad (47)$$

When Eqs. (45) and (47) are compared, it is not guaranteed that inside the nested integral one can automatically neglect Eq. (47) while keeping Eq. (45).

In the following, we repeat the analysis of [27] using the Hamiltonian given in Eq. (42). We also compare the results with those in [9], which were obtained via a different in-in method.

The analysis in [27] is based on the commutator approach [54], in which

$$\langle\zeta_{\mathbf{p}}(\tau_0)\zeta_{\mathbf{p}'}(\tau_0)\rangle = -\int^{\tau_0} d\tau_1 \int^{\tau_1} d\tau_2 \langle[\mathbf{H_3}(\tau_2),[\mathbf{H_3}(\tau_1),\zeta_{\mathbf{p}}(\tau_0)\zeta_{\mathbf{p}'}(\tau_0)]]\rangle. \tag{48}$$

Depending on where H_c and H_d are located in the nested integrals, we obtain

$$\langle\zeta_{\mathbf{p}}\zeta_{\mathbf{p}'}\rangle = \langle\zeta_{\mathbf{p}}\zeta_{\mathbf{p}'}\rangle_{[c,c]} + \langle\zeta_{\mathbf{p}}\zeta_{\mathbf{p}'}\rangle_{[d,c]} + \langle\zeta_{\mathbf{p}}\zeta_{\mathbf{p}'}\rangle_{[c,d]}, \tag{49}$$

in which, for example,

$$\langle\zeta_{\mathbf{p}}\zeta_{\mathbf{p}'}\rangle_{[c,c]} = -\int^{\tau_0} d\tau_1 \int^{\tau_1} d\tau_2 \langle[\mathbf{H_c}(\tau_2),[\mathbf{H_c}(\tau_1),\zeta_{\mathbf{p}}(\tau_0)\zeta_{\mathbf{p}'}(\tau_0)]]\rangle, \tag{50}$$

and so on.

In the analysis of [27] it was argued that the first two terms in Eq. (49) cancel each other to the volume order $p^3/q^3 \ll 1$, while the last term in Eq. (49) is subleading. Indeed, as we check specifically below, the conclusion that $\langle\zeta_{\mathbf{p}}\zeta_{\mathbf{p}'}\rangle_{[c,d]}$ is subleading is correct. However, we show that the cancellation between the first two terms in Eq. (49) is not exact, which is the source of the discrepancy between the results of [9,27].

To perform the in-in analysis, as in [9,27], we only consider the contributions for the time interval $\tau_s \le \tau \le \tau_e$. Performing all the contractions and incorporating the symmetric factors, one obtains

$$\langle\zeta_{\mathbf{p}}\zeta_{\mathbf{p}'}\rangle'_{[c,c]} = -8M_P^4 \int_{\tau_s}^{\tau_e} d\tau_1 \int_{\tau_s}^{\tau_1} d\tau_2 \times \int \frac{d^3\mathbf{q}}{(2\pi)^3} \mathrm{Im}\left[X_1^*(\tau_2)(c_1\delta(\tau_1)+2\beta(\tau_1))\right], \tag{51}$$

in which the coefficient c_1 is added for bookkeeping, as we discuss below. Here and below, $\langle\cdots\rangle'$ means that we absorbed the overall factor $(2\pi)^3\delta^3(\mathbf{q}+\mathbf{p})$. In addition,

$$X_1(\tau) \equiv \eta\epsilon a^2 \zeta_p^*(\tau_0)\zeta_p(\tau)\zeta_q'(\tau)^2, \tag{52}$$

$$\delta(\tau) \equiv 2\epsilon\eta a^2 \zeta_q'(\tau)^2 \mathrm{Im}\left[\zeta_p^*(\tau_0)\zeta_p(\tau)\right], \tag{53}$$

and

$$\beta(\tau) \equiv 2\epsilon\eta a^2 \zeta_q'(\tau)\zeta_q(\tau)\mathrm{Im}\left[\zeta_p^*(\tau_0)\zeta_p'(\tau)\right]. \tag{54}$$

Note that there are additional subleading terms containing $\zeta_p'(\tau_2)$ which are not included in Eq. (51). This is because these terms are suppressed by a factor p^2 compared to the term denoted by $X_1(\tau_2)$.[1]

Looking at the expressions of $\delta(\tau)$ and $\beta(\tau)$, we note that $\delta(\tau)$ originated from the commutator (47) while $\beta(\tau)$ originated from the commutator (45). Consequently, in the analysis of [27], who uses Eq. (46) instead of Eq. (47), the term containing $\delta(\tau)$ does not exist. This corresponds to setting $c_1 = 0$. However, in our analysis we have $c_1 = 1$. We have verified that if we set $c_1 = 0$, then Eq. (51) agrees exactly with the corresponding result in [27] [Eq. (50) in [27]].

Proceeding similarly, we obtain

$$\langle\zeta_{\mathbf{p}}\zeta_{\mathbf{p}'}\rangle'_{[d,c]} = 8M_P^4 \int_{\tau_s}^{\tau_e} d\tau_1 \int_{\tau_s}^{\tau_1} d\tau_2 \int \frac{d^3\mathbf{q}}{(2\pi)^3} q^2 \times \mathrm{Im}\left[Y^*(\tau_2)(c_1\delta(\tau_1)+2\beta(\tau_1))\right], \tag{55}$$

in which

$$Y(\tau) \equiv \eta\epsilon a^2 \zeta_p^*(\tau_0)\zeta_p(\tau)\zeta_q(\tau)^2. \tag{56}$$

As in the previous case, if we set $c_1 = 0$, the above result agrees exactly with $\langle\zeta_{\mathbf{p}}\zeta_{\mathbf{p}'}\rangle_{[d,c]}$ obtained in [27].

Finally, calculating $\langle\zeta_{\mathbf{p}}\zeta_{\mathbf{p}'}\rangle_{[c,d]}$, we obtain

$$\langle\zeta_{\mathbf{p}}\zeta_{\mathbf{p}'}\rangle'_{[c,d]} = -16M_P^4 \int_{\tau_s}^{\tau_e} d\tau_1 \int_{\tau_s}^{\tau_1} d\tau_2 \epsilon(\tau_1)a(\tau_1)^2 \times \int \frac{d^3\mathbf{q}}{(2\pi)^3} q^2 \mathrm{Im}\left[\zeta_p^*(\tau_0)\zeta_p(\tau_1)\right] \times \mathrm{Im}\left[X_1^*(\tau_2)\zeta_q(\tau_1)^2\right].$$

Combining the results for $\langle\zeta_{\mathbf{p}}\zeta_{\mathbf{p}'}\rangle_{[c,c]}$, $\langle\zeta_{\mathbf{p}}\zeta_{\mathbf{p}'}\rangle_{[d,c]}$, and $\langle\zeta_{\mathbf{p}}\zeta_{\mathbf{p}'}\rangle_{[c,d]}$, the total one-loop correction at the cubic order is obtained to be

$$\langle\zeta_{\mathbf{p}}\zeta_{\mathbf{p}'}\rangle'_{\mathbf{H_3}} = 8M_P^4 \int_{\tau_s}^{\tau_e} d\tau_1 \int_{\tau_s}^{\tau_1} d\tau_2 \int \frac{d^3\mathbf{q}}{(2\pi)^3} \mathcal{F}(\tau_1,\tau_2;q), \tag{57}$$

in which

$$\mathcal{F}(\tau_1,\tau_2;q) \equiv \mathrm{Im}\left\{X_1^*(\tau_2)\left[(c_1\delta(\tau_1)+2\beta(\tau_1))\times(1-f_q^*(\tau_2)) - c_1 f_q(\tau_1)\delta(\tau_1)\right]\right\} \tag{58}$$

and

[1]In the analysis of [9], these subleading terms are denoted by X_2; see the discussion after Eq. (A.10) in [9].

$$f_q(\tau) \equiv q^2 \frac{Y}{X_1} = \frac{q^2 \zeta_q^2}{\zeta_q'^2}. \qquad (59)$$

As mentioned before, our result in Eq. (63) reduces to the result of [27] if we set $c_1 = 0$. In addition, performing the nested integral, one can show that the last term in $\mathcal{F}$, containing $f_q(\tau_1)\delta(\tau_1)$, is subleading, which agrees with the conclusion in [27] that the contribution of $\langle \zeta_{\mathbf{p}} \zeta_{\mathbf{p}'} \rangle_{[c,d]}$ is subleading compared to $\langle \zeta_{\mathbf{p}} \zeta_{\mathbf{p}'} \rangle_{[c,c]}$ and $\langle \zeta_{\mathbf{p}} \zeta_{\mathbf{p}'} \rangle_{[d,c]}$.

Finally, Eq. (57) agrees with our earlier result in [9], which was obtained using a somewhat different method to implement the in-in analysis. More specifically, in [9] the in-in analysis is performed as follows:

$$\begin{aligned} \langle \zeta_{\mathbf{p}}(\tau_0) \zeta_{\mathbf{p}'}(\tau_0) \rangle_{\mathbf{H}_3} &= \langle \zeta_{\mathbf{p}}(\tau_0) \zeta_{\mathbf{p}'}(\tau_0) \rangle_{(2,0)} \\ &+ \langle \zeta_{\mathbf{p}}(\tau_0) \zeta_{\mathbf{p}'}(\tau_0) \rangle_{(1,1)} \\ &+ \langle \zeta_{\mathbf{p}}(\tau_0) \zeta_{\mathbf{p}'}(\tau_0) \rangle_{(0,2)}, \end{aligned} \qquad (60)$$

in which

$$\begin{aligned} &\langle \zeta_{\mathbf{p}}(\tau_0) \zeta_{\mathbf{p}'}(\tau_0) \rangle_{(2,0)} \\ &= -\int_{-\infty}^{\tau_0} d\tau_1 \int_{-\infty}^{\tau_1} d\tau_2 \langle \mathbf{H}_3(\tau_2) \mathbf{H}_3(\tau_1) \zeta_{\mathbf{p}}(\tau_0) \zeta_{\mathbf{p}'}(\tau_0) \rangle \\ &= \langle \zeta_{\mathbf{p}}(\tau_0) \zeta_{\mathbf{p}'}(\tau_0) \rangle^{\dagger}_{(0,2)} \end{aligned} \qquad (61)$$

and

$$\begin{aligned} &\langle \zeta_{\mathbf{p}}(\tau_0) \zeta_{\mathbf{p}'}(\tau_0) \rangle_{(1,1)} \\ &\quad = \int_{-\infty}^{\tau_0} d\tau_1 \int_{-\infty}^{\tau_0} d\tau_2 \langle \mathbf{H}_3(\tau_1) \zeta_{\mathbf{p}}(\tau_0) \zeta_{\mathbf{p}'}(\tau_0) \mathbf{H}_3(\tau_2) \rangle. \end{aligned} \qquad (62)$$

Combining all contributions one obtains the same result as Eq. (57); see [9] for detailed derivations.

B. Loop cancellation?

Our goal here is to examine the loop cancellation at the cubic order, as advocated for in [27]. Plugging the mode functions (32) and (34) into Eq. (57) and performing the nested integrals for the range $\tau_s \le \tau_2 \le \tau_1 \le \tau_e$ and $-\frac{1}{\tau_s} \le q \le -\frac{1}{\tau_e}$, we obtain

$$\langle \zeta_{\mathbf{p}} \zeta_{\mathbf{p}'} \rangle'_{\mathbf{H}_3} = \frac{9 c_1 (h-12)}{8h} (\Delta N e^{6\Delta N}) \frac{H^4}{2\pi^2 M_P^4 \epsilon_i^2 p^3}. \qquad (63)$$

Correspondingly, the correction in one-loop power spectrum from the cubic Hamiltonian $\Delta \mathcal{P}_{\mathbf{H}_3}$ is obtained to be

$$\Delta \mathcal{P}_{\mathbf{H}_3} \equiv \frac{p^3}{2\pi^2} \langle \zeta_{\mathbf{p}}^2 \rangle_{\mathbf{H}_3} = \frac{18 c_1 (h-12)}{h} (\Delta N e^{6\Delta N}) \mathcal{P}_{\mathrm{CMB}}^2. \qquad (64)$$

As expected, we see that if $c_1 = 0$, then the loop correction cancels to leading order, as advocated for in [27]. However, the consistent analysis requires $c_1 = 1$ and there is no loop cancellation.

The above result is qualitatively consistent with the result obtained [1]. However, with $h = -6$, which is the case studied in [1], the above result is larger than the result obtained in [1] by a factor of 6. There may be a number of reasons for this numerical discrepancy. Note that in [1] they used the new variables ζ_n defined in [29] with $\zeta = \zeta_n + \mathcal{O}(\zeta_n^2)$, while here we work with ζ. It is possible that the nonlinear relation between ζ and ζ_n induces quartic interactions from the starting cubic interactions, which were not taken into account in [1]. In addition, the boundary term was not included in the analysis of [1], which may also contribute to the numerical mismatch.

The conclusion is that there is no cancellation in the one-loop correction at the cubic order. The source of the disagreement with the conclusion of [27] is that $[\zeta_{\mathbf{p}}(\tau_0), \zeta_{\mathbf{q}}(\tau)] \neq 0$, as summarized in Eq. (47). This is because one cannot neglect the roles of the would-be decaying mode, which grows exponentially during the USR phase.

Now we can compare our result in Eq. (63) with the corresponding result obtained in [9]. We see that Eq. (63) agrees exactly with the result obtained in [9] when the integration[2] is over $\tau_s \le \tau_2 \le \tau_1 \le \tau_e$; see Eq. (5.39) in [9]. This is not surprising, since the starting cubic Hamiltonian in both [9,27] is the same, as given in Eq. (42).

Finally, we comment that the contribution of the quartic Hamiltonian in loop correction was calculated in [9], which we present here:

$$\langle \zeta_{\mathbf{p}}^2 \rangle'_{\mathbf{H}_4} = \frac{3}{8h} (h^2 + 6h + 36) (\Delta N e^{6\Delta N}) \frac{H^4}{2\pi^2 M_P^4 \epsilon_i^2 p^3}. \qquad (65)$$

We see that it has a somewhat different dependence on the sharpness parameter h such that the quartic one-loop corrections scales linearly with h for $|h| \gg 1$.

Combining the cubic and quartic one-loop corrections from Eqs. (63) and (65), and, setting $c_1 = 1$, the total one-loop correction is given by

$$\Delta \mathcal{P}_{\mathbf{H_3 + H_4}} = (6h + 54)(\Delta N e^{6\Delta N}) \mathcal{P}_{\mathrm{CMB}}^2. \qquad (66)$$

We see that the total one-loop correction scales linearly with h. There is no cancellation in the total one-loop

[2]In performing the nested integrals, two different strategies were considered in [9]. In the first strategy, one calculates the nested integral considering only the modes which become superhorizon during the USR phase. This means cutting the time integral in the range $-\frac{1}{q} \le \tau_2 \le \tau_1 \le \tau_e$ so the lower bound of the integral is $-\frac{1}{q}$ instead of τ_s. The second strategy is similar to that followed here (as in [27]), integrating over all modes, whether subhorizon or superhorizon during the USR phase, corresponding to $\tau_s \le \tau_2 \le \tau_1 \le \tau_e$. This corresponds to simply setting the lower bound of the time integral to be τ_s, as given in Eq. (57).

correction for a general value of h except at $h = -9$. Of course, there are subleading terms in one-loop contributions which were not included in our analysis here, so we believe that, even for $h = -9$, the one-loop cancellation does not occur. Finally, for $h = -6$, which is the case studied in [1,2], the total one-loop correction in Eq. (66) is larger than the result in [1,2] by a factor of 2. Having said this, it is interesting that the final result, once the effects of the boundary terms and the quartic interaction are incorporated, is qualitatively in agreement with the results of [1,2].

As argued in [1,2], the loop corrections in the form of Eq. (66) can get out of control if one enhances the short-scale power spectrum during the USR phase by a factor of 10^7 to generate the desired PBH abundance. Furthermore, as argued in [9], this gets even worse if one considers extremely sharp transitions with $h \to -\infty$. However, for a mild transition with $h \sim \eta_V$, the loop corrections will be slow-roll suppressed and the model is reliable for PBH formation [3,11].

V. SUMMARY AND DISCUSSIONS

In this work we have revisited the question of one-loop corrections in the setup which contains an intermediate phase of USR inflation. First, we have provided physical arguments on the reality of loop effects. More specifically, one may worry that large loop corrections on long modes induced from small scales may be in conflict with the notion of the decoupling of scales. We have tried to clarify this puzzle. We have argued that the nonlinear couplings between the long and short modes generate a second order source term for the evolution of the long mode perturbations. On the other hand, the long mode rescales the background coordinate, so its effects can be viewed as a modulation of the short mode power spectrum. These two effects combine to induce a nontrivial backreaction on the long mode itself, which can be viewed as the source of the loop corrections [4]. In order for the loop corrections to be noticeable, we require a significant scale dependence for the power spectrum of the short modes. This is guaranteed in the USR phase, as the power spectrum of the modes which leave the horizon during the USR phase experiences a rapid rise like $\mathcal{P}_\zeta \propto q^4$.

In the first part of this work we have found that our expression for the loop correction in Eq. (28) has the same structure as advocated for in [28]. However, we disagree with the argument in [28] that the contribution of $q_{\max}$ is negligible after performing the $i\varepsilon$ prescription. Indeed, a natural prescription for $q_{\max}$ is $q_{\max} = q_e$, as advocated for in [1,2]. This leads to the expected result in Eq. (31). However, if we follow the prescription of [28] and push $q_{\max}$ to the maximum allowed value, then the power spectrum has a quadratic divergence in the UV region, with rapid small oscillations superimposed on top of it. We have argued that these oscillations are harmless, as they are much smaller than the overall quadratic divergence. Indeed, the situation here is similar to the standard QFTs in which one has to employ a renormalization scheme, such as the dimensional regularization approach, to regularize and renormalize the divergent power spectrum. For this to be consistent, one requires the loop corrections to be perturbatively under control order by order. Consequently, we need the fractional one-loop correction in Eq. (31) with the amplitude $e^{6\Delta N}\mathcal{P}_{\rm CMB}$ to be small.

In the second part of this work we have revisited the claim in [27] that the loop contributions cancel out to leading order when using the cubic Hamiltonian. We have highlighted the important roles played by the would-be decaying mode during the USR phase. As is well known, the decaying mode grows exponentially during the USR period, which is the main reason behind the violation of Maldacena's consistency condition [34]. Correspondingly, one cannot simply take for granted that $[\zeta_{\mathbf{p}}(\tau_0), \zeta_{\mathbf{q}}(\tau)] \neq 0$, so one should use Eq. (47) instead of Eq. (46). The contribution of Eq. (47) in our analysis is captured by the term δ in Eq. (58). To follow the contribution of the term δ, we have inserted the fiducial parameter c_1 into the follow-up analysis. We have verified that if $c_1 = 0$, then one reproduces the result of [27] in which the loop corrections by cubic interactions cancel out to leading order. However, in the correct treatment with $c_1 = 1$, the loop correction does not cancel out, as seen explicitly in Eq. (63).

As the one-loop corrections are genuine and are not canceled out, one has to worry about their cosmological implications. In particular, it may not be easy to generate PBHs in the models employing an intermediate phase of USR inflation, as highlighted in [1,2]. The amplitude of loop corrections scales linearly with the sharpness parameter h. Correspondingly, for sharp transitions the loop corrections can get out of control for $\Delta N > 1$. However, as shown in [11], the loop corrections will be slow-roll suppressed if the transition is mild. Another interesting question is the loop effects on the bispectrum. Experience with the case of the power spectrum suggests that the loop corrections can have significant impacts in the f_{NL} parameter on large CMB scales as well. This is a nontrivial question since the corresponding in-in analysis involves higher order nested integrals. Another question of interest is to look at two loops and higher order loop corrections for both the power spectrum and the bispectrum. As we see in Eq. (66), at the one-loop level, the loop corrections scale linearly with h for large values of $|h|$. It is interesting to examine the dependence of the two loops and higher order loops on h. These nontrivial dependences on the sharpness parameter may put additional constraints on the model for it to be perturbatively under control. We would like to come back to these questions in the future.

ACKNOWLEDGMENTS

We thank Antonio Riotto, Mohammad Hossein Namjoo, Jacopo Fumagalli, and Sina Hooshangi for the helpful discussions and correspondence. We are grateful to Jason Kristiano for many insightful comments and discussions. This work is supported by INSF of Iran under Grant No. 4025208.

[1] J. Kristiano and J. Yokoyama, arXiv:2211.03395.
[2] J. Kristiano and J. Yokoyama, arXiv:2303.00341.
[3] A. Riotto, arXiv:2301.00599.
[4] A. Riotto, arXiv:2303.01727.
[5] S. Choudhury, M. R. Gangopadhyay, and M. Sami, arXiv: 2301.10000.
[6] S. Choudhury, S. Panda, and M. Sami, Phys. Lett. B **845**, 138123 (2023).
[7] S. Choudhury, S. Panda, and M. Sami, J. Cosmol. Astropart. Phys. 11 (2023) 066.
[8] S. Choudhury, S. Panda, and M. Sami, J. Cosmol. Astropart. Phys. 08 (2023) 078.
[9] H. Firouzjahi, J. Cosmol. Astropart. Phys. 10 (2023) 006.
[10] H. Motohashi and Y. Tada, J. Cosmol. Astropart. Phys. 08 (2023) 069.
[11] H. Firouzjahi and A. Riotto, arXiv:2304.07801.
[12] G. Tasinato, Phys. Rev. D **108**, 043526 (2023).
[13] G. Franciolini, A. J. Iovino, M. Taoso, and A. Urbano, arXiv: 2305.03491.
[14] H. Firouzjahi, Phys. Rev. D **108**, 043532 (2023).
[15] S. Maity, H. V. Ragavendra, S. K. Sethi, and L. Sriramkumar, arXiv:2307.13636.
[16] S. L. Cheng, D. S. Lee, and K. W. Ng, arXiv:2305.16810.
[17] J. Fumagalli, S. Bhattacharya, M. Peloso, S. Renaux-Petel, and L. T. Witkowski, arXiv:2307.08358.
[18] A. Nassiri-Rad and K. Asadi, arXiv:2310.11427.
[19] D. S. Meng, C. Yuan, and Q. g. Huang, Phys. Rev. D **106**, 063508 (2022).
[20] S. L. Cheng, D. S. Lee, and K. W. Ng, Phys. Lett. B **827**, 136956 (2022).
[21] P. Ivanov, P. Naselsky, and I. Novikov, Phys. Rev. D **50**, 7173 (1994).
[22] J. Garcia-Bellido and E. Ruiz Morales, Phys. Dark Universe **18**, 47 (2017).
[23] M. Biagetti, G. Franciolini, A. Kehagias, and A. Riotto, J. Cosmol. Astropart. Phys. 07 (2018) 032.
[24] M. Y. Khlopov, Res. Astron. Astrophys. **10**, 495 (2010).
[25] O. Özsoy and G. Tasinato, Universe **9**, 203 (2023).
[26] C. T. Byrnes and P. S. Cole, arXiv:2112.05716.
[27] J. Fumagalli, arXiv:2305.19263.
[28] Y. Tada, T. Terada, and J. Tokuda, arXiv:2308.04732.
[29] J. M. Maldacena, J. High Energy Phys. 05 (2003) 013.
[30] W. H. Kinney, Phys. Rev. D **72**, 023515 (2005).
[31] M. J. P. Morse and W. H. Kinney, Phys. Rev. D **97**, 123519 (2018).
[32] W. C. Lin, M. J. P. Morse, and W. H. Kinney, J. Cosmol. Astropart. Phys. 09 (2019) 063.
[33] P. Creminelli and M. Zaldarriaga, J. Cosmol. Astropart. Phys. 10 (2004) 006.
[34] M. H. Namjoo, H. Firouzjahi, and M. Sasaki, Europhys. Lett. **101**, 39001 (2013).
[35] J. Martin, H. Motohashi, and T. Suyama, Phys. Rev. D **87**, 023514 (2013).
[36] X. Chen, H. Firouzjahi, M. H. Namjoo, and M. Sasaki, Europhys. Lett. **102**, 59001 (2013).
[37] X. Chen, H. Firouzjahi, E. Komatsu, M. H. Namjoo, and M. Sasaki, J. Cosmol. Astropart. Phys. 12 (2013) 039.
[38] M. Akhshik, H. Firouzjahi, and S. Jazayeri, J. Cosmol. Astropart. Phys. 07 (2015) 048.
[39] M. Akhshik, H. Firouzjahi, and S. Jazayeri, J. Cosmol. Astropart. Phys. 12 (2015) 027.
[40] S. Mooij and G. A. Palma, J. Cosmol. Astropart. Phys. 11 (2015) 025.
[41] R. Bravo, S. Mooij, G. A. Palma, and B. Pradenas, J. Cosmol. Astropart. Phys. 05 (2018) 024.
[42] B. Finelli, G. Goon, E. Pajer, and L. Santoni, Phys. Rev. D **97**, 063531 (2018).
[43] S. Passaglia, W. Hu, and H. Motohashi, Phys. Rev. D **99**, 043536 (2019).
[44] S. Pi and M. Sasaki, Phys. Rev. Lett. **131**, 011002 (2023).
[45] H. Firouzjahi and A. Riotto, Phys. Rev. D **108**, 123504 (2023).
[46] Y. F. Cai, X. Chen, M. H. Namjoo, M. Sasaki, D. G. Wang, and Z. Wang, J. Cosmol. Astropart. Phys. 05 (2018) 012.
[47] Y. F. Cai, X. H. Ma, M. Sasaki, D. G. Wang, and Z. Zhou, J. Cosmol. Astropart. Phys. 12 (2022) 034.
[48] F. Arroja and T. Tanaka, J. Cosmol. Astropart. Phys. 05 (2011) 005.
[49] C. T. Byrnes, P. S. Cole, and S. P. Patil, J. Cosmol. Astropart. Phys. 06 (2019) 028.
[50] P. S. Cole, A. D. Gow, C. T. Byrnes, and S. P. Patil, arXiv:2204.07573.
[51] P. Carrilho, K. A. Malik, and D. J. Mulryne, Phys. Rev. D **100**, 103529 (2019).
[52] O. Özsoy and G. Tasinato, Phys. Rev. D **105**, 023524 (2022).
[53] S. Pi and J. Wang, J. Cosmol. Astropart. Phys. 06 (2023) 018.
[54] S. Weinberg, Phys. Rev. D **72**, 043514 (2005).
[55] L. Senatore and M. Zaldarriaga, J. High Energy Phys. 12 (2010) 008.
[56] G. L. Pimentel, L. Senatore, and M. Zaldarriaga, J. High Energy Phys. 07 (2012) 166.