YITP-18-128

Primordial Black Holes and Local Non-Gaussianity in Canonical Inflation

Samuel Passaglia,[1, *] Wayne Hu,[1] and Hayato Motohashi[2]

[1] *Kavli Institute for Cosmological Physics, Department of Astronomy & Astrophysics,*
Enrico Fermi Institute, University of Chicago, Chicago, IL 60637
[2] *Center for Gravitational Physics, Yukawa Institute for Theoretical Physics, Kyoto University, Kyoto 606-8502, Japan*

(Dated: December 21, 2018)

Primordial black holes (PBHs) cannot be produced abundantly enough to be the dark matter in canonical single-field inflation under slow roll. This conclusion is robust to local non-Gaussian correlations between long and short-wavelength curvature modes, which we show have no effect in slow roll on local primordial black hole abundances. For the prototypical model which evades this no go, ultra-slow roll (USR), local non-Gaussian correlations have at most an order unity effect on the variance of PBH-producing curvature fluctuations for models that would otherwise fail to form sufficient PBHs. Moreover, the transition out of USR, which is necessary for a successful model, suppresses even this small enhancement unless it causes a large increase in the inflaton kinetic energy in a fraction of an e-fold, which we call a large and fast transition. Along the way we apply the in-in formalism, the δN formalism, and gauge transformations to compute non-Gaussianities and illuminate different aspects of the physical origin of these results. Local non-Gaussianity does not weaken the Gaussian conclusion that PBHs as dark matter in canonical single-field inflation require a complicated and fine-tuned potential shape with an epoch where slow roll is transiently violated.

I. INTRODUCTION

Primordial black holes [1–5] (PBHs) can form in the early universe from the collapse upon horizon reentry of perturbations generated during inflation. PBHs of mass $\sim 10^{-11} M_\odot$ could comprise the dark matter [6–9] if one can evade current astrophysics-dependent constraints from neutron-star capture in globular clusters [10, 11], while PBHs of mass $\sim 10\ M_\odot$ could be responsible for LIGO black-hole merger events [12–16].

PBHs must be formed on physical length scales far removed from the CMB and large-scale-structure modes, where the perturbations are too small, and therefore constraining their abundance and mass provides complementary information about the inflationary epoch. In particular, Ref. [17] showed that PBHs cannot be produced with sufficient abundance to be the dark matter through Gaussian fluctuations in canonical inflation, i.e. by a single scalar field with a canonical kinetic term, without violating the slow-roll (SR) assumption. Therefore any confirmed detection of PBHs as a significant mass fraction of the dark matter would provide evidence beyond canonical slow-roll inflation, e.g. for a violation of slow roll after CMB modes exit the horizon (e.g., [18–20]), a non-canonical kinetic term for the inflaton (e.g., [21]), a multi-field inflationary scenario (e.g., [22, 23]), or some other non-standard cosmological scenario (e.g., [24]).

PBH abundances are sensitive to the full probability distribution of the density contrast averaged on horizon scales at reentry [25–27]. One source of non-Gaussianity in this distribution is the local and position-dependent modulation of the power spectrum induced by the non-Gaussian coupling of short-wavelength PBH modes to long-wavelength modes [28, 29]. In §II we show that this effect vanishes in freely-falling coordinates in slow roll, and therefore Ref. [17]'s no-go theorem for PBH formation in canonical inflation is robust to local non-Gaussianity: any such model which produces large quantities of PBHs must break the slow-roll approximation well before the end of inflation.

In §III we study the prototypical model which evades the no-go theorem by violating the slow-roll approximation – the so-called ultra-slow-roll (USR) inflation [30] – to show analytically that even beyond slow roll, local non-Gaussianity in canonical single-field inflation can have only an order unity effect on the local statistical properties of fluctuations for models which fail to form PBHs as the dark matter under Gaussian assumptions. Along the way we show how the non-Gaussianity in USR can be understood intuitively via the δN formalism [31–36], and we discuss its relationship to locally observable quantities in freely-falling coordinates, specifically the local power spectrum that is relevant to PBHs. In App. A 1 we review how non-Gaussianities are computed numerically through the in-in formalism, and in App. A 2 we clarify how this computation for USR is related through boundary terms to the analytic computation via gauge transformation, usually called a field redefinition in the literature [37].

Observationally viable implementations of USR inflation must have the USR phase be transient. In §IV, we explore the phenomenology of such transient USR models to show that in most circumstances the order-unity effect in pure USR is suppressed. In particular, in §IV A we study a transient USR model proposed in the literature for PBH formation, inflection-point inflation, and show that non-Gaussianities have a much smaller than order unity effect on the local power spectrum.

To understand why the local power spectrum is not enhanced in the inflection-point model and to study gen-

[*] passaglia@uchicago.edu

erally the non-Gaussianity in transient USR, we first review in §IV B the analytic results of Ref. [38] for infinitely sharp potential transitions from USR to SR. We then generalize that study in §IV C to different types of transitions from USR with the help of an adjustable toy model.

We show that transient USR recovers the results of pure USR only when the transition to SR causes a large increase in the inflaton kinetic energy in a fraction of an e-fold, which we call a **large** and **fast** transition. Inflection-point inflation is, by contrast, an example of a **small** and **slow** transition. Thus there is only a limited, but well defined, class of transient models that can have significant non-Gaussian changes to the local power spectrum of inflationary fluctuations, and even these cannot enhance the formation of PBHs sufficiently to be the dark matter unless the model is at least on the threshold of being able to produce them under Gaussian assumptions already. We conclude in §V.

II. NO GO FOR SLOW ROLL

Neglecting the non-Gaussianity of perturbations, Ref. [17] showed that in canonical single field inflation the comoving curvature power spectrum

$$\Delta_\zeta^2 \equiv \frac{k^3}{2\pi^2} P_\zeta \tag{1}$$

must reach at least

$$\Delta_\zeta^2 \sim 10^{-2} \tag{2}$$

within ~ 42 e-folds from the epoch CMB scales exited the horizon, at which $\Delta_\zeta^2 \simeq 10^{-9}$, for the dark matter to be entirely composed of PBHs. In slow roll, the power spectrum satisfies

$$\Delta_\zeta^2 \simeq \frac{H^2}{8\pi^2\epsilon}, \tag{3}$$

with $\epsilon \equiv -d\ln H/dN$, N the negative increasing e-folds to the end of inflation, and the reduced Planck mass $M_{\rm Pl} \equiv (8\pi G)^{-1/2} = 1$ here and throughout. Therefore such an enhancement of Δ_ζ^2 requires a slow-roll violation of at least $\eta \equiv d\ln\epsilon/dN \sim 1$ after horizon exit of the CMB modes but well before the end of inflation.

In this section we update this slow-roll no-go theorem to include local non-Gaussianity, which modulates short-wavelength power in a long-wavelength mode. In particular since the formation of a PBH depends on the density fluctuation averaged on the horizon scale at reentry of the perturbations, horizon scale power that is modulated by superhorizon wavelength fluctuations can in principle enhance formation. We study whether such a modulation can make it possible to produce a substantial fraction of the dark matter in PBHs with slow-roll inflation.

In the presence of a long-wavelength fluctuation ζ_L, low pass filtered for comoving wavenumbers $k \le k_L$, the power spectrum at $k_S \gg k_L$ becomes position dependent

$$P_\zeta(k_S, x) = P_\zeta(k_S)\left[1 + \frac{d\ln P_\zeta(k_S)}{d\zeta_L}\zeta_L(x)\right]. \tag{4}$$

By multiplying by and averaging over the long-wavelength mode,

$$\langle \zeta_L(x) P_\zeta(k_S, x)\rangle_{\zeta_L} \simeq \int \frac{d^3k_L}{(2\pi)^3} B_\zeta(k_L, k_S, k_S), \tag{5}$$

we can relate the power spectrum response to the curvature bispectrum B_ζ,

$$\frac{d\ln P_\zeta(k_S)}{d\zeta_L} \simeq \frac{B_\zeta(k_L,k_S,k_S)}{P_\zeta(k_S)P_\zeta(k_L)} \simeq \frac{12}{5} f_{\rm NL}(k_L,k_S,k_S). \tag{6}$$

Here $f_{\rm NL}$ is the standard dimensionless non-Gaussianity parameter

$$f_{\rm NL}(k_1,k_2,k_3) \equiv \frac{5}{6}\frac{B_\zeta(k_1,k_2,k_3)}{P_\zeta(k_1)P_\zeta(k_2) + \text{perm.}}, \tag{7}$$

in which '+ perm.' denotes the two additional cyclic permutations of indices and the approximation (6) assumes the squeezed limit $k_S \gg k_L$.

In single-field inflation, $f_{\rm NL}(k_L,k_S,k_S)$ has a constrained form when ζ is conserved above the horizon. The curvature perturbation is equivalent to a field fluctuation in spatially flat gauge $\zeta = -\delta\phi/\phi'$, with primes denoting derivatives with respect to e-folds $' = d/dN$ here and throughout. Therefore for a constant ζ, the field fluctuation evolves according to

$$\delta\phi' = \frac{\phi''}{\phi'}\delta\phi, \tag{8}$$

and the phase-space trajectory of the long-wavelength field perturbation follows that of the background itself. Short-wavelength modes evolving in a long-wavelength perturbation then also follow the phase-space trajectory of the background, with the only difference being the local e-folds which determines the relationship between physical and comoving wavenumber (see Fig 1a).

Single-field inflation on the slow-roll attractor (8) therefore satisfies the consistency relation [37]

$$\lim_{k_L/k_S \to 0} \frac{12}{5} f_{\rm NL}(k_L,k_S,k_S) = -\frac{d\ln\Delta_\zeta^2(k_S)}{d\ln k_S}. \tag{9}$$

This implies a modulation of the small-scale power spectrum due to the long-wavelength mode according to Eqs. (4) and (6) as

$$P_\zeta(k_S, x) = P_\zeta(k_S)\left[1 - \frac{d\ln\Delta_\zeta^2(k_S)}{d\ln k_S}\zeta_L(x)\right]. \tag{10}$$

This modulation is zero at the scale where the power spectrum peaks and corresponds to a dilation of scales rather than an amplitude enhancement.

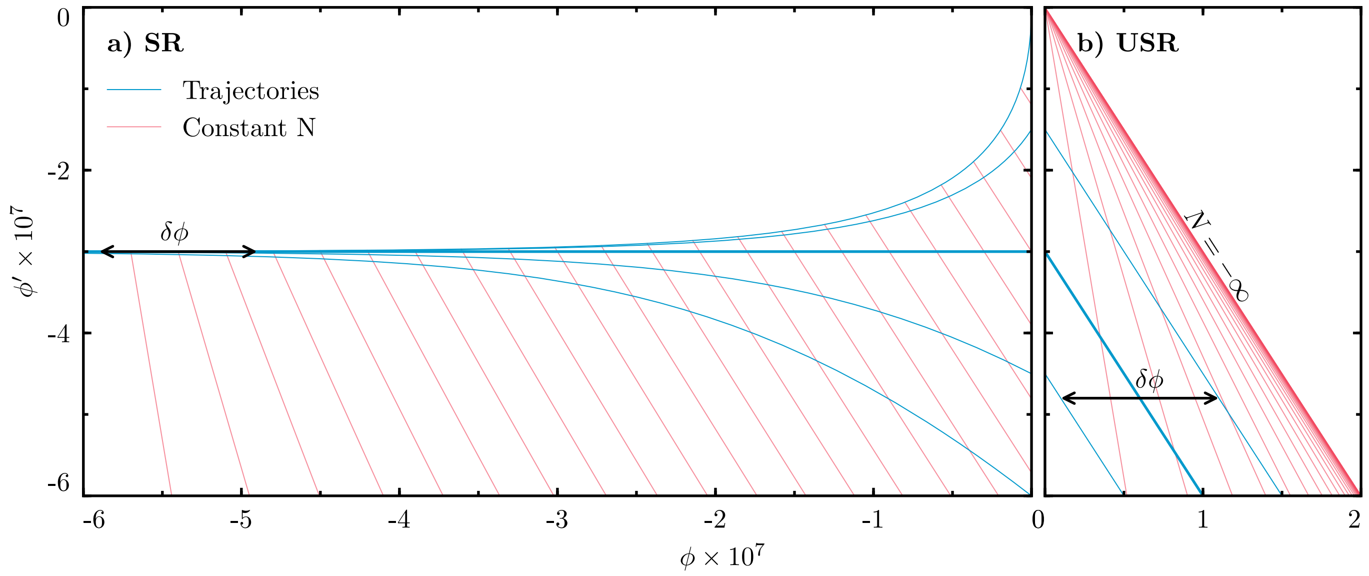

FIG. 1. Phase space diagram for a) slow roll (SR) and b) ultra-slow roll (USR). Shown are background trajectories (blue lines), lines of constant e-folds (red lines) to the end of inflation (left edge of panels), and field fluctuations $\delta\phi$ (arrows). SR trajectories converge to the attractor for different initial kinetic energies at $\phi = 0$. SR field fluctuations $\delta\phi$ follow the attractor trajectory and can be absorbed into a change in e-folds leaving a change in the relationship between local and global coordinates, but no local imprint on observables once clocks are synchronized to a fixed field value at the end of inflation. USR field fluctuations can still be absorbed into a local background but no longer the background of the unperturbed universe (thick blue line). Since different USR trajectories experience different numbers of e-folds to the end, the power spectrum becomes position dependent, with $f_{\rm NL}$ reflecting the e-folding asymmetry between positive and negative $\delta\phi$ or $\partial^2 N/\partial\phi^2$.

In general the physical effect of a dilation of scales is to change the mass scale of PBHs rather than enhance their abundance. However in the slow-roll case, there is actually a change in neither abundance nor mass scale. Though the dilation (10) does occur in global comoving coordinates, in single-field inflation a freely-falling observer will not see this dilation locally.

For a given perturbed metric, the standard Fermi normal coordinates (FNC) [39] can be constructed with respect to a central timelike geodesic of a comoving observer [40, 41], such that $g_{\mu\nu}^{\rm FNC} \simeq \eta_{\mu\nu}$ up to tidal corrections. In order to absorb the effects of superhorizon perturbations out to the horizon scale of a local observer, as required for PBH calculations, we utilize conformal Fermi normal coordinates ($\overline{\rm FNC}$) [42]. $\overline{\rm FNC}$ are constructed such that $g_{\mu\nu}^{\overline{\rm FNC}} \simeq a^2\eta_{\mu\nu}$, i.e. a conformally flat, locally Friedmann-Lemaître-Robertson-Walker (FLRW) form where the global scale factor a of the background universe is evaluated at the proper time of the central observer.

As shown in Ref. [42], for single-field slow-roll inflation the bispectrum in $\overline{\rm FNC}$ is related to the comoving-gauge bispectrum by an additional term proportional to the tilt of the power spectrum as

$$\lim_{k_L/k_S\to 0} B_{\bar\zeta}(k_L,k_S,k_S) = P_\zeta(k_L)P_\zeta(k_S)\frac{d\ln\Delta^2_\zeta(k_S)}{d\ln k_S} + B_\zeta(k_L,k_S,k_S), \tag{11}$$

where barred symbols denote quantities in the $\overline{\rm FNC}$ frame. This additional term neatly cancels the comoving-gauge squeezed bispectrum from the consistency relation (9) and thus in single-field slow-roll inflation

$$\lim_{k_L/k_S\to 0} B_{\bar\zeta}(k_L,k_S,k_S) = 0. \tag{12}$$

There is therefore no modulation of the power spectrum in $\overline{\rm FNC}$

$$P_{\bar\zeta}(k_S,x) = P_{\bar\zeta}(k_S), \tag{13}$$

and the small-scale power spectrum in $\overline{\rm FNC}$ does not depend on the value of the long-wavelength perturbation. All local observers therefore see the same small-scale power spectrum regardless of their position in the long-wavelength mode.

Note that it is only the squeezed non-Gaussianity which vanishes in local observations. Other operators can produce different shapes of the bispectrum. However, in canonical slow-roll inflation these terms are proportional to ϵ and thus extremely suppressed if one hopes to enhance the power spectrum by shrinking ϵ.

Physically, the cancellation in Eq. (12) occurs because the bispectrum from the consistency relation encodes the effect on small-wavelength modes of evolving in a separate universe with a background evolution defined by the long-wavelength mode. Once the long-wavelength mode is frozen, this effect is just to change coordinates in the separate universe relative to global coordinates. When making local observations, an observer knows nothing of the global coordinates and instead makes measurements in coordinates corresponding to the separate universe. The formation of PBHs is a local process and so their properties also do not depend on their position in the long-wavelength mode.

This lack of local modulation can be also understood from the phase-space diagram Fig. 1a. Relative to the end of inflation at a fixed field value, perturbed trajectories in slow roll are indistinguishable from the background trajectory and thus observers making measurements relative to the end of inflation cannot from any local measurement decide whether they inhabit different regions of a long-wavelength curvature perturbation.

This leads us to our first conclusion: squeezed non-Gaussianity cannot produce PBHs as a significant fraction of the dark matter in canonical single-field slow-roll inflation. For such PBHs to form in canonical single-field inflation, the slow-roll approximation must be violated, at least transiently, to either produce large Gaussian or non-Gaussian fluctuations. In this sense, the slow-roll no-go theorem shown in Ref. [17] is robust and does not change.

Models that evade this no-go result typically have a period when the inflaton rolls on a very flat potential where Hubble friction is insufficient to keep the inflation on the slow-roll attractor. The ultra-slow-roll model, where the inflaton potential is perfectly flat, provides the prototypical example for such studies as we shall see next.

III. ULTRA-SLOW-ROLL INFLATION

Ultra-slow roll [30] is a model of single-field inflation which greatly enhances the scalar power spectrum while also breaking the single-field consistency relation (9) for the squeezed bispectrum by violating the attractor condition (8) [43, 44]. It is therefore possible to spatially modulate the local power in small scale density fluctuations relevant for PBHs with long-wavelength modes. In this section we examine whether this non-Gaussian modulation can significantly enhance the PBH abundance in ultra-slow roll.

USR is characterized by a potential which is sufficiently flat before its end, which we denote with $\phi = 0$, that the Klein-Gordon equation takes the form

$$\ddot{\phi} \simeq -3H\dot{\phi}, \tag{14}$$

where here and throughout overdots denote derivatives with respect to the coordinate time t. If the potential energy dominates then $H \simeq$ const. and Eq. (14) then implies $\phi'' \simeq -3\phi'$ and hence $\phi' \simeq -3\phi + \text{const.}$, defining a family of trajectories in the phase-space diagram, as depicted by the blue trajectories in Fig. 1b. Therefore, the phase-space trajectory of the background evolution depends on the initial kinetic energy and does not exhibit attractor behavior.

For an exactly flat potential at $\phi > 0$, an inflaton with insufficient initial kinetic energy will not cross the plateau to reach $\phi = 0$, neglecting stochastic effects. In Fig. 1b we focus on classical trajectories that can reach $\phi = 0$ within finite e-folds, and hence the upper right triangle region is inaccessible.

The solution to Eq. (14) is $\dot{\phi} \propto a^{-3}$ and so $\epsilon \propto a^{-6}$ and $\eta = -6$. Since the analytic solution of the Mukhanov-Sasaki equation for ζ in the superhorizon limit is given by

$$\zeta \simeq c_1 + c_2 \int \frac{dt}{a^3 \epsilon}, \tag{15}$$

with integration constants c_1 and c_2, it is dominated by the second mode which grows in USR since $(a^3\epsilon)^{-1} \propto a^3$ rather than decays as it does in slow roll. With $H \simeq$ const., Eq. (15) gives $\zeta \propto a^3$ and hence in the spatially flat gauge $\delta\phi = -\zeta\phi' = \text{const.}$, implying that $\delta\phi' = 0$, unlike the case of the slow-roll attractor (8).

The power spectrum in this model depends on the value of ϵ at the end of USR [see Eq. (A17)],

$$\Delta_\zeta^2 \simeq \frac{H^2}{8\pi^2 \epsilon_{\rm end}}, \tag{16}$$

and thus can be very large if $\epsilon_{\rm end} \ll 1$. One can employ a gauge transformation from spatially flat gauge to comoving gauge to show that the squeezed-limit non-Gaussianity takes the form (see Appendix A 2)

$$\lim_{k_L/k_S \to 0} \frac{12}{5} f_{\rm NL}(k_L, k_S, k_S) = 6. \tag{17}$$

Since the USR power spectrum is scale invariant, the large value of $f_{\rm NL}$ in USR violates the consistency relation.

The physical origin of this large value for $f_{\rm NL}$ can be seen from the phase-space diagram Fig. 1b. Due to the initial kinetic energy dependence of the background evolution, a USR perturbation cannot be mapped into a change in the background clock along the same phase-space trajectory. Instead, long-wavelength perturbations $\delta\phi$ carry no corresponding $\delta\phi'$ and so shift the USR trajectory to one with a different relationship between ϕ and ϕ'. On this shifted trajectory, the short-wavelength power spectrum attains a different value at the end of USR. More generally, if a local measurement is sensitive to ϕ' at the end of inflation, as in the case of $\Delta_\zeta^2(k_S)$, then different observers will produce different measurements depending on their position in the long-wavelength mode.

This graphical representation of $f_{\rm NL}$ can be turned in to a computational method through the so-called δN formalism [31–34]. When the expansion shear for a local observer is negligible, as it is in USR above the horizon, the nonlinear evolution of the curvature fluctuation follows the evolution of local e-folds. On spatially flat hypersurfaces, the field fluctuation can be absorbed into a new conformally flat FLRW background on scales much shorter than the wavelength and so the local e-folds may be calculated from the Friedmann equation of a separate universe. The position dependent power spectrum is therefore the second order change in e-folds due to a short-wavelength $\delta\phi_S$ on top of a long-wavelength $\delta\phi_L$. Since in USR these perturbations leave ϕ' unchanged, the non-Gaussianity parameter can be computed from the e-folds as a function of phase-space position of the background $N(\phi,\phi')$ as

$$\frac{12}{5}f_{\rm NL} = 2\frac{\partial^2 N}{\partial\phi^2}\bigg/\left(\frac{\partial N}{\partial\phi}\right)^2 , \tag{18}$$

at fixed ϕ'.

The consequence of this formula can be visualized through Fig. 1b as the effect of perturbations on phase-space trajectories. Around a chosen background trajectory, the long-wavelength perturbation is reabsorbed into a new background, a horizontal shift to a new trajectory. Short-wavelength perturbations living in this new background induce a second shift in the trajectory, hence the second derivative. Visually, the fact that for the same amplitude of field fluctuation $|\delta\phi|$, a positive fluctuation intersects more surfaces of constant N than a negative fluctuation indicates a large $f_{\rm NL}$. Refs. [38, 43, 45, 46] follow this approach to analytically compute its value in complete agreement with the in-in approach App. A 1 or the gauge-transformation App. A 2. We shall again exploit the δN formalism in §IV.

Despite the violation in the consistency relation, the coordinate transformation for the bispectrum Eq. (11) still holds and the transformation from global comoving coordinates to $\overline{\rm FNC}$ leads to the same additional tilt-dependent term in the bispectrum as in the canonical case so long as the transformation to $\overline{\rm FNC}$ is performed when modes are frozen outside the horizon after the end of inflation.[*1] After this time, the construction follows Ref. [42] exactly. This procedure of transforming coordinate systems after inflation is followed for slow-roll inflation in Ref. [47] to compute the next-to-leading order term in the bispectrum transformation. Practically, it corresponds to the clock-synchronization condition that all local observers make their measurements at fixed proper time after the end of inflation.

Given the scale invariance of the spectrum, the tilt-dependent transformation from comoving gauge to $\overline{\rm FNC}$ leaves neither an enhancement of the local power in the long-wavelength mode nor a modulation of the mass of the PBHs. On the other hand, since the transformation term no longer cancels with the comoving-gauge $f_{\rm NL}$ itself, a large value of the latter can in principle enhance PBH formation locally.

If $f_{\rm NL}(k_L,k_S,k_S)$ is described by the USR result Eq. (17), then the local power spectrum can be enhanced by a factor $12/5\times f_{\rm NL}\times\zeta_L = 6\times\zeta_L$. Therefore the non-Gaussian response enhances the local power spectrum by an order unity quantity unless the long-wavelength mode is large, i.e.

$$\zeta_L \gtrsim 10^{-1}. \tag{19}$$

However, the scale invariance of USR would then imply

$$\Delta_S^2 = \Delta_L^2 \sim \langle\zeta_L^2(x)\rangle \gtrsim 10^{-2}, \tag{20}$$

which satisfies the criterion Eq. (2) for PBH formation, and therefore PBHs would already be produced at scale k_S even before accounting for the non-Gaussian response. Note that the conversion from Δ_ζ^2 to spatial variance involves a summation over k_L and gives a logarithmic factor which depends on the total e-folds of USR. In a realistic model this logarithmic factor must be finite so as to also satisfy constraints from the CMB.

This result is the second main conclusion of this work: in a USR model which does not produce a significant PBH abundance under the Gaussian approximation, the non-Gaussian response enhances the local power spectrum by at most

$$\frac{\Delta P_\zeta}{P_\zeta} \lesssim 1, \tag{21}$$

and therefore the non-Gaussian response does not qualitatively change Gaussian conclusions. Of course as they originate from rare fluctuations, PBHs can change in their abundance but these changes can be reabsorbed into model parameters that make no more than an order unity change in the power spectrum. In particular local non-Gaussianity cannot make a model that falls far short of making PBHs the dark matter under the Gaussian assumption into one that does.

Since inflation has to end and observational constraints should be satisfied on CMB scales, the simple picture presented here must be modified to account for transitions into and out of USR. In §IV we shall explore whether even this level of enhancement still holds in such models of transient USR inflation.

IV. TRANSIENT ULTRA-SLOW ROLL

In addition to a graceful exit problem, USR inflation is incompatible with the measured tilt of the CMB power spectrum [48] and is in tension with constraints on local non-Gaussianities in the CMB [49], and therefore any USR phase must begin after CMB modes exit the horizon and must take care not to grow those modes after horizon exit.

[*1] $\overline{\rm FNC}$ can still be established during the USR phase but are more closely related to spatially flat gauge than comoving gauge in temporal synchronization (see also App. A 2). In spatially flat gauge, a superhorizon field fluctuation $\delta\phi$ can be absorbed into a new, nearly conformally flat FLRW background, as we exploit with the δN formalism.

One model proposed in the literature for PBH production with a transient USR phase is inflection-point inflation [17, 20]. In §IV A, we show that the transition out of USR in inflection-point inflation induces

$$f_{\rm NL}(k_L, k_S, k_S) \ll 1, \tag{22}$$

and therefore non-Gaussianities do not enable PBHs to be the dark matter in inflection-point inflation.

This numerical result can be understood from Ref. [38]'s analytic study of infinitely sharp potential transitions between USR and SR, which we review briefly in §IV B. Transitions where the inflaton velocity monotonically decreases to reach an attractor solution lead to squeezed non-Gaussianity that is proportional to the potential slow-roll parameters on the attractor. Conversely, transitions where the inflaton instantly goes from having too much kinetic energy for the potential it evolves on to suddenly having insufficient kinetic energy for a now much steeper potential conserve the USR non-Gaussianity. We call the later transitions **large**, which we will define specifically below [see Eq. (35)].

In §IV C, we generalize the analysis of Ref. [38] to potentials which do not have an infinitely sharp break, and in particular we study how quickly the inflaton must traverse the potential feature to reproduce the USR result. We show that to conserve the USR result Eq. (17) the transition must be **fast** in that it completes in a small fraction of an e-fold [see Eq. (41)].

We conclude that the large $f_{\rm NL}$ of USR will only be preserved if the transition to SR is both large and fast. For all other cases, the enhancement to the local power spectrum

$$\frac{\Delta P_\zeta}{P_\zeta} \ll 1, \tag{23}$$

and so non-Gaussianity in transient USR does not generally affect the conclusions on PBH formation.

A. Slow–Small Transition: Inflection-Point Inflation

Inflection-point inflation is characterized by a potential which supports a slow-roll phase when CMB scales exit the horizon followed by a slow-roll violation and subsequent ultra-slow-roll phase which enhances the power spectrum at small scales. This USR phase is generally unstable and lasts just a few e-folds before the inflaton loses enough kinetic energy to lock onto the attractor solution of the potential and slow-roll inflation resumes [50]. We call this transition **slow** because the inflaton kinetic energy decreases monotonically to the slow-roll value, and **small** because the potential slow-roll parameters on the attractor are comparable to the kinetic energy at the end of the USR phase.

We consider an inflection potential of the form explored in Ref. [17] following Ref. [20],

$$V(\phi) = \frac{\lambda v^4}{12} \frac{x^2(6-4ax+3x^2)}{(1+bx^2)^2}, \tag{24}$$

where $x = \phi/v$. We study this model with the parameters $\{a, b-1, \lambda, v\} = \{3/2,\ 4\times10^{-5},\ 7\times10^{-8},\ 0.658\}$. In terms of the auxiliary variables of Refs. [17, 20], this model has $\{\beta,\ \Delta N_{\rm SR}\} = \{4\times10^{-5},\ 125\}$.

These parameters are finely tuned to significantly suppress ϵ after the CMB scale $k_0 = 0.05$ Mpc^{-1} exits the horizon 55 e-folds before the end of inflation while also preventing the inflection point from trapping the inflaton for too many e-folds. Nonetheless our qualitative results for the non-Gaussianity are not sensitive to the specific functional form of the potential nor the parameter set above.

Note that even with fine tuning, this model does not fit observational constraints from the CMB (e.g. [48]) because the power spectrum is too red (scalar slope $n_s = 0.91$) due to the proximity of the inflection point to CMB scales. This additional red tilt implies a larger value of ϵ at CMB scales and hence a larger relative suppression of ϵ and growth of the power spectrum during the USR phase. Without this enhancement, the inflection model falls far short of forming PBHs as the dark matter [17] and so we choose these parameters to study whether models on the threshold of forming sufficient PBHs for Gaussian fluctuations can be made to do so through non-Gaussianity in the model.

Along the inflaton trajectory, the potential (24) has a single inflection point, where $d^2V/d\phi^2 = 0$ is satisfied, between two close points where $dV/d\phi = 0$. In this region the slope of the potential is tiny, and hence the USR condition $|dV/d\phi| \ll |\phi'|H^2$ is satisfied briefly, after which slow roll quickly resumes. The evolution of the slow-roll parameters ϵ and η in this model are shown in Fig. 2. The model exhibits a transient period where $\epsilon \propto a^{-6}$ and thus the USR result $\eta \simeq -6$ is temporarily achieved.

The upper panel of Fig. 3 shows the power spectrum Δ^2_ζ produced by inflection-point inflation with the potential (24), computed by numerically solving the Mukhanov-Sasaki equation (A3). Modes which exit the horizon well before the USR phase do not grow outside the horizon and their power spectrum satisfies the slow-roll result (3). Modes which exit the horizon shortly before the USR phase do, however, grow outside the horizon leading to the rise before peak power in the USR phase.

This behavior can be understood in more detail from the exact, but formal, solution of the Mukhanov-Sasaki equation (A3) [51]

$$\zeta' = -\frac{1}{a^3\epsilon H}\left[\int \frac{da}{a} a^3 \left(\frac{k}{aH}\right)^2 (\epsilon H)\zeta + \text{const.}\right]. \tag{25}$$

In the SR phase, ϵ is roughly constant and the growing integrand provides the leading contribution

$$\zeta' \simeq -\left(\frac{k}{aH}\right)^2 \zeta, \qquad \text{(SR)}, \tag{26}$$

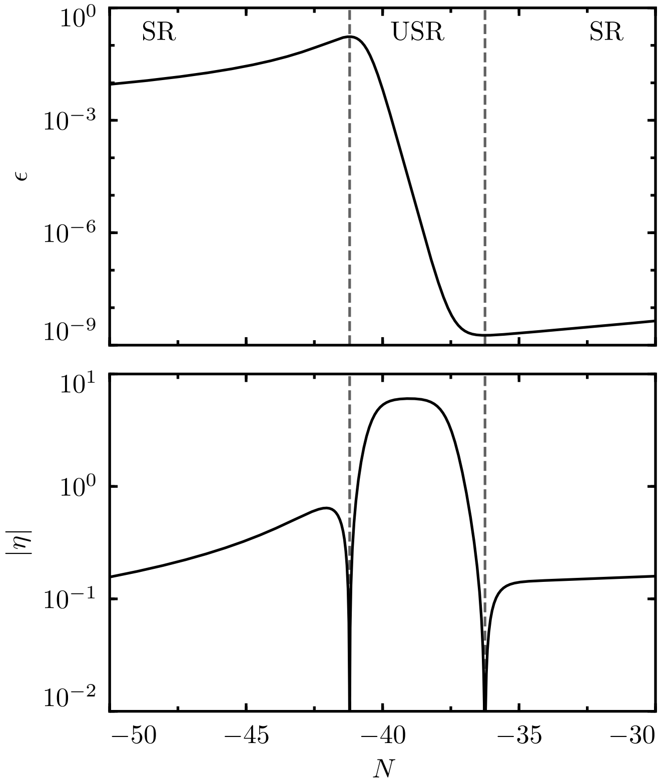

FIG. 2. Inflection-point model background parameters $\epsilon \equiv -d\ln H/dN$ and $\eta \equiv d\ln\epsilon/dN$. η experiences two zero-crossings as ϵ reaches critical points entering and exiting the USR period, which we use to delineate the USR phase from the SR phases. Here the transient period lasts for ~ 5 e-folds, but only achieves $\eta \simeq -6$ for a shorter period.

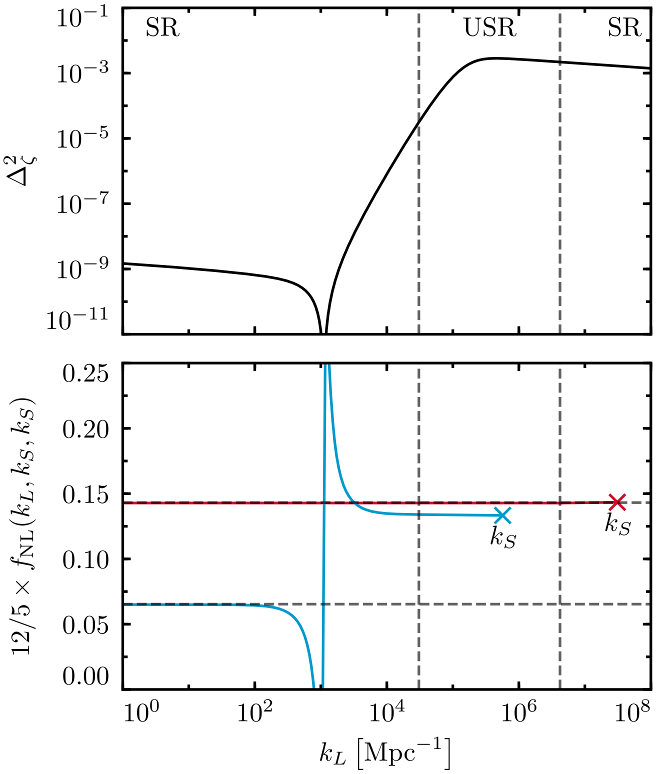

FIG. 3. Inflection-point model power spectrum Δ^2_ζ and non-Gaussianity parameter $f_{\rm NL}(k_L, k_S, k_S)$. The vertical dashed lines delineate modes which cross the horizon during the SR and USR periods (see Fig. 2). For $f_{\rm NL}$, ×'s denote values for $k_S > k_L$ with blue a mode that exits the horizon during USR and red after USR. The horizontal dashed lines denote the consistency relation expectation for the two modes and the spike in $f_{\rm NL}$ reflects the near zero in Δ^2_ζ rather than a large bispectrum (see §IV A for further discussion).

and hence the curvature perturbation $\zeta \propto e^{\frac{1}{2}\left(\frac{k}{aH}\right)^2}$ freezes out to a constant for $k/(aH) \ll 1$ as in Eq. (15). On the other hand, in USR phase, since $\zeta \propto a^3$ outside the horizon from (15), it immediately holds that

$$\zeta' \propto a^3, \qquad \text{(USR)}, \tag{27}$$

outside the horizon. One can also see that (27) is consistent with (25) as follows. With $\zeta \propto a^3$, the integral in (25) acts as $\simeq \int d\ln a\, a^{-2}$, which is dominated by early times and hence converges to a constant, whereas the prefactor grows as $\propto a^3$, resulting in (27). Thus for a mode which spends $N_{\rm SR}$ e-folds outside the horizon in slow roll, it takes $N_{\rm USR} = 2/3 \times N_{\rm SR}$ e-folds of USR inflation to raise ζ' back to order unity. Therefore at a fixed duration $N_{\rm USR}$ of USR inflation, modes which exit the horizon more than $3/2 \times N_{\rm USR}$ e-folds before USR remain constant while modes which exit within $3/2 \times N_{\rm USR}$ grow outside the horizon. After the USR phase, modes freeze in and the smooth change in the slope of the potential assures a slow increase in ϵ and a smooth transition of the power spectrum to the final SR phase.

The power spectrum shown in the upper panel of Fig. 3 exhibits a near-zero minimum $\Delta^2_\zeta \sim 8 \times 10^{-16}$. Similar behavior occurs in other models in which the growing mode overtakes the constant mode (see, e.g., Ref. [52] and §IV C). This phenomenon can be also understood in detail from the formal solution Eq. (25), in which it can be seen that in slow roll the superhorizon mode approaches its slow-roll freezeout value with decreasing amplitude, i.e. with

$$\text{Arg}\left[\frac{\zeta'}{\zeta}\right] = \pi + \mathcal{O}\left(\frac{k}{aH}\right), \qquad \text{(SR)}, \tag{28}$$

where the order of the correction follows from using the approximate SR form (A4) in (25).

While $\zeta \propto a^3$ in the USR superhorizon limit, at the onset of USR, the curvature perturbation must reach this limit from the SR side. Let $a = a_*$ at the onset of the USR phase, then the curvature evolves as

$$\zeta' = \zeta'\Big|_{a_*} \times \left(\frac{a}{a_*}\right)^3, \qquad \text{(USR)}, \tag{29}$$

with the boundary condition $\zeta'|_{a_*}$ given approximately by the SR solution for a smooth transition. Given the relative sign in the leading order SR expression (26), this represents an increase in the decay rate of $|\zeta|$ and thus before modes can grow as $\zeta \propto a^3$ in USR they must reverse sign.

There is a mode which experiences just enough evolution outside the horizon by the end of the USR phase to go from its freezeout value to near-zero. The corresponding value of the power spectrum at the minimum is determined by the small out-of-phase component, i.e. how close Eq. (28) is to π and therefore how far outside the horizon this mode is when USR begins. Thus the longer the USR phase is, the deeper the minimum is.

Modes which exit after this minimum are dominated by their superhorizon growth, and as modes exit the horizon closer to the USR phase they grow for a longer period and thus the power spectrum grows with increasing k. While a prolonged USR phase leads to a constant Δ_ζ^2 for modes which exit the horizon during USR [cf. Eq. (16)], the inflection model touches the $\eta \simeq -6$ phase only briefly and the power spectrum therefore exhibits a peak $\Delta_\zeta^2(k_{\rm Peak}) = 2.8 \times 10^{-3}$. This peak falls a factor of a few short of the value $\Delta_\zeta^2 \sim 10^{-2}$ required for PBHs to form all the dark matter (see §II and note that a model with the right tilt at CMB scales must fall much further short of this requirement [17]). After the USR phase, the model returns to the slow-roll attractor and Δ_ζ^2 is once more described by Eq. (3).

It is now interesting to ask whether the power spectrum of the upper panel of Fig. 3 can be locally enhanced by a factor of a few to exceed the threshold (2) for PBH dark matter. According to Eqs. (4) and (6), for the power spectrum at a short-wavelength scale k_S to be significantly enhanced, we require a large long-wavelength perturbation ζ_L and a large correlation $f_{\rm NL}$.

In the lower panel of Fig. 3, we plot $f_{\rm NL}(k_L, k_S, k_S)$ as a function of the long-wavelength mode k_L for two different values of the short-wavelength mode k_S. The red upper curve shows $f_{\rm NL}$ for a short-wavelength mode which exits the horizon after the end of USR, while the blue lower curve shows $f_{\rm NL}$ for a short-wavelength mode which exits the horizon during USR. The upper and lower horizontal dashed lines show the consistency relation expectation $f_{\rm NL}$ in the limit $k_L/k_S \to 0$.

The numerically computed bispectrum for a short-wavelength mode which exits the horizon after USR, the red upper curve, agrees with the consistency relation. In other words, the short-wavelength perturbation ζ_S retains no memory that, while it was inside the horizon, the long-wavelength perturbation ζ_L outside the horizon grew in USR. This is because $f_{\rm NL}(k_L, k_S, k_S)$ is set when k_S exits the horizon and the modes ζ_L are already frozen at this time. Because the transformation of the bispectrum to $\overline{\rm FNC}$ involves a subtraction of the consistency relation component, Eq. (11), we conclude that short-wavelength modes which exit after the USR phase show no response to long-wavelength modes in local coordinates and therefore no enhancement of local PBH abundance.

For a short-wavelength mode which exits the horizon during USR, the blue curve, the above logic does not hold. The numerically computed bispectrum $f_{\rm NL}(k_L, k_S, k_S)$ does not agree with the consistency relation when ζ_L grows outside the horizon. For such triangles, $12/5 \times f_{\rm NL} \simeq 0.13$ while the consistency relation predicts $12/5 \times f_{\rm NL} \simeq 0.065$.

Conversely, for the frozen ζ_L modes that correspond to modes that exited the horizon well before USR, the consistency relation for $f_{\rm NL}$ does hold. This is a successful test of our numerical computation, since in this limit the long-wavelength mode remains constant outside the horizon and just shifts the local coordinates for the small-wavelength mode along the background trajectory.

Fig. 3 also shows that when k_S exits the horizon during USR, the near-zero of Δ_ζ^2 induces a feature on $f_{\rm NL}(k_L, k_S, k_S)$. This is due to the division by the power spectrum in the definition of $f_{\rm NL}$, Eq. (7). In particular, when k_S exits the horizon ζ_L has not yet reached its final (tiny) value set at the end of USR and thus a non-zero bispectrum $B_\zeta(k_L, k_S, k_S)$ can be obtained. After the end of USR, ζ_L is very small and thus $f_{\rm NL}$ is amplified. The physical effect of this feature is negligible since, to obtain the power spectrum response, $f_{\rm NL}$ should be multiplied by ζ_L, which has a minimum at this feature.

More generally, the USR phase does enhance $f_{\rm NL}$ relative to the consistency relation value. Hence the non-Gaussianity in $\overline{\rm FNC}$, $\bar{f}_{\rm NL}$, is non-zero. There is therefore an enhancement of the PBH abundances due to local non-Gaussianity, which is not the case in single-field inflation on the attractor. However, for the k_S shown in Fig. 3 in blue, both $f_{\rm NL}$ and $d \ln \Delta_\zeta^2 / d \ln k$ are so small that, once multiplied by $\zeta_L \ll 1$, the position-dependent effect on ζ_S is insignificant.

Quantitatively, we can summarize the PBH abundance enhancement in this model by choosing k_S and k_L in the USR phase, where $f_{\rm NL}$ is nearly constant. In particular, to eliminate the tilt-dependent coordinate effects on the abundance and to maximize the Gaussian part of the power spectrum, we can choose $k_S = k_L = k_{\rm Peak}$. This triangle is not squeezed but since in USR $f_{\rm NL}$ is the same for all triangle shapes, this triangle does serve as a summary statistic for local non-Gaussianity in inflection-point inflation.

Doing so, we compute numerically that $12/5 \times f_{\rm NL}(k_{\rm Peak}, k_{\rm Peak}, k_{\rm Peak}) = 0.13$. To obtain the response, we set ζ_L to the peak value $\zeta^{\rm RMS}(k_{\rm Peak}) = \sqrt{\Delta_\zeta^2(k_{\rm Peak})} \simeq 0.05$. Local non-Gaussianity can therefore enhance the local power spectrum by a factor of at most ~ 0.006. We therefore conclude that in inflection-point inflation, the power spectrum enhancement due to non-Gaussianity is negligible,

$$\frac{\Delta P_\zeta}{P_\zeta} \ll 1. \tag{30}$$

In this model, we do not recover the USR squeezed limit result $12/5 \times f_{\rm NL} \simeq 6$ and therefore do not enhance the small-scale power spectrum by an order-unity quantity. This is a reflection of the analytic result of Ref. [38]

that transitions from a USR phase to a SR phase which are monotonically decreasing in the field velocity suppress the USR non-Gaussianity.

In the following sections, we will show that this suppression of non-Gaussianity is generic to transition models, except for the special case where the transition is both fast and large.

B. Infinitely Fast Transitions

Infinitely fast transitions from USR to SR were considered in Refs. [38] and [50], and Ref. [38] established analytically that the final level of non-Gaussianity is sensitive to the way USR is exited.

Specifically Ref. [38] considered the case where a pure ultra-slow-roll potential is joined to a slow-roll potential $V_{\rm SR}$ at a field position which we label ϕ_2 for ease of generalization later:

$$V(\phi) = \begin{cases} V_{\rm SR}(\phi_2), & \phi > \phi_2 \quad \text{(USR)} \\ V_{\rm SR}(\phi), & \phi \le \phi_2 \quad \text{(SR)} \end{cases} \tag{31}$$

and hence the potential has an infinitely sharp discontinuity in slope at ϕ_2. We call this an infinitely fast transition from USR to SR because the inflaton rolls over this discontinuity instantaneously. $V_{\rm SR}(\phi)$ can be characterized in general by the potential slow roll parameters for $\phi \le \phi_2$

$$\epsilon_V \equiv \frac{1}{2}\left(\frac{1}{V}\frac{dV}{d\phi}\right)^2, \quad \eta_V \equiv \frac{1}{V}\frac{d^2V}{d\phi^2}. \tag{32}$$

The transition can be characterized by the strictly positive amplitude parameter[*2]

$$h \equiv \sqrt{\frac{\epsilon_V(\phi_2^-)}{\epsilon(\phi_2^+)}}, \tag{33}$$

where

$$\phi_2^\pm \equiv \lim_{\omega\to 0} \phi_2 \pm \omega, \tag{34}$$

which is a ratio between the potential slow-roll parameter at the beginning of the SR phase and the Hubble slow-roll parameter ϵ at the end of the USR phase.

If $h = 1$, the kinetic energy at the end of USR is just enough to keep the field on the attractor of the SR phase. The $h \ll 1$ limit therefore corresponds to the small transition, a monotonic transition from USR to the SR attractor where the inflaton continues to slow down before hitting the attractor and hence the power spectrum continues to evolve. Conversely for $h \gg 1$, Ref. [38] showed that the perturbations freeze out at $N(\phi_2)$. We call this a **large** transition because the inflaton instantly goes from having too much kinetic energy for the potential it evolves on to suddenly having insufficient kinetic energy for a now much steeper potential

$$h \gg 1 \implies \epsilon_V \gg \epsilon = \frac{1}{2}\phi'^2 , \tag{35}$$

Since perturbations do not freezeout immediately for a finite value of h the final level of non-Gaussianity is not given by Eq. (17) but rather can be shown analytically to be [38]

$$\lim_{k_L/k_S\to 0} \frac{12}{5} f_{\rm NL}(k_L, k_S, k_S) = 2\frac{h(3h+\eta_V)}{(h+1)^2}, \tag{36}$$

for scales k_L, k_S which cross the horizon during USR. Eq. (36) yields the USR result Eq. (17) only in the limit $h \gg 1$, and thus for infinitely fast transitions the USR non-Gaussianity is conserved only when the transition is large.

The enhancement of the local power spectrum is suppressed for small transitions, but we shall next see that it is also suppressed if the transition is not sufficiently fast. Therefore, the transition needs to be large and fast to recover the USR non-Gaussianity. In contrast, the inflection model of the previous section is an example where the transition is both small and slow.

C. Fast/Slow–Large/Small Transitions

In order to study in more detail the phenomenology of transient USR inflationary phases beyond the slow-small transition of §IV A and the infinitely fast limit of §IV B, we construct a toy inflationary model which begins in SR, enters a USR phase, and then transitions back to SR. We implement this with a potential where the slope of an otherwise linear potential makes two transitions across adjustable widths in field space

$$\frac{dV}{d\phi}(\phi) = \frac{\beta}{2}\left[1+\tanh\left(\frac{\phi-\phi_1}{\delta_1}\right)\right] + \frac{\gamma}{2}\left[1+\tanh\left(\frac{\phi_2-\phi}{\delta_2}\right)\right], \tag{37}$$

and hence

$$V(\phi) = V_0 + \frac{\beta}{2}\left[\phi + \log\left\{\cosh\left(\frac{\phi-\phi_1}{\delta_1}\right)\right\}\right] + \frac{\gamma}{2}\left[\phi - \log\left\{\cosh\left(\frac{\phi_2-\phi}{\delta_2}\right)\right\}\right]. \tag{38}$$

This potential describes three phases with finite transitions, which is a natural generalization of the model (31) with two phases with instant transition considered in §IV B. The model parameters $\{\delta_1,\ \delta_2\}$ and $\{\phi_1,\ \phi_2\}$ determine the widths and positions of two transitions, respectively. The limit $\delta_1, \delta_2 \to 0$ amounts to instant transitions, where the potential is composed of a flat plateau

[*2] The h defined in (33) is equivalent to Ref. [38]'s $-h/6$.

of amplitude V_0 for $\phi_2 < \phi < \phi_1$ in between two linear pieces of slope $\{\beta,\ \gamma\}$, which we set positive. By modifying these parameters, we can set the duration of the USR phase as well as the circumstances of its beginning and end. By constructing the transition in $dV/d\phi(\phi)$ rather than in $V(\phi)$ directly, we trivially obtain a monotonic potential where the field always rolls downhill.

Since this is a toy model that we introduce to illustrate the fast/slow and large/small distinction, we do not attempt to accurately fit measurements at CMB scales or to appropriately end inflation. Once all modes which we are interested in have frozen out in the latter slow-roll phase, we end inflation by hand. By adjusting the potential before and after the plateau, one could turn this toy model into a model which can fit CMB observations and produce PBHs while ending inflation gracefully without changing the conclusions we draw below.

For $\phi > \phi_1$ the potential has a positive slope $\simeq \beta$ and the inflaton follows the slow-roll attractor. In order for the inflaton to leave the slow-roll attractor and enter a USR phase, the transition must be sufficiently sharp that the inflaton enters the flat region of the potential with excess kinetic energy. Thus we fix the entry parameters $\{\beta,\ \phi_1,\ \delta_1\} = \{10^{-14},\ 0,\ 10^{-2}\}$ to guarantee such a transition. By having inflation start on the slow-roll attractor, we are freed from having to specify initial conditions during USR.

The region $\phi_2 < \phi < \phi_1$ marks the USR phase where the potential is approximately flat. We fix the amplitude of the potential in the flat plateau $V_0 = 2 \times 10^{-14}$, chosen to ensure that USR modes in our fiducial model are still perturbative, i.e. $\Delta^2_\zeta \lesssim 1$ for the durations of USR we consider here.

Finally, for $\phi < \phi_2$ the potential has a positive tilt $\simeq \gamma$ and the inflaton returns to the slow-roll attractor. Among the remaining parameters $\{\phi_2,\ \delta_2,\ \gamma\}$, ϕ_2 determines the duration of the USR period, and $\{\delta_2, \gamma\}$ set the circumstances of the exit from USR. ϕ_2 in particular must be very finely tuned to allow several e-folds of USR inflation while still reaching the transition point in a reasonable amount of time. The instant transition of §IV B corresponds to taking $\delta_2 \to 0$ and to focusing on the inflaton behavior around $\phi = \phi_2$.

We generalize Ref. [38]'s analysis to transitions of finite width between the flat and slow-roll potentials by allowing $\delta_2 \neq 0$. We start by generalizing the definition for the start and end of the transition, Eq. (34). We choose the end of the transition ϕ_2^- from the potential through

$$\phi_2^- \equiv \phi_2 - 2\delta_2. \tag{39}$$

The beginning of the transition, ϕ_2^+, is not simply $\phi_2 + 2\delta_2$ since the USR phase persists while $\epsilon_V \ll \epsilon$. Instead, we choose to define the beginning of the transition through the deviation from the USR analytic solution,

$$1 - \frac{\phi'_{\rm USR}}{\phi'}\bigg|_{\phi_2^+} = 0.05 \times \left(1 - \frac{\phi'_{\rm USR}}{\phi'}\bigg|_{\phi_2^-}\right), \tag{40}$$

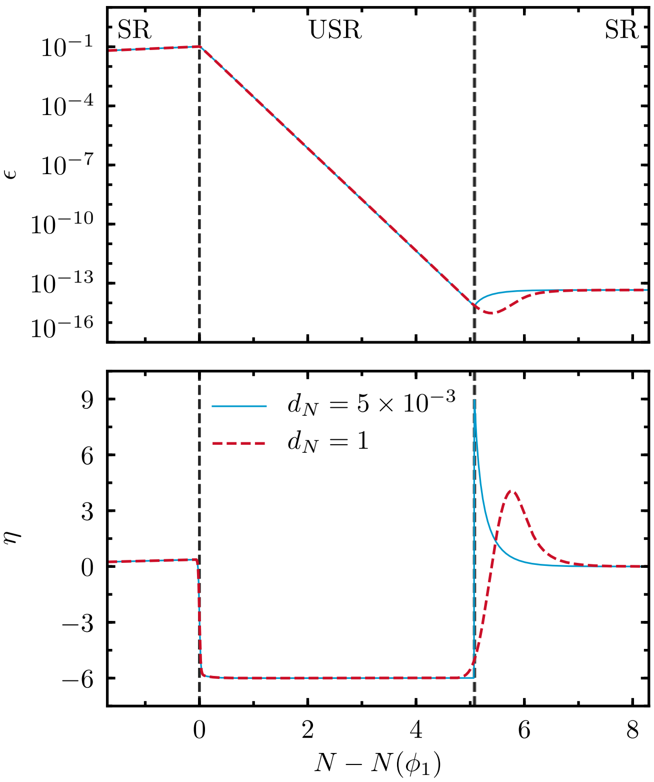

FIG. 4. Transition model background parameters $\epsilon \equiv -d\ln H/dN$ and $\eta \equiv d\ln\epsilon/dN$ for a fast vs. slow transition, denoted by the e-fold width d_N, with a fixed large transition ($h = 2.5$). The vertical lines mark $N(\phi_1)$ and $N(\phi_2^+)$. The specific parameter choices used for these models are described in §IV C.

where $\phi'_{\rm USR}$ is the analytic solution for the field velocity in USR, and ϕ' is the actual field velocity, which is evaluated numerically. By computing the field velocity deviation relative to the change at the end of the transition ϕ_2^- we guarantee that ϕ_2^+ can be defined even for small and fast transitions.

In other words, ϕ_2^- is roughly where the potential completes its transition, and ϕ_2^+ is roughly where the field velocity begins to leave the USR solution. While the specific criteria chosen here are arbitrary, they are useful for classifying transition regimes and in the $\delta_2 \to 0$ limit the choices here return the limit Eq. (34) up to percent level factors.

From these definitions for ϕ_2^+ and ϕ_2^- we compute h by evaluation of Eq. (33) and we quantify the duration of the transition from USR inflation to the beginning of the relaxation process

$$d_N \equiv N(\phi_2^-) - N(\phi_2^+). \tag{41}$$

The situation of §IV B corresponds to the limit $d_N \to 0$ and we now generalize this result by exploring the impact of the duration d_N on the resultant non-Gaussianity. In Fig. 4, we show the background parameters ϵ and η for two models with a large transition $h = 2.5$, one fast

($d_N = 5 \times 10^{-3}$) and the other slow ($d_N = 1$). For these models γ is fixed at 6×10^{-21}, while the fast transition has $\{\phi_2,\ \delta_2\} = \left\{-0.1580281699,\ 2.12 \times 10^{-10}\right\}$ and the slow transition has $\{\phi_2,\ \delta_2\} = \left\{-0.1580282187,\ 3.6 \times 10^{-8}\right\}$. Notice the amount of fine tuning in ϕ_2 required to achieve sub-percent level control of the transition amplitude and duration.

Defining $n \equiv N - N(\phi_2^-)$ as the positive increasing number of e-folds elapsed since the potential transition, in the fast transition limit $d_N \to 0$ the solution for η after the transition point behaves according to the analytic result [38]

$$\eta(n > 0) = \frac{6(h-1)}{1 + h(e^{3n} - 1)}\ . \tag{42}$$

This is reflected in the behavior of the blue solid curve in Fig. 4, which behaves as Eq. (42) up to a $\sim 0.3\%$ difference in the h parameter as defined here compared to the h parameter in the exact $\delta_2 \to 0$ limit.

On the other hand, in the case where d_N is large, the red dashed curve of Fig. 4, the numerical solution for η deviates significantly from this analytic form. This can be understood by Taylor expanding Eq. (42) around the transition point $n = 0$,

$$\eta(n > 0) = 6(h-1)\,(1 - 3hn) + \mathcal{O}\left(n^2\right), \tag{43}$$

from which we can see that after ϕ_2^-, η evolves on a timescale $n \sim h^{-1}$. Thus if the transition timescale d_N is larger than this timescale, the evolution of η will differ from the analytic solution (42).

The behavior of η is important because it comes directly into the source of squeezed non-Gaussianity in the in-in formalism, Eq. (A9), and controls the freezeout of perturbations through the evolution of ϵ in Eq. (25). Therefore, the timescale d_N plays an important role in changing the non-Gaussianity produced in USR.

The upper panel of Fig. 5 shows the power spectra for the same large-fast and large-slow models as Fig. 4. Once more, the red upper curve shows the slow transition, while the blue lower curve shows the fast transition. The vertical dashed line shows a mode which exits the horizon at the onset of the transition $N(\phi_2^+)$.

The power spectra in these models show many of the same features as the inflection-point power spectrum, Fig. 3, and therefore we focus on the USR and transition regions of the plot, where unlike the inflection model these models have power spectrum plateaus for modes which exit during USR. This plateau is modulated by small oscillations sourced by the first feature in the potential. The USR to SR transition feature in the potential also induces power spectrum oscillations, with the fast transition model having more oscillations due to the sharper source. The slow transition model has a larger power spectrum than the fast transition as ϵ reaches a lower level in this model (see Fig. 4) and thus the modes grow more.

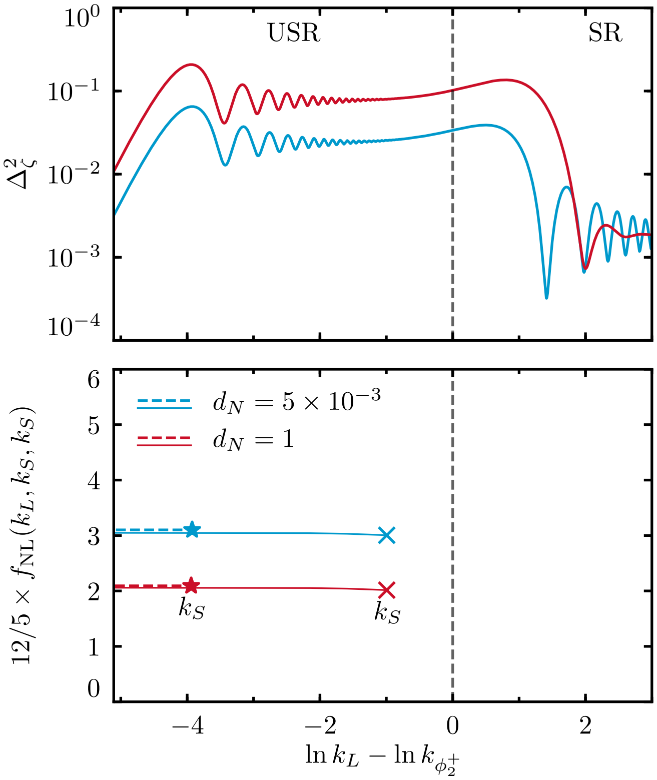

FIG. 5. Transition model power spectrum Δ^2_ζ and non-Gaussianity parameter $f_{\rm NL}(k_L, k_S, k_S)$ for the large-fast (blue) and large-slow (red) transition models of Fig. 4. The left edge corresponds roughly to a mode which exits the horizon at the beginning of USR. Conventions for displaying $f_{\rm NL}$ are the same as in Fig. 3 except both k_S values for each model cross the horizon before the end of USR (vertical line, $k_{\phi_2^+}$) (see §IV C for further discussion).

The lower panel of Fig. 5 shows the non-Gaussianity for these models for two different values of the short-wavelength mode, one marked with a cross and the other a star and both exiting during the USR phase, as a function of the long-wavelength mode, curves with correspondingly solid and dashed lines. All triangles yield the same value of $f_{\rm NL}$ when the legs exit during the USR phase up to corrections of order $k_S/k_{\phi_2^+}$, consistent with the result in the inflection model §IV A and the exact USR result of §III. However, for neither model does the level of the non-Gaussianity agree with the analytic result for USR Eq. (17), $5f_{\rm NL}/12 = 6$. This is the result of Ref. [38], that the residual level of $f_{\rm NL}$ depends on the value of the transition parameter h, and in particular the fast transition model yields the result expected from Eq. (36) for a transition with $h = 2.5$, $f_{\rm NL} \simeq 3.1$. However, the slow model has the same $h = 2.5$ as the fast model, yet a smaller value of $f_{\rm NL}$. This is due to the slow nature of the transition in the $d_N = 1$ model. Fast transitions yield Eq. (36), while slow transitions suppress the non-Gaussianity.

We model this effect with the ansatz that a given tran-

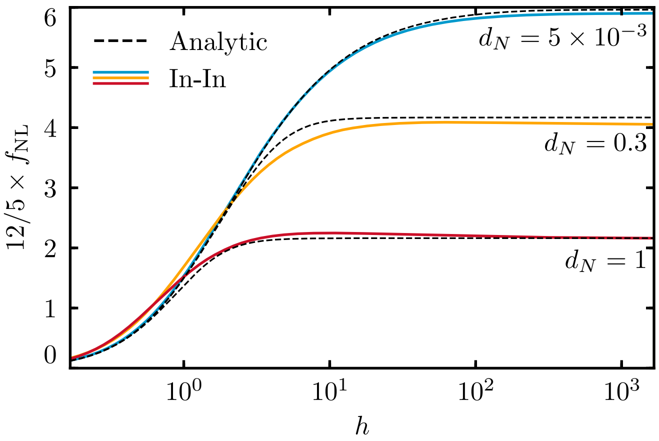

FIG. 6. Transition model non-Gaussianity parameter $f_{\rm NL}$ as a function of amplitude h for various e-fold width d_N with both k_L and k_S exiting the horizon during USR. Colored lines show the numerical in-in computation and dashed black lines show the calibrated analytic prediction from Eqs. (45) and (45). To reach the USR result $12 f_{\rm NL}/5 = 6$, the transition must be large $h \gg 1$ and fast $d_N \ll 1$ (see §IV C for further discussion).

sition length d_N sets an upper bound to the transition amplitude, independent of h. We can define an effective transition amplitude

$$h_{\rm eff} \equiv \left[\left(\frac{1.5}{d_N} \right)^{-3} + h^{-3} \right]^{-1/3} \tag{44}$$

where the exponent serves merely to interpolate between the two limits and the factor of 1.5 comes from calibrating the results to the form of (36)

$$\frac{12}{5} f_{\rm NL}(h_{\rm eff}) = 2 \frac{h_{\rm eff}(3h_{\rm eff} + \eta_V)}{(h_{\rm eff}+1)^2}, \tag{45}$$

where in this toy model with a linear SR potential we set $\eta_V = 0$.

In Fig. 6, we compare the ansatz (44) coupled with the analytic formula (45) (black dashed curves) to the full numerical in-in computation of the bispectrum (colored solid curves) for modes that exit the horizon during USR as a function of the transition amplitude h for three different transition speeds d_N. Every point along these lines corresponds to a different set of parameters for our toy model.

The analytic formula (44) agrees well with the in-in computation for all values of h and d_N. For small-slow transitions (small h and large d_N), where $f_{\rm NL}$ is small and becomes proportional to the potential slow-roll parameters on the attractor, there is a slight difference between the numerical and analytic $f_{\rm NL}$. This we attribute to small differences in the value of h as defined for the infinitely fast transitions and as defined for slow transitions since the errors decrease for smaller values of d_N. For large transitions $h \gg 1$, where $f_{\rm NL}$ is largest, the analytic ansatz produces a slight overestimate of $f_{\rm NL}$ as the transitions become faster. This is due to non-linearities between the true $h_{\rm eff}$ and $1/d_N$ in the large h limit which our ansatz does not model. Since these differences are minor, we conclude from the analytic formula that to produce a large level of non-Gaussianity after the transition from USR to SR requires $h_{\rm eff} \gg 1$ and thus the transition must be large, $h \gg 1$, and fast, $d_N \ll 1$.

Just as in the SR (§II) and exact USR case (§III), the effect on $f_{\rm NL}$ of the transition from USR to SR can be understood visually from the way phase-space trajectories intersect constant N surfaces. Fig. 7a shows the phase-space trajectories (blue lines) and constant N surfaces (red lines) for a narrow $\delta_2 = 2.12 \times 10^{-10}$ and therefore faster transitions, such that the fast model of Fig. 4 and Fig. 5 corresponds to a trajectory in this space. Fig. 7b shows the phase space for a wider $\delta_2 = 3.6 \times 10^{-8}$ and therefore slower transitions, and the slow model of Fig. 4 and Fig. 5 evolves through this space.

Trajectories near the top of each panel have the inflaton speed up after the transition and hence have a large h, with the large-small dividing line of $h = 1$ denoted by thick blue lines. Notice also that the union of the panels of Fig. 1 gives the limit of infinitely fast transitions, with the exception that here constant N surfaces are plotted relative to the transition feature $N(\phi_2)$ rather than the end of inflation $N = 0$. Trajectories are evenly spaced in ϕ at the point where they cross the bottom edge at $\phi' = -6$ in a range that reflects a reasonable amount of USR e-folds as we describe next.

Due to the smooth nature of the potential (38), for any finite $\phi > \phi_2$ the potential slope $dV/d\phi$ has a finite positive value. Thus, unlike in the exact USR case (Fig. 1b) or the infinitely fast case, all trajectories with any finite ϕ' for $\phi > \phi_2$ will eventually cross ϕ_2.

Black dashed curves in Fig. 7 depict the envelope of such trajectories, neglecting stochastic effects, and correspond to initial conditions on the attractor on the very nearly flat potential. Consequently, the constant e-fold surfaces become increasingly tightly packed and eventually space-filling, in contrast to the empty upper right triangle in Fig. 1b. We choose not to continue showing trajectories which take such large numbers of e-folds to traverse the nearly flat plateau of the potential.

By the same δN arguments of §III we can immediately see from these phase spaces why a large ($h \gg 1$) transition is necessary to conserve the USR non-Gaussianity. Here δN refers to the change in the total number of e-folds elapsed to a fixed field position on the SR side for a shift in the initial field position $\delta\phi_i$ on the USR side which then shifts the whole trajectory. Note that δN combines the change from the USR and SR sides.

Let us first consider the fast case in the top panel. Around a central trajectory with large h (upper trajectories), the crossing rate $\partial N/\partial\phi_i$ is strongly asymmetric to the sign of $\delta\phi$, i.e. there is a large second derivative $\partial^2 N/\partial\phi_i^2$ and hence a large $f_{\rm NL}$ according to Eq. (18).

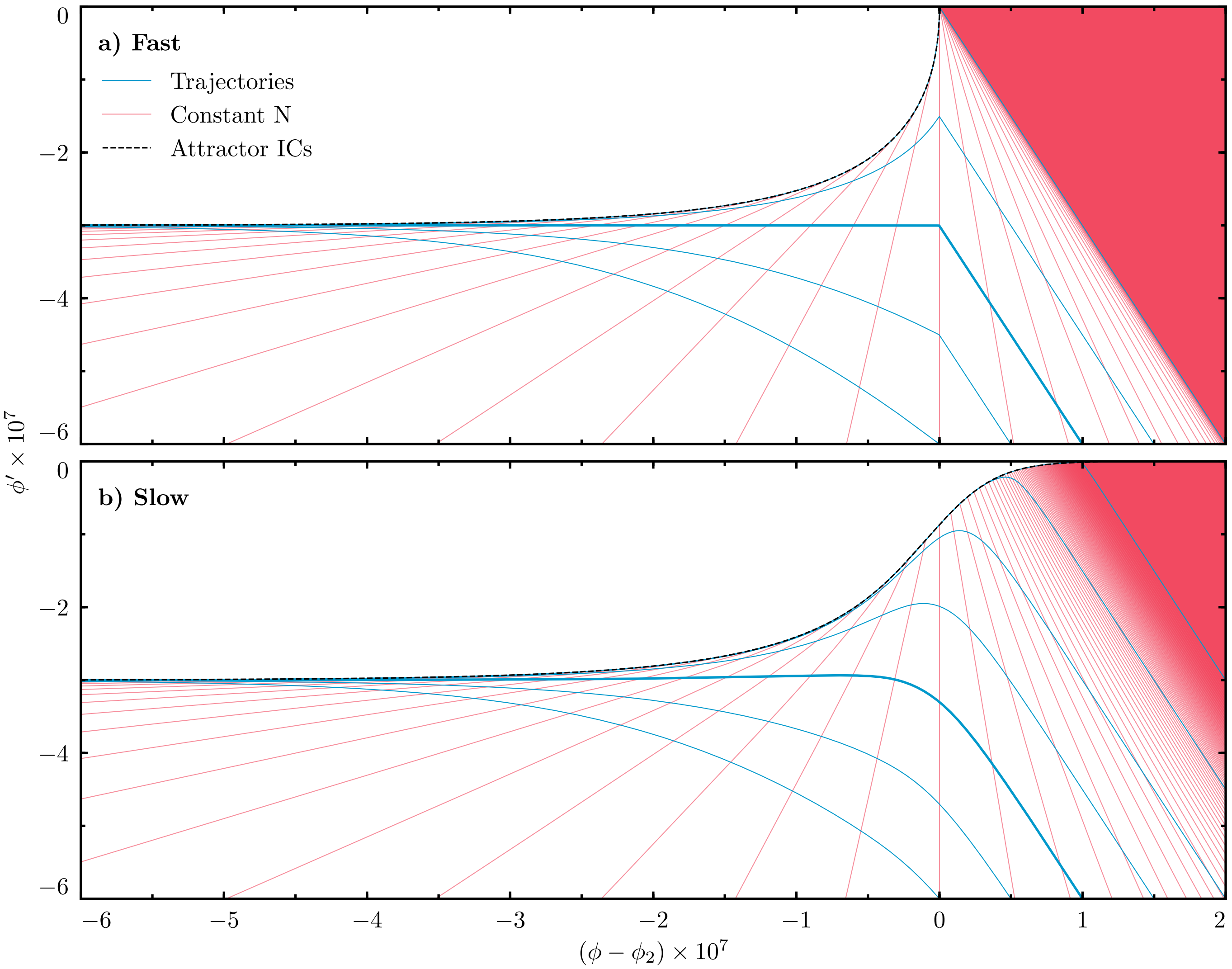

FIG. 7. Phase-space diagram for the USR to SR transition models with model trajectories (blue lines) and constant e-fold surfaces (red lines) relative to the transition at ϕ_2. The top panel a) has $\delta_2 = 2.12 \times 10^{-10}$, such that most trajectories correspond to fast transitions, while the bottom panel b) has $\delta_2 = 3.6 \times 10^{-8}$, such that most trajectories correspond to slow transitions. In each panel, higher trajectories represent larger transitions, with $h = 1$ as the large-small dividing line (thick blue). Constant N surfaces become space-filling in the top right corners. A large $f_{\rm NL}$ requires a fast-large transition as can be visualized by δN, the change in e-folds given a shift in the initial field $\delta\phi_i$ that takes the local background to a new trajectory (see §IV C for further discussion).

This is due to the much larger contribution to the rate of surfaces crossed in the USR side where the asymmetry is larger than the SR side where the asymmetry is small.

On the other hand, the asymmetry around a small h trajectory (lower trajectories) is small and therefore the non-Gaussianity is small. This is due to the smaller contribution to $\partial N/\partial\phi_i$ on the USR side relative to the SR side. In other words the power spectrum continues to grow on the SR side at small h which suppresses the non-Gaussianity from the USR side.

By comparing the fast and slow cases, we can visually see that the transition duration sets an effective maximum transition amplitude $h_{\rm eff}$. Above a certain value of h, the trajectory joins the slow-roll attractor before the transition and therefore will have a highly suppressed non-Gaussianity comparable to the small-slow transition of §IV A.

In Fig. 8 we formalize these heuristic arguments by computing the δN formula (18) for different trajectories in these phase spaces. We organize the trajectories by their velocity at ϕ_2, and thus the horizontal axis of Fig. 8 corresponds to the vertical axis of the Fig. 7 at the transition point. We then also compare the δN result to our analytic expectation for $f_{\rm NL}$, Eqs. (44) and (45).

The δN computation cross-validates our analytic formula which was calibrated to the in-in calculations, showing excellent agreement for all methods of computation across the fast-slow and large-small transition space. The

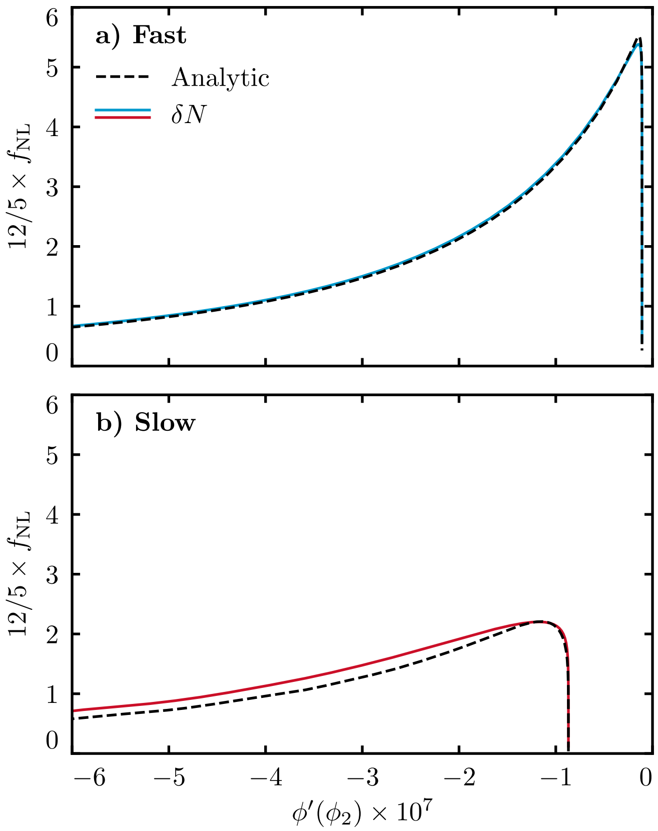

FIG. 8. Non-Gaussianity for trajectories in the corresponding phase spaces of Fig. 7 computed via the δN formula (18) compared to our calibrated analytic result Eqs. (44) and (45). The horizontal axis here corresponds to the vertical axis of Fig. 7. Only for fast and large transitions does the transition model reproduce the USR result, and for fixed δ_2 there is a maximum value of $f_{\rm NL}$ attainable.

δN computation thus also confirms that a large $f_{\rm NL}$ requires a large-fast transition.

For large $\phi'(\phi_2)$, after the peak non-Gaussianity, there is a sharp cliff in Fig. 8 beyond which $f_{\rm NL}$ becomes suppressed. This cliff corresponds to trajectories which reach the attractor before the transition, thus inhabiting the space-filling regions of Fig. 7, and the edge of the cliff is given by the black dashed envelope of Fig. 7. These trajectories are slow transitions even though they have the same narrow field space width δ_2 and appear in the 'fast' panel. These cases behave in the same way as those in the 'slow' panel once the duration of the transition in e-folds d_N is accounted for.

Of course even for a large and fast transition, for which $5/12\, f_{\rm NL} \to 6$, the response of the small scale power spectrum to the long-wavelength mode is still dependent on the value of the long-wavelength mode, and in particular the argument of §III still holds, that

$$\frac{\Delta P_\zeta}{P_\zeta} \lesssim 1, \tag{46}$$

unless the model already produces PBHs with a Gaussian distribution.

V. CONCLUSION

Canonical slow-roll inflation cannot produce primordial black holes in a large enough quantity to be the dark matter. While perturbations do exhibit a small level of local non-Gaussianity which couples short-wavelength PBH fluctuations to the long-wavelength modes they live in and can in principle enhance local abundances at peaks of long-wavelength modes, transforming to a freely-falling coordinate system shows that locally measured PBH abundances are completely insensitive to this non-Gaussianity because it is generated by the reverse coordinate transformation to begin with.

Therefore any confirmation that the dark matter is in the form of PBHs would rule out canonical slow-roll inflation. The only way to rescue canonical inflation would be to violate slow roll, and a phase of ultra-slow-roll inflation after CMB scales exit the horizon is the natural way to do this. We showed by gauge transformation and by the δN formalism, which can be illustrated graphically and contrasted to the SR case, that non-Gaussianities are large in USR phase when perturbations freeze out instantly at some fixed field position.

The same coordinate transformation machinery as in slow roll confirms that USR non-Gaussianities can locally enhance PBH abundances. However the effect is very mild, giving at most an order unity enhancement of the local power spectrum. Such enhancements can only make models that are already on the border line of succeeding to produce PBHs as the dark matter under Gaussian assumptions actually succeed. For such cases, generally a small change in parameters that prolong the USR phase would equally well produce PBHs under Gaussian assumptions.

Even more importantly, USR phase has to end in some way. Ref. [38] established that the non-Gaussianity is very sensitive to how this period ends using cases where the transition is infinitely fast. By exact computation in the in-in formalism and validation with the δN formalism, we mapped the entire range of possible endings to USR to show that only a small class of transitions conserve the large USR non-Gaussianity through the transition to slow roll. These are the transitions which are fast, in that the potential exhibits a sharp feature that is traversed by the inflaton in much less than an e-fold, and large, in that the inflaton needs to gain significant velocity after transiting the feature. All other types of transitions suppress the non-Gaussianity significantly.

Producing primordial black holes as dark matter in canonical single-field inflation requires a complicated and fine-tuned potential shape with a transient violation of slow-roll, a conclusion which is not weakened by the inclusion of local non-Gaussian effects.

Note added: While this work was being completed, Ref. [53] appeared which explores the equilateral non-Gaussianity in transitions which are slow and small in our terminology, similar to the inflection model of §IV A. Unlike that work, we focus on whether squeezed non-Gaussianity can make a model that fails to form sufficient PBHs for the dark matter under the Gaussian assumption into one that does rather than how non-Gaussianity changes the abundance of a rare tracer population of PBHs.

ACKNOWLEDGMENTS

We thank Macarena Lagos, Meng-Xiang Lin, and Marco Raveri for fruitful discussions. SP and WH were supported by U.S. Dept. of Energy contract DE-FG02-13ER41958, NASA ATP NNX15AK22G, and the Simons Foundation. SP was additionally supported by the Kavli Institute for Cosmological Physics at the University of Chicago through grant NSF PHY-1125897 and an endowment from the Kavli Foundation and its founder Fred Kavli. HM was supported by Japan Society for the Promotion of Science (JSPS) Grants-in-Aid for Scientific Research (KAKENHI) Nos. JP17H06359 and JP18K13565.

Appendix A: Bispectrum Computation

1. Numerical Methods

We describe in this Appendix the numerical computation of the background evolution, the modefunctions, and the cubic interactions during inflation in order to study the precise predictions of inflationary scenarios which transition between SR and USR phases.

After numerically solving for the background evolution using the canonical Klein-Gordon and Friedmann equations, we use the quadratic action for the comoving curvature perturbation ζ

$$S_2 = \int d^4x \; \mathcal{L}_2, \tag{A1}$$

with the quadratic Lagrangian density

$$\mathcal{L}_2 \equiv a^3\epsilon \left[\dot\zeta^2 - \frac{1}{a^2}\left(\partial\zeta\right)^2\right], \tag{A2}$$

to determine the evolution of the modefunctions through the Mukhanov-Sasaki equation of motion

$$\frac{1}{a^2\epsilon}\frac{d}{ds}\left(a^2\epsilon\frac{d\zeta_k}{ds}\right) + k^2\zeta_k = 0, \tag{A3}$$

where $s \equiv \int_t^{t_{\rm end}} dt/a$, with Bunch-Davies initial conditions at $ks \gg 1$ of the form

$$\zeta_k^0 = \frac{1}{2a\sqrt{k\epsilon}}\left(1 + \frac{i}{ks}\right)e^{iks}. \tag{A4}$$

From these modefunctions we can construct the Fourier space interaction-picture field operators

$$\hat\zeta^I_{\mathbf{k}} = \zeta_k \hat a(\mathbf{k}) + \zeta_k^* \hat a^\dagger(-\mathbf{k}), \tag{A5}$$

where the creation and annihilation operators satisfy the usual commutation relation

$$[\hat a(\mathbf{k}), \hat a^\dagger(\mathbf{k}')] = (2\pi)^3\delta(\mathbf{k}-\mathbf{k}'). \tag{A6}$$

The power spectrum can be evaluated from the modefunctions at a time t_* taken to be after all the relevant modes have frozen out (e.g., after the end of USR phase) as

$$\begin{aligned}\langle\hat\zeta^I_{\mathbf{k}}(t_*)\hat\zeta^I_{\mathbf{k}'}(t_*)\rangle &= (2\pi)^3\delta^3(\mathbf{k}+\mathbf{k}')|\zeta_k(t_*)|^2 \\ &\equiv (2\pi)^3\delta^3(\mathbf{k}+\mathbf{k}')P_\zeta(k).\end{aligned} \tag{A7}$$

The tree-level three-point correlation function is then computed in the in-in formalism as [37, 54–56]

$$\begin{aligned}&\langle\hat\zeta_{\mathbf{k_1}}\hat\zeta_{\mathbf{k_2}}\hat\zeta_{\mathbf{k_3}}\rangle \\ &\simeq 2\,\mathrm{Re}\left[-i\int_{-\infty(1+i\epsilon)}^{t_*} dt\langle\hat\zeta^I_{\mathbf{k_1}}(t_*)\hat\zeta^I_{\mathbf{k_2}}(t_*)\hat\zeta^I_{\mathbf{k_3}}(t_*)H_I(t)\rangle\right],\end{aligned} \tag{A8}$$

where $\hat\zeta$ is the full field operator and in which the interaction Hamiltonian H_I can be calculated at cubic order from the cubic Lagrangian $\mathcal{L}_3$ by $H_I \simeq -\int d^3x\mathcal{L}_3$ [57]. $\mathcal{L}_3$ itself is given by [54, 58]

$$\begin{aligned}\mathcal{L}_3 \equiv\;& a^3\epsilon\frac{d}{dt}\left(\epsilon+\frac{\eta}{2}\right)\zeta^2\dot\zeta - \frac{d}{dt}\left[a^3\epsilon\left(\epsilon+\frac{\eta}{2}\right)\zeta^2\dot\zeta\right] \\ &+ \epsilon\zeta(\mathcal{H}_2 + 2\mathcal{L}_2) \\ &- \frac{d}{dt}\left[\frac{a^3\epsilon}{H}\zeta\dot\zeta^2 + a^3\frac{\epsilon^2}{2H}\dot\zeta\partial_a\zeta\partial_a(\partial^{-2}\dot\zeta)\right],\end{aligned} \tag{A9}$$

where $\mathcal{H}_2$ is the quadratic Hamiltonian density

$$\mathcal{H}_2 = a^3\epsilon\left[\dot\zeta^2 + \frac{1}{a^2}\left(\partial\zeta\right)^2\right]. \tag{A10}$$

We have neglected here operators which do not contribute to the squeezed limit and for a canonical scalar (cf. [58]). We have also neglected here boundary operators whose contribution at t_* is suppressed by relative factors of k/aH. Similarly, terms in the third line of (A9) are suppressed by the extra $\dot\zeta$ factor so long as t_* is taken to be after all modes have frozen out, i.e. after the transition, in the context of transient USR. We take this approach in the main paper. In App. A 2, we consider this extra boundary contribution if the correlator is evaluated during USR.

From translational and rotational invariance the three-point correlator is related to the bispectrum B_ζ through

$$\langle\hat\zeta_{\mathbf{k_1}}\hat\zeta_{\mathbf{k_2}}\hat\zeta_{\mathbf{k_3}}\rangle = (2\pi)^3\delta^3(\mathbf{k_1}+\mathbf{k_2}+\mathbf{k_3})B_\zeta(k_1,k_2,k_3), \tag{A11}$$

from which is constructed the conventional non-Gaussianity parameter, Eq. (7).

In this work we solve these formulae numerically to compute the non-Gaussian response of perturbations in generic inflationary scenarios. We use these numerical results to calibrate an analytic formula for models with smooth USR to SR transitions in §IV C. In addition, we exploit the δN formalism, which requires only numerical solutions for various background evolutions, to validate and visualize our results. In the circumstance of pure USR inflation an analytic approach can be followed which we consider next.

2. USR Gauge vs. Field Redefinition

Under some conditions the bispectrum can be calculated much more simply than the numerical approach presented in App. A 1. In this Appendix, we compute the bispectrum in USR (§III) analytically. This can be done by transforming between spatially flat and comoving gauges since the flat potential implies negligible field interactions in the former. We explicitly show this to be the case below and clarify subtleties due to the shift in evaluation time induced by the gauge transformation on cubic interactions from the boundary terms. This gauge transformation is sometimes phrased as a field redefinition in the literature [37], since the dependence on the evaluation time drops out in ordinary slow roll where the curvature is frozen outside the horizon. Even in USR the three point correlations can still can be deduced directly from the two point correlations from this perspective (see., e.g., Refs. [43–45, 59]) but our approach clarifies the role of the cubic interaction terms involved in their direct computation.

We start the computation in spatially flat-gauge, where the bispectrum computation is trivial, and then transform to comoving gauge, which is the relevant gauge for observations. The relationship between the time coordinate in these gauges is

$$t_{\rm SF} = t_{\rm C} + \frac{\zeta}{H} \tag{A12}$$

where $t_{\rm SF}$ denotes the time coordinate in spatially flat gauge, $t_{\rm U}$ the time coordinate in comoving gauge, and ζ is the comoving-gauge curvature perturbation.[*3] Letting ζ_N be the rescaled field perturbation in spatially-flat gauge, $\zeta_N \equiv -H\delta\phi/\dot\phi$, the relationship between the perturbations is

$$\zeta = \zeta_N + f(\zeta_N) \tag{A13}$$

where

$$f(\zeta_N) = \frac{\eta}{4}\zeta_N^2 + \frac{1}{H}\zeta_N\dot\zeta_N \tag{A14}$$

to the lowest order in ζ_N and up to terms suppressed outside the horizon by factors of $k/(aH)$. Eq. (A13) can be viewed equivalently as simply a field redefinition.

In spatially flat gauge, all interactions should be suppressed by ϵ, which is extremely small in USR. We therefore assume for now that the spatially-flat gauge fields are free fields,

$$\langle \zeta_N\zeta_N\zeta_N\rangle = 0. \tag{A15}$$

We can use Eq. (A13) to find that in USR

$$\begin{aligned}\langle \zeta_{k_1}\zeta_{k_2}\zeta_{k_3}\rangle = {}& (2\pi)^3\delta^3(k_1+k_2+k_3)\\ &\times \frac{3H^4(k_1^3+k_2^3+k_3^3)}{16\epsilon_{\rm end}^2 k_1^3k_2^3k_3^3},\end{aligned} \tag{A16}$$

where we have evaluated correlators of the form $\langle\zeta\zeta\rangle$ and $\langle\zeta\dot\zeta\rangle$ using (A4) with the USR scalings $\epsilon = \epsilon_*(a_*/a)^6$ and $s \simeq (aH)^{-1}$ which yield (e.g. [43])

$$\zeta_k = -\frac{iH}{2\sqrt{\epsilon_* k^3}}\left(\frac{a}{a_*}\right)^3(1-iks)\,e^{iks}, \tag{A17}$$

and the subscript $*$ here indicates the values of the parameters at an arbitrary reference time during USR phase. Note that choosing $a_* = a_{\rm end}$, the end of the USR phase, leads to Eq. (16) for the power spectrum as well.

This result for the bispectrum leads to the well-known result that the squeezed bispectrum in USR inflation violates the consistency relation prediction of zero. This result is often computed from a field redefinition of the form of Eq. (A13). However, this can be confusing as the basis of this derivation, $\langle\zeta_N\zeta_N\zeta_N\rangle = 0$, is a statement about perturbations of a free field in spatially flat gauge rather than a statement about interactions of the redefined field in comoving gauge. To see this, we can take opposite approach and start with the action in comoving gauge, transform it to spatially flat gauge, and show how and why $\langle\zeta_N\zeta_N\zeta_N\rangle = 0$.

Neglecting all terms suppressed by factors of ϵ, factors of $k/(aH)$, or irrelevant in the squeezed limit, the action for the comoving curvature perturbation ζ in USR is given by

$$\begin{aligned}S[\zeta] = {}& \int_{\mathcal{M}_{\rm C}} d^4x\; a^3\epsilon\left[\dot\zeta^2 - \frac{1}{a^2}(\partial\zeta)^2\right]\\ &+ \int_{\partial\mathcal{M}_{\rm C}} d^3x\; a^3\epsilon\left(-\frac{\eta}{2}\zeta^2\dot\zeta - \frac{1}{H}\zeta\dot\zeta^2\right)\end{aligned} \tag{A18}$$

where the first line is the quadratic action and the second line is the cubic action, $\mathcal{M}_{\rm C}$ denotes the bulk of the spacetime $\vec{x} \in (-\vec\infty, +\vec\infty)$ and $t \in (0, t^{\rm C}_{\rm end})$, and $\partial\mathcal{M}_{\rm C}$ denotes the temporal boundary of $\mathcal{M}_{\rm C}$ at $t^{\rm C}_{\rm end}$. The cubic

[*3] In USR, ζ grows and so does the difference between the time coordinate of spatially flat gauge and comoving gauge. Hence even though a growing curvature violates the separate universe criteria for comoving slicing [51], spatially flat slicing observers are nearly freely-falling and see superhorizon field perturbations as a local FLRW background.

action here can be obtained from Eq. (A9) by eliminating $\dot\eta$ and ϵ-suppressed terms, which are both driven to zero in USR.

Note that plugging the cubic portion of this action into the in-in formula Eq. (A8) with the modefunctions above yields the same result as Eq. (A16).

Let us transform the action (A18) in two ways. Transforming the action (A18) using Eq. (A12) as a field redefinition alone gives

$$S[\zeta_N] = \int_{\mathcal{M}_\mathrm{C}} d^4x\ \mathcal{L}_2[\zeta_N] + \int_{\partial\mathcal{M}_\mathrm{C}} d^3x\ (2a^3\epsilon\dot\zeta_N f) + \int_{\partial\mathcal{M}_\mathrm{C}} d^3x\ a^3\epsilon\left(-\frac{\eta}{2}\zeta_N^2\dot\zeta_N - \frac{1}{H}\zeta_N\dot\zeta_N^2\right), \tag{A19}$$

where we have integrated by parts and used the equation of motion from the quadratic action $\mathcal{L}_2[\zeta_N]$ defined in (A2).

After substituting Eq. (A14) for the f term associated with the nonlinear field redefinition, we obtain

$$S[\zeta_N] = \int_{\mathcal{M}_\mathrm{C}} d^4x\ \mathcal{L}_2[\zeta_N] + \int_{\partial\mathcal{M}_\mathrm{C}} d^3x\ \frac{a^3\epsilon}{H}\zeta_N\dot\zeta_N^2, \tag{A20}$$

where we see that a cubic boundary interaction remains in the action. Unlike in the ordinary SR case, it contributes significantly in the USR case since $\zeta_N \propto a^3$ outside the horizon.

Therefore, if we had used Eq. (A20) for the cubic interactions in the in-in formula, we would have computed

$$\langle\zeta_N\zeta_N\zeta_N\rangle \neq 0, \tag{A21}$$

which is inconsistent with what we expected.

In contrast, if we interpret Eq. (A12) as a gauge transformation, then we also have to apply the transformation of the temporal boundary $\partial\mathcal{M}_\mathrm{C} = \partial\mathcal{M}_\mathrm{SF} + \mathcal{O}(\zeta_N)$ to the action (A20). For the intrinsically cubic terms, this transformation yields a higher order correction, but for the quadratic term it produces a cubic boundary interaction

$$\int_{\mathcal{M}_\mathrm{C}} d^4x\ \mathcal{L}_2[\zeta_N] = \int_{\mathcal{M}_\mathrm{SF}} d^4x\ \mathcal{L}_2[\zeta_N] - \int_{\partial\mathcal{M}_\mathrm{SF}} d^3x\ \frac{\zeta_N}{H}\mathcal{L}_2[\zeta_N]\,, \tag{A22}$$

which precisely cancels the remaining cubic order term after ignoring higher order terms of ϵ and $k/(aH)$. We thus end up with the action of free field form

$$S[\zeta_N] = \int_{\mathcal{M}_\mathrm{SF}} d^4x\ a^3\epsilon\left[\dot\zeta_N^2 - \frac{1}{a^2}\left(\partial\zeta_N\right)^2\right], \tag{A23}$$

up to the cubic order of ζ. Since there is no cubic interaction for ζ_N, we have

$$\langle\zeta_N\zeta_N\zeta_N\rangle = 0. \tag{A24}$$

This is consistent with the intuition that a flat potential for the scalar field produces no interaction terms.

Therefore, the transformation (A12) is better viewed as a gauge transformation than a field redefinition. In practice, realistic inflationary models have an end of the USR stage. We therefore can always choose to evaluate the bispectrum after all relevant modes have frozen out and in this case the $\zeta\dot\zeta^2$ boundary term and the subtleties about evaluation times become irrelevant. This is the approach which is described in App. A 1.

[1] Y. B. Zel'dovich and I. D. Novikov, Sov. Astron. **10**, 602 (1967).

[2] B. J. Carr and S. W. Hawking, Mon. Not. Roy. Astron. Soc. **168**, 399 (1974).

[3] B. J. Carr, Astrophys. J. **201**, 1 (1975).

[4] P. Meszaros, Astron. Astrophys. **37**, 225 (1974).

[5] G. F. Chapline, Nature (London) **253**, 251 (1975).

[6] B. Carr, F. Kuhnel, and M. Sandstad, Phys. Rev. **D94**, 083504 (2016), arXiv:1607.06077 [astro-ph.CO].

[7] H. Niikura *et al.*, (2017), arXiv:1701.02151 [astro-ph.CO].

[8] B. Carr, M. Raidal, T. Tenkanen, V. Vaskonen, and H. Veerme, Phys. Rev. **D96**, 023514 (2017), arXiv:1705.05567 [astro-ph.CO].

[9] F. Khnel and K. Freese, Phys. Rev. **D95**, 083508 (2017), arXiv:1701.07223 [astro-ph.CO].

[10] F. Capela, M. Pshirkov, and P. Tinyakov, Phys. Rev. **D87**, 123524 (2013), arXiv:1301.4984 [astro-ph.CO].

[11] R. R. Lane, L. L. Kiss, G. F. Lewis, R. A. Ibata, A. Siebert, T. R. Bedding, and P. Székely, Mon. Not. Roy. Astron. Soc. **400**, 917 (2009), arXiv:0908.0770 [astro-ph.GA].

[12] S. Bird, I. Cholis, J. B. Muoz, Y. Ali-Hamoud, M. Kamionkowski, E. D. Kovetz, A. Raccanelli, and A. G. Riess, Phys. Rev. Lett. **116**, 201301 (2016), arXiv:1603.00464 [astro-ph.CO].

[13] M. Sasaki, T. Suyama, T. Tanaka, and S. Yokoyama, Phys. Rev. Lett. **117**, 061101 (2016), [Erratum: Phys. Rev. Lett. **121**, 059901 (2018)], arXiv:1603.08338 [astro-ph.CO].

[14] M. Zumalacárregui and U. Seljak, Phys. Rev. Lett. **121**, 141101 (2018), arXiv:1712.02240 [astro-ph.CO].

[15] M. Sasaki, T. Suyama, T. Tanaka, and S. Yokoyama, Class. Quant. Grav. **35**, 063001 (2018), arXiv:1801.05235 [astro-ph.CO].

[16] LIGO Scientific and Virgo Collaborations (2018), arXiv:1811.12907 [astro-ph.HE].

[17] H. Motohashi and W. Hu, Phys. Rev. **D96**, 063503 (2017), arXiv:1706.06784 [astro-ph.CO].

[18] G. Ballesteros and M. Taoso, Phys. Rev. **D97**, 023501 (2018), arXiv:1709.05565 [hep-ph].

[19] M. Kawasaki, A. Kusenko, Y. Tada, and T. T. Yanagida, Phys. Rev. **D94**, 083523 (2016), arXiv:1606.07631 [astro-ph.CO].

[20] J. García-Bellido and E. Ruiz Morales, Phys. Dark Univ. **18**, 47 (2017), arXiv:1702.03901 [astro-ph.CO].

[21] A. Y. Kamenshchik, A. Tronconi, T. Vardanyan, and G. Venturi, (2018), arXiv:1812.02547 [gr-qc].

[22] T. Suyama and J. Yokoyama, Phys. Rev. **D84**, 083511 (2011), arXiv:1106.5983 [astro-ph.CO].

[23] M. Kawasaki, N. Kitajima, and T. T. Yanagida, Phys. Rev. **D87**, 063519 (2013), arXiv:1207.2550 [hep-ph].

[24] J.-W. Chen, J. Liu, H.-L. Xu, and Y.-F. Cai, Phys. Lett. **B769**, 561 (2017), arXiv:1609.02571 [gr-qc].

[25] S. Young and C. T. Byrnes, JCAP **1308**, 052 (2013), arXiv:1307.4995 [astro-ph.CO].

[26] S. Young, C. T. Byrnes, and M. Sasaki, JCAP **1407**, 045 (2014), arXiv:1405.7023 [gr-qc].

[27] G. Franciolini, A. Kehagias, S. Matarrese, and A. Riotto, JCAP **1803**, 016 (2018), arXiv:1801.09415 [astro-ph.CO].

[28] S. Young and C. T. Byrnes, JCAP **1504**, 034 (2015), arXiv:1503.01505 [astro-ph.CO].

[29] Y. Tada and S. Yokoyama, Phys. Rev. **D91**, 123534 (2015), arXiv:1502.01124 [astro-ph.CO].

[30] W. H. Kinney, Phys. Rev. **D72**, 023515 (2005), arXiv:gr-qc/0503017 [gr-qc].

[31] A. A. Starobinsky, JETP Lett. **42**, 152 (1985), Pis'ma v ZhETF. 42, 124 (1985).

[32] D. S. Salopek and J. R. Bond, Phys. Rev. **D42**, 3936 (1990).

[33] M. Sasaki and E. D. Stewart, Prog. Theor. Phys. **95**, 71 (1996), arXiv:astro-ph/9507001 [astro-ph].

[34] N. S. Sugiyama, E. Komatsu, and T. Futamase, Phys. Rev. **D87**, 023530 (2013), arXiv:1208.1073 [gr-qc].

[35] G. Domenech, J.-O. Gong, and M. Sasaki, Phys. Lett. **B769**, 413 (2017), arXiv:1606.03343 [astro-ph.CO].

[36] A. A. Abolhasani and M. Sasaki, JCAP **1808**, 025 (2018), arXiv:1805.11298 [astro-ph.CO].

[37] J. M. Maldacena, JHEP **05**, 013 (2003), arXiv:astro-ph/0210603 [astro-ph].

[38] Y.-F. Cai, X. Chen, M. H. Namjoo, M. Sasaki, D.-G. Wang, and Z. Wang, JCAP **1805**, 012 (2018), arXiv:1712.09998 [astro-ph.CO].

[39] F. K. Manasse and C. W. Misner, J. Math. Phys. **4**, 735 (1963).

[40] L. Senatore and M. Zaldarriaga, JHEP **09**, 148 (2013), arXiv:1210.6048 [hep-th].

[41] L. Senatore and M. Zaldarriaga, JCAP **1208**, 001 (2012), arXiv:1203.6884 [astro-ph.CO].

[42] E. Pajer, F. Schmidt, and M. Zaldarriaga, Phys. Rev. **D88**, 083502 (2013), arXiv:1305.0824 [astro-ph.CO].

[43] M. H. Namjoo, H. Firouzjahi, and M. Sasaki, EPL **101**, 39001 (2013), arXiv:1210.3692 [astro-ph.CO].

[44] J. Martin, H. Motohashi, and T. Suyama, Phys. Rev. **D87**, 023514 (2013), arXiv:1211.0083 [astro-ph.CO].

[45] X. Chen, H. Firouzjahi, E. Komatsu, M. H. Namjoo, and M. Sasaki, JCAP **1312**, 039 (2013), arXiv:1308.5341 [astro-ph.CO].

[46] C. Pattison, V. Vennin, H. Assadullahi, and D. Wands, JCAP **1710**, 046 (2017), arXiv:1707.00537 [hep-th].

[47] G. Cabass, E. Pajer, and F. Schmidt, JCAP **1701**, 003 (2017), arXiv:1612.00033 [hep-th].

[48] N. Aghanim *et al.* (Planck), (2018), arXiv:1807.06209 [astro-ph.CO].

[49] P. A. R. Ade *et al.* (Planck), Astron. Astrophys. **594**, A17 (2016), arXiv:1502.01592 [astro-ph.CO].

[50] C. Pattison, V. Vennin, H. Assadullahi, and D. Wands, (2018), arXiv:1806.09553 [astro-ph.CO].

[51] W. Hu and A. Joyce, Phys. Rev. **D95**, 043529 (2017), arXiv:1612.02454 [astro-ph.CO].

[52] M. Cicoli, V. A. Diaz, and F. G. Pedro, JCAP **1806**, 034 (2018), arXiv:1803.02837 [hep-th].

[53] V. Atal and C. Germani, (2018), arXiv:1811.07857 [astro-ph.CO].

[54] P. Adshead, W. Hu, and V. Miranda, Phys. Rev. **D88**, 023507 (2013), arXiv:1303.7004 [astro-ph.CO].

[55] S. Weinberg, Phys. Rev. **D72**, 043514 (2005), arXiv:hep-th/0506236 [hep-th].

[56] P. Adshead, R. Easther, and E. A. Lim, Phys. Rev. **D80**, 083521 (2009), arXiv:0904.4207 [hep-th].

[57] P. Adshead, R. Easther, and E. A. Lim, Phys. Rev. **D79**, 063504 (2009), arXiv:0809.4008 [hep-th].

[58] S. Passaglia and W. Hu, Phys. Rev. **D98**, 023526 (2018), arXiv:1804.07741 [astro-ph.CO].

[59] X. Chen, H. Firouzjahi, M. H. Namjoo, and M. Sasaki, EPL **102**, 59001 (2013), arXiv:1301.5699 [hep-th].