Inflation in Motion:
Unitarity Constraints in Effective Field Theories with
Broken Lorentz Symmetry

Tanguy Grall\textsuperscript{a} and Scott Melville\textsuperscript{a,b}
\textsuperscript{a}Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Wilberforce Road, Cambridge CB3 0WA, U.K.
\textsuperscript{b}Emmanuel College, University of Cambridge, St Andrews Street, Cambridge CB2 3AP, U.K.
E-mail: tg418@cam.ac.uk, scott.melville@damtp.cam.ac.uk

Abstract: During inflation, there is a preferred reference frame in which the expansion of the background spacetime is spatially isotropic. In contrast to Minkowski spacetime, observables can depend on the velocity of the system with respect to this cosmic rest frame. We derive new constraints from radiative stability and unitarity on effective field theories with such spontaneously broken Lorentz symmetry. In addition to a maximum energy scale, there is now also a critical velocity at which the theory breaks down. The theory therefore has different resolving power in time and in space, and we show that these can only coincide if cubic Lorentz-violating interactions are absent. Applying these bounds to the Effective Field Theory of Inflation, we identify the region of parameter space in which inflation can be both single-field and weakly coupled on subhorizon scales. This can be implemented as a theoretical prior, and we illustrate this explicitly using Planck observational constraints on the primordial bispectrum.
1 Introduction

Unitarity is a fundamental pillar of quantum field theory. The requirement that probability amplitudes correctly normalise is essential if we are to make sense of our theoretical computations. Studying the restrictions that this places on a physical theory has proven invaluable in guiding past efforts to construct viable models of the unknown—for example constraining the pion cross section before the development of QCD [1, 2], and bounding the Higgs mass before the LHC [3]. In cosmology, we are faced with the challenge of modelling the earliest moments of our Universe. Can we similarly leverage unitarity to guide our efforts, and improve our understanding of the physics responsible for inflation?

The Effective Field Theory (EFT) approach to inflation [4] provides a model-independent framework with which to analyse the non-Gaussianities produced during inflation. Rather
than specify the matter content (e.g. a particular scalar potential) responsible for an FLRW expansion in the early Universe, instead it describes fluctuations about this background from the point of view of the spontaneously broken Lorentz symmetry—this allows the EFT of Inflation to capture a wide variety of different theories which all share the same background and linear cosmology, without committing to a particular underlying (UV complete) model. However, this also means that existing (Lorentz-invariant) techniques from particle physics, including the usual implementation of unitarity, cannot be applied directly to the EFT of Inflation. As a step towards implementing unitarity in the EFT of Inflation, in this work we study the scattering of sufficiently subhorizon modes (whose propagation is not affected by the expanding spacetime, but whose interactions need not be Lorentz-invariant), and are able to identify the region of parameter space in which this scattering is unitary.

Our main result is the construction of unitarity bounds for EFTs with broken boosts, using a new partial wave expansion for $2 \to 2$ scattering which accounts for a non-zero centre-of-mass velocity. We also show how simple scaling arguments, from the power counting of loops and the optical theorem for $n \to n$ scattering, can be extended to include the effects of broken boosts. This equips us with the necessary tools to analyse the scattering of subhorizon modes during inflation, and we identify the region of parameter space in which subhorizon physics can be approximately single-field (and weakly coupled), which can be compared with Planck’s observational constraints on primordial non-Gaussianity.

**Inflationary Correlators**

Metric fluctuations about an FLRW background, $ds^2 = -dt^2 + a^2(t)d\mathbf{x}^2/c_T^2$ (where $c_T$ is the tensor sound speed), can be described at leading order in derivatives by the EFT [4],

$$S_{\text{LO}}[\delta g] = \int d^4x \sqrt{-g} \left\{ \frac{M_P^2}{2} (R - 6H^2) + M_P^2 \dot{H} (-2 + \delta g^{00}) + \sum_{n=2}^{\infty} \frac{M_n^4}{n!} (\delta g^{00})^n \right\}, \quad (1.1)$$

where the scales $M_0$ and $M_1$ have been fixed so that the background is stable. Since temporal diffeomorphisms have been spontaneously broken, $\delta g^{\mu\nu}$ now propagates a scalar mode, $\pi$ (in addition to the usual tensor modes of General Relativity). The decay constant $f_\pi$ associated with this symmetry breaking is set by $M_1$, namely $f_\pi^2 = 2c_s M_P^2 |\dot{H}|$, while the $\pi$ sound speed is set by $M_2$, namely $M_2^4 = \frac{1}{3} f_\pi^4 (1 - c_s^2)/c_s^3$, where $c_s = c_\pi/c_T$ is the ratio of the scalar sound speed to the tensor sound speed. $\pi$ is related to the usual scalar curvature perturbations, $\zeta = -H \pi/f_\pi^2$, whose power spectrum and higher order correlations seed the structure observed in the CMB.

Although $\pi$ mixes non-trivially with $\delta g^{00}$, if we focus on scalar modes with sufficiently

---

1 Rather than work in units in which $c_T = 1$, we will keep factors of $c_T$ explicit so that dimensional analysis can be carried out in time and space separately.
large energies they decouple from the tensors,

\[ \frac{\omega}{f_\pi} \gg \frac{f_\pi}{M_P} \quad \text{(equivalently, } M_P \to \infty \text{ with } f_\pi \text{ fixed}). \quad (1.2) \]

In this limit, the leading order interactions for the scalar perturbations produced during inflation can be written succinctly in terms of \( Z_{\pi\pi} = Z^{\mu\nu} \partial_{\mu\pi} \partial_{\nu\pi}, \)

\[ S_{\text{LO}}[\pi] = \int d^4 x \sqrt{-g} \left\{ -\frac{1}{2} Z_{\pi\pi} + \frac{\alpha_1}{f_\pi} \pi^3 - \frac{\alpha_2}{f_\pi} \pi Z_{\pi\pi} + \frac{\beta_1}{f_\pi} \pi^4 - \frac{\beta_2}{f_\pi} \pi^2 Z_{\pi\pi} + \frac{\beta_3}{f_\pi} Z_{\pi\pi}^2 \right\} \quad (1.3) \]

where \( Z^{\mu\nu} = \text{diag}(-1, c_s^2/a^2 \delta^{ij}) \) is the kinetic matrix for fluctuations that propagate with sound speed \( c_s \) on the FLRW background, and the coefficients \( \{\alpha_i, \beta_j\} \) are fixed in terms of \( c_s \) and the scales \( \{M_3, M_4\} \) from (1.1). The decay constant \( f_\pi \) has been used to normalise the interactions so that inflationary correlators are given by\(^2\),

\[ \langle \hat{\zeta}_{k_1} \hat{\zeta}_{k_2} \rangle_{\text{in-in}} \sim \left( \frac{H^4}{f_\pi^4} \right)^{2} p_{k_1}, \quad (1.4) \]

\[ \langle \hat{\zeta}_{k_1} \hat{\zeta}_{k_2} \hat{\zeta}_{k_3} \rangle_{\text{in-in}} \sim \left( \frac{H^4}{f_\pi^4} \right)^{3} \alpha f_{k_1 k_2}, \quad (1.5) \]

\[ \langle \hat{\zeta}_{k_1} \hat{\zeta}_{k_2} \hat{\zeta}_{k_3} \hat{\zeta}_{k_4} \rangle_{\text{in-in}} \sim \left( \frac{H^4}{f_\pi^4} \right)^{3} \left( \beta g_{k_1 k_2 k_3} + \alpha^2 h_{k_1 k_2 k_3} \right), \quad (1.6) \]

where \( p_{k_1}, f_{k_1 k_2}, g_{k_1 k_2 k_3} \) and \( h_{k_1 k_2 k_3} \) are known functions of the momenta \([5–7]\) (with an overall momentum-conserving delta-function removed). Measurements of the power spectrum (1.4) fix \( f_\pi = (58.64 \pm 0.33) H [8] \), while bounds on primordial non-Gaussianities, the bispectrum (1.5) and trispectrum (1.6), place constraints on the coefficients \( \alpha_i \) and \( \beta_j \), respectively. Our goal in this work is to assess under what conditions \( S_{\text{LO}}[\pi] \) mediates unitary scattering amplitudes between \( \pi \) quanta, and hence apply unitarity as a theoretical constraint on the coefficients \( \alpha_j \) and \( \beta_j \) to complement observational searches for primordial non-Gaussianity.

### Inflationary Amplitudes

There are a number of obstacles which prevent directly applying Lorentz-invariant amplitude techniques to inflation. Even working within the decoupling limit (1.2), the expanding spacetime background spontaneously breaks the Poincaré symmetry which underpins our usual definition of a scattering \( \hat{S} \)-matrix (energy is no longer conserved, particle production can occur, plane waves are no longer well-defined asymptotic states, there is no known LSZ procedure to relate field correlators to in-out observables, etc.). Overcoming these issues would be a monumental task. Here, we propose to take but the first step beyond the usual Lorentz-invariant Minkowski spacetime setting. By focussing on *subhorizon modes*, i.e. fluctuations

\(^2\)The factors of \( H \) in the correlators can be most easily seen by switching to conformal time, in which the mode functions of \( \pi \) scale as \( \sim H^{3/2}, \sqrt{-g} \sim (H\eta)^{-1} \) and \( g^{\mu\nu} \sim (H\eta)^2 \), so the interaction coefficients scale as \( \sqrt{-g} \alpha_i \left( g^{\mu\nu} \right)^3 / f_\pi^2 \sim \alpha_i H^2 / f_\pi^2 \) and \( \sqrt{-g} \beta_i \left( g^{\mu\nu} \right)^4 / f_\pi^4 \sim \beta_i H^4 / f_\pi^4 \).
whose time derivatives are much larger than any time derivative of the background, it is possible to define in-out scattering amplitudes. Formally, the FLRW mode function for an on-shell π particle\(^3\) is \(f_\pi(t) \propto H_\nu^{(1)}(c_\pi |p|/aH)\), a Hankel function of order \(\nu\) (\(= 3/2\) for a light scalar on quasi-de Sitter), but on sufficiently subhorizon scales the effects of the expanding spacetime become unimportant and \(\pi\) behaves like a scalar on flat space, with plane-wave mode functions (that conserve energy and momentum). The \(|p|\) at which this happens is set by \(\nu\), since \(H_\nu^{(1)}(z) \sim e^{iz/\sqrt{z}}\) when \(z \gg |\nu^2 - 1/4|\) (up to an overall phase factor). When considering scattering over timescales \(Ht \ll 1\) (so that \(c_\pi |p|/aH = \text{const} - c_\pi |p|t/a + \ldots\)), a \(\pi\) particle with this momentum is described by a plane-wave mode function,\(^\text{Subhorizon Modes:}\)

\[
\frac{c_\pi |p|}{aH} \gg \left| \nu^2 - \frac{1}{4} \right| \quad \Rightarrow \quad f_\pi(t) \propto e^{ic_\pi |p|t/a}.
\]

These approximate plane waves will characterise our in- and out- states, and allow us to apply the LSZ procedure and compute scattering amplitudes as if on flat space. Throughout this work, we will refer to (1.7) as “subhorizon scales”, i.e. the regime in which \(\pi\) modes behave as approximately plane waves.

However with \(f_\pi\) fixed this EFT retains the effects of symmetry breaking: in the decoupling limit (1.2) and for subhorizon modes (1.7), the action \(S_{\text{LO}}[\pi]\) in (1.3) describes a single scalar field on a flat background with (spontaneously) broken boosts. Unlike in Lorentz-invariant EFTs, the scattering amplitude may now depend on the Lorentz-frame in which it is evaluated. That is to say, once coordinates have been fixed so that the background expansion is isotropic (the cosmic rest frame), there is no longer any freedom to boost away the centre-of-mass motion—this is depicted in Figure 1. As a result, the scattering amplitudes computed from \(S_{\text{LO}}[\pi]\) will depend on both the total incoming energy, \(\omega_s\), and the total incoming momentum, \(p_s\). For instance, the tree-level \(2 \rightarrow 2\) amplitude has the form,

\[
A_{\text{tree}}^{2 \rightarrow 2} = \frac{\omega_s^4}{f_\pi^4} F \left( \frac{c_\pi |p_s|}{\omega_s} \right)
\]

where \(F\) is a dimensionless function of the ratio \(\rho_s = c_\pi |p_s|/\omega_s\) (related to the velocity of the centre-of-mass, \(\rho_s = v_{\text{CoM}}/c_\pi\)) and depends linearly on the coefficients \(\beta_i\) and \(\alpha_i\alpha_j\) in (1.3). The precise question we wish to address is: when is \(A_{\text{tree}}^{2 \rightarrow 2}\) a viable (unitary) description of the subhorizon physics during inflation? Since this amplitude stems from the EFT of Inflation, which has assumed only that there is a single light degree of freedom (arising from the symmetry breaking), this is equivalent to asking: when is inflation approximately single-field (and weakly coupled) on subhorizon scales?

\(^3\pi\) fluctuations propagating on an FLRW background are canonically quantised as [9],

\[
\pi(t, x) = \int \frac{d^3k}{(2\pi c_\pi)^3/2} e^{ik \cdot x} \left[ \hat{a}_k f_k(t) + \hat{a}^\dagger_k f^*_k(t) \right]
\]

where \([\hat{a}_k, \hat{a}^\dagger_k] = -i\) fixes the normalisation of the above mode function as \(\sqrt{\frac{c_\pi}{4\pi}} e^{-i\omega/2}\).
To answer this question, we derive new unitarity constraints at a finite centre-of-mass velocity. The existing Lorentz-invariant implementation of unitarity (the partial-wave expansion) relies on using boosts to set $|p_s| = 0$, which can no longer be done for (1.8)—instead, we have developed a more general “spherical-wave expansion” which can accommodate $|p_s| \neq 0$. Using these spherical-wave amplitudes, we identify the region of $\{c_s, \alpha_1\}$ parameter space in which the $S_{LO}[\pi]$ interactions are unitary (for subhorizon modes in the decoupling limit). Since Lorentz-boosts are spontaneously broken, the EFT cutoffs in energy and momentum need no longer be the same (for instance the UV completion has states characterised by both a mass and a sound speed). In particular, there is a maximum energy and a maximum momentum,

$$\omega_{\text{max}} = 2f_\pi \left( \frac{30\pi c_s^4}{1 - c_s^2} \right)^{1/4}, \quad c_\pi |p_s|_{\text{max}} = 2f_\pi \left( \frac{2\pi}{15\alpha_1^2} \right)^{1/4} \tag{1.9}$$
at which the EFT breaks down. Demanding that both of these cutoff are large enough for the EFT to resolve subhorizon scales \(1.7\) can either be used as a theoretical prior—for instance assuming that subhorizon physics is single-field and weakly coupled improves the Planck 68% confidence interval on equilateral and orthogonal \(f_{NL}\) by a factor of \(\approx 3\)—or can be used as a way to test with future data whether other light fields play an important role on subhorizon scales.

We also point out the curious feature that Lorentz-violating cubic interactions (such as \(\dot{\pi}^3\)) lead to a minimum interaction energy, since otherwise their exchange contributions to \(A_{2\to2}^{\text{free}} \sim \omega_s^6/s\) grow faster at small \(s\) than is allowed by unitarity. For \(\alpha_1 \dot{\pi}^3/f_\pi^2\), scattering at fixed \(|p_s|\) leads to the unitarity requirement,

\[
s > s_{\text{min}} = \frac{|p_s|^6}{f_\pi^2} \frac{25}{64} \frac{\alpha_1^2}{4\pi}.
\]

Since \(s = \omega_s^2 - c_\pi^2 |p_s|^2\), this means that it is only possible for the EFT to have comparable cutoffs in energy and momentum if \(\alpha_1\) vanishes (or is made smaller than the \(\pi\) mass and other slow-roll suppressed corrections which we have neglected).

### Loops and Power Counting

Since our analysis relies on studying the high-energy behaviour of amplitudes from \(S_{\text{LO}}\), it is essential that higher-order EFT corrections can be neglected. The precise form and size of these corrections depends on the underlying UV physics, and cannot be determined within the EFT. However, given knowledge of \(S_{\text{LO}}\) (e.g. measurements of \(c_s, \alpha_1, \ldots\)), it is possible to place lower bounds on these corrections from radiative stability—i.e. that quantum corrections to \(S_{\text{HO}}\) (in particular the running induced by loops of \(S_{\text{LO}}\)) are at most an order one effect. Placing such bounds on \(S_{\text{HO}}\) is important because \(S_{\text{LO}}\) can only ever describe energy scales at which \(S_{\text{HO}}\) can be neglected, and in general it is not possible to simply tune \(S_{\text{HO}}\) to zero (over a wide range of scales) because they are regenerated by loops of \(S_{\text{LO}}\).

We therefore complement our analysis of unitarity in theories with broken boosts with an exploration of radiative stability, and show how to power count the higher-order corrections to the EFT of Inflation. In particular, the background scale \(H\) is separated from the decay constant \(f_\pi\) by less than two orders of magnitude, so it may seem at first sight that describing subhorizon physics (processes with \(\omega \gg H\)) is not possible. However, the EFT cutoff (and the scale at which symmetry is restored) is related to \(f_\pi\) by a factor of the field coupling, and in particular can be as large as \(\Lambda^4 \sim 16\pi^2 f_\pi^4 \sim (200H)^4\). This is precisely the analogue of chiral perturbation theory (\(\chi\PT\)) for pions, in which the decay constant (\(f_\pi^\text{pion} = 93\ \text{MeV}\)) is very close to the pion mass (\(m_\pi^\text{pion} \approx 140\ \text{MeV}\)), and so it is only possible to describe pion scattering consistently thanks to the fact that the chiral symmetry breaking scale is \(\Lambda_{\chi_{SB}} = 4\pi f_\pi \sim 1.2\text{GeV}\) [10] (see e.g. [11] for further discussion of power counting in \(\chi\PT\)).
Synopsis

In section 2, we derive constraints from radiative stability on the higher-order EFT corrections, beginning with the familiar Lorentz-invariant examples of a single scalar field and a massive vector field before moving on to our new analysis of a scalar with broken boosts and then the EFT of Inflation. In section 3, we relate the EFT interaction coefficients to the scale at which (perturbative) unitarity is lost, again providing a review of the well-known Lorentz-invariant case before turning to how this can be extended to theories with broken boosts, and finally comparing our results with the Planck constraints on $f_{NL}$. We derive constraints from radiative stability and unitarity independently, so that sections 2 and 3 may be read in either order. In section 4 we summarise and discuss possible future work.

2 Radiative Stability Bounds

In order to construct a useful effective field theory, one must employ a consistent “power counting scheme” (a set of rules for deciding which operators are the most important) which is “radiatively stable” (preserved under quantum corrections). In particular, when we truncate our EFT and include only a particular set of leading order interactions, $S_{LO}$, they will inevitably generate higher order corrections, $S_{HO}$. In this section, we first review radiatively stable power counting schemes in Lorentz-invariant theories, and then discuss how these can be extended to theories in which boosts are broken, finally arriving at a consistent power counting scheme for the EFT of Inflation in which the interaction coefficients are naturally bounded in terms of the sound speed of scalar fluctuations, $c_s$.

2.1 EFTs with Lorentz Invariance

We begin by briefly reviewing the constraints imposed by radiative stability in effective field theories with Lorentz invariance, building up from a single scalar field to a massive vector field. This provides a simple example of estimating loop corrections in a theory with a nonlinearly realised gauge symmetry (which parallels our approach to the EFT of Inflation in section 2.3).

Single Scalar Field

The first step in constructing an effective field theory action is to identify a basis of local operators built from the desired degrees of freedom and consistent with the desired symmetries. For example, to construct a Lorentz-invariant theory using a single scalar field with a shift symmetry, such a basis would be all possible Lorentz scalars built from the field and its derivatives,

$$\mathcal{L}[\phi, \partial_{\mu}] = Z(\partial \phi)^2 + \frac{(\partial \phi)^4}{(M_4)^4} + \frac{(\partial \phi)^6}{(M_6)^6} + ... + \frac{1}{2} \frac{(\partial \partial \phi)^2}{(M_2)^2} + \frac{(\partial \partial \phi)^4}{(M_4)^8} + ...$$  (2.1)

The second step is to consider which choices of the various scales, $\{M_n, M'_n, ...\}$, are radiatively stable. This is important because while large hierarchies between the scales can seem
acceptable classically, in the quantum theory these interaction coefficients run and hierarchies are typically washed out. For example, a loop of two \((\partial \phi)^4/M_4^4\) vertices requires a counterterm that schematically looks like \(\sim (\partial \partial \phi)^4/M_4^5\), causing \(M_4'\) to flow down to \(M_4\) upon renormalisation—this means that once a \((\partial \phi)^4/M_4^4\) interaction is included in our theory, it is not possible to switch off higher order terms like \((\partial \partial \phi)^4\) (over a range of scales) since they are generated by loops.

To systematically assess the impact of quantum corrections, it is convenient to adopt a particular “power counting scheme”, in which every scale \(\{M_n, M_n', ...\}\) is replaced with a dimensionless (order unity) constant according to a set of rules: for example, in the well-known “single-scale-single-coupling” scheme\(^4\) every derivative is suppressed by a single scale \(\Lambda\) and every field is suppressed by a second scale \(\Lambda_\phi = \Lambda/g_\phi\) (which can be written in terms of \(\Lambda\) and a single dimensionless coupling \(g_\phi\), such that \[S[\phi] = \int d^4x \Lambda^4/g_\phi^2 L[\phi/\Lambda, \partial/\Lambda] \]
, where the overall scale in front of \(L\) is chosen to give a canonical kinetic term. Adopting any particular power counting scheme always represents a slight loss of generality from (2.1), since not all UV-complete theories produce an action of this form in the IR (e.g. single-scale-single-coupling only captures those characterised by a single heavy mass, \(\Lambda\), and a single coupling strength \(g\) between heavy and light physics), but with the significant gain that any loop correction may now be systematically estimated. In the case of the single-scale-single-coupling scheme, all interactions are of the form \(O_{ab} = g_\phi^{a-2} \partial^{2b} \phi^a/\Lambda^{a+2b-4}\) (multiplied by an order unity Wilson coefficient), and the \(L\)-loop diagram made from \(V_{ab}\) such vertices produces a counterterm\(^5\),

\[
\Delta O_{ED} \sim \left( \frac{\Lambda^4}{g_\phi^2} \right) \sum_{a} V_{ab} \left( \frac{g_\phi}{\Lambda \partial} \right)^{a} \sum_{b} V_{ab} \left( \frac{\partial^2}{\Lambda^2} \right)^{b} \sum_{a} V_{ab} \left( \frac{\partial^4}{16\pi^2} \right)^{L} (\partial \phi)^{2-2L+\sum_{a} V_{ab}(a-2)} = \left( \frac{g_\phi}{4\pi} \right)^{2L} g_\phi^{E-2} \partial^D \phi^E \frac{\Lambda^{E+2D-4}}{\Lambda^{E+2D-4}}.
\]

(2.2)

In this power counting scheme, radiative stability (i.e. that running from loop corrections is at most an order one correction to the tree-level Wilson coefficient) requires that \(g_\phi \lesssim 4\pi\). This demonstrates something important: the scale suppressing the fields, \(\Lambda/g_\phi\), cannot be made arbitrarily lower than the scale suppressing the derivatives, \(\Lambda\), because a loop can trade \(\phi\)’s for derivatives. Note that since derivatives can never be turned back into \(\phi\)’s, a hierarchy like \(\Lambda/g_\phi \gg \Lambda\) is perfectly consistent with radiative stability.

\(^{4}\)This scheme is also known as SILH, after its original use for a “Strongly Interacting Light Higgs”\[12\]. Setting \(g_\phi = 4\pi\) recovers the “Naive Dimensional Analysis” introduced in\[13\] (see also\[10, 14–16\]), and \(g_\phi = 1\) is the traditional Weinberg power counting for EFTs\[17\].

\(^{5}\)The total number of \(\phi\) lines in such a graph is \(\sum_{a} V_{ab} = 2I + E\), where \(I\) is the number of internal \(\phi\) propagators and \(E\) is the number of outgoing \(\phi\) legs. Since the Euler formula for planar graphs gives the number of internal lines \(I = L - 1 + \sum_{a} V_{ab}\), this fixes \(E\) in terms of \(V_{ab}\) and \(L\). In a regularisation scheme which respects the EFT counting (e.g. dimensional regularisation), the \(\Lambda\) scales can arise only from each vertex factor. The typical size of each loop integral is \(d^4k/(2\pi)^4 \sim 1/(4\pi)^2\), and then finally the total number of derivatives, \(D\), is determined by the total dimension of the operator being 4. See e.g.\[18\] for a review.
The strategy which we shall adopt in this paper is to separate out the leading-order interactions of interest and treat them like (2.1), assuming nothing about their overall size or hierarchies (which is determined by the underlying UV physics), but then to include higher-order corrections for which we will assume a simple power counting scheme, determined by a single scale and a small number of dimensionless parameters (this is arguably the most agnostic one can be, since one must always assume something about the higher-order corrections if one is to trust leading-order computations within the EFT). For example, if we were to focus on a leading order interaction like \((\partial \phi)^4\) for our single scalar field, we would write,

\[
S[\phi] = \int d^4x \left( -\frac{1}{2}(\partial \phi)^2 + \frac{C}{f_\phi^4}(\partial \phi)^4 + \frac{\epsilon_0 \Lambda^4}{g_0^2}L_{\text{HO}} \left[ \frac{g_\phi \phi}{\Lambda}, \frac{\partial}{\Lambda} \right] \right)
\]

where \(C/f_\phi^4\) is a free Wilson coefficient, higher order interactions are suppressed in derivatives by \(\Lambda\) and in fields by \(\Lambda/g_\phi\), and \(\epsilon_0\) controls any overall hierarchy between \(L_{\text{LO}}\) and \(L_{\text{HO}}\). Loops from \(L_{\text{HO}}\) generate corrections to \(L_{\text{HO}}\) which \(\sim \epsilon_0^{-1} (g_\phi/4\pi)^{2L}\), and so radiative stability requires \(g_\phi \lesssim 4\pi\) and \(\epsilon_0 \lesssim 1\). But since loops of \((\partial \phi)^4\) can also generate higher order interactions, radiative stability also requires that the scale \(\Lambda\) cannot be made arbitrarily small (it must be at least as large as the scale \(C/f_\phi^4\)) and that the hierarchy \(\epsilon_0\) cannot be made arbitrarily small (it must be within at least one loop factor of \(C/f_\phi^4\)),

\[
g_\phi^2 \Lambda^4 \gtrsim \frac{C}{f_\phi^4} \quad \text{and} \quad \epsilon_0^2 g_\phi^2 \Lambda^4 \gtrsim \frac{g_\phi^2 C}{16\pi^2 f_\phi^4}.
\]

Conversely, loops from \(L_{\text{HO}}\) can never renormalise either \((\partial \phi)^2\) or \((\partial \phi)^4\) in any scheme which respects the EFT power-counting\(^6\), so a hierarchy like \(C/f_\phi^4 \ll g_\phi^2/\Lambda^4\) is radiatively stable. These bounds will be important in Section 3 when we compute high-energy scattering amplitudes using only leading-order interactions like \((\partial \phi)^4\), as they tell us under what conditions the higher order corrections may be safely neglected (in a radiatively stable way).

**Massive Vector Field**

Before moving on to theories with broken boosts, let us exemplify how radiative stability is implemented when gauged symmetries are spontaneously broken. Consider a massive vector field \(A^\mu\), with a power counting of the form,

\[
S_U[A^\mu] = \int d^4x \left\{ -\frac{1}{4} F^{\mu\nu} F_{\mu\nu} - \frac{1}{2} m^2 A^2 + \frac{M^4}{g_A^2} \mathcal{L}_{\text{gauge}} \left[ \frac{g_A F^{\mu\nu}}{M^2}, \frac{\partial \mu}{M} \right] + \frac{\epsilon_* M^4}{g_A^2} \mathcal{L} \left[ \frac{g_A A^\mu}{M}, \frac{\partial \mu}{M} \right] \right\}
\]

\(^6\)On dimensional grounds, loops which are regularised in a scheme which do not introduce a new heavy scale (e.g. dimensional regularisation) can only produce interactions with at least an additional four derivatives (or a factor of \(m_5^2/\Lambda^2\), which we assume to be negligible), and so all leading order interactions of form \((\partial \phi)^n\) and \(\partial^2 (\partial \phi)^n\) are not renormalised within the EFT. We refer to [19] for further discussion of this type of non-renormalization in various scalar and gravitational field theories.
where \( F^{\mu\nu} = \partial^\mu A^\nu - \partial^\nu A^\mu \) and we have separated the interactions into those that respect the would-be symmetry \( A_\mu \to A_\mu + \partial_\mu \xi \) in the absence of the mass term, and those that explicitly break this symmetry. It is tempting to conclude, by analogy with the scalar example above, that this action is radiatively stable for any \( \epsilon_* \lesssim 1 \) and \( g_A \lesssim 4\pi \). However this is not the case. Since the propagator for \( A^\mu \) does not have a canonical \( 1/p^2 \) fall-off, naive estimates of loop divergences must be considered more carefully.

To bring the propagator of \( A^\mu \) into a canonical form, we can perform the so-called “Stuckelberg procedure” to replace \( S_U[A^\mu] \) with an equivalent action \( S[\tilde{A}^\mu, \phi] \) in which \( \tilde{A}^\mu \) enjoys a gauge invariance. This is achieved via the replacement,

\[
A^\mu = \tilde{A}^\mu + \frac{\partial^\mu \phi}{m}, \quad (2.6)
\]

where the normalisation of \( \phi \) has been fixed so that the mass term \( -\frac{1}{2} m^2 A^2 \) leads to a canonical kinetic term \( -\frac{1}{2} (\partial \phi)^2 \). There is also a mixing term, \( m \tilde{A}^\mu \partial_\mu \phi \), but at sufficiently high energies, \( \partial \gg m \), the two fields decouple (as required by the Goldstone Equivalence Theorem). The action \((2.5)\) in the decoupling limit \((m \to 0 \text{ with } mM \text{ fixed})\) then corresponds to an effective description of \( \phi \) only,

\[
S[\phi] = \int d^4 x \left\{ -\frac{1}{2} (\partial \phi)^2 + \frac{\epsilon_* M^4}{g_A^2} \mathcal{L} \left[ \frac{g_A \partial \phi}{m M}, \frac{\partial}{M} \right] \right\}. \quad (2.7)
\]

It is now clear that \( g_A \sim 4\pi \) is not radiatively stable, since an \( L \)-loop diagram with \( V_a \) vertices of \( \mathcal{O}_a = \epsilon_* M^4 / g_A^2 (g_A \partial \phi / mM)^a \) requires a counterterm which is only smaller than the tree-level vertices in \( S[\phi] \) if,

\[
\left( \frac{\epsilon_* M^2}{m^2} \right)^{1+\sum V_a} \left( \frac{g_A^2 M^2}{16\pi^2 m^2} \right)^L \lesssim 1. \quad (2.8)
\]

This example highlights the difficulties with power counting directly in the unitary gauge \((2.7)\)—radiative stability actually requires that \( g_A \lesssim m/M \) and \( \epsilon_* \lesssim (m/M)^2 \). This is made transparent by defining new couplings, \( g_\phi = (M/m) g_A \) and \( \epsilon_\phi = (M/m)^2 \epsilon_* \), which brings the decoupling limit action into the single-field-single-coupling form, and is radiatively stable for all \( \epsilon_\phi \lesssim 1 \) and \( g_\phi \lesssim 4\pi \).

### 2.2 EFTs with Broken Boosts

We will now move on to the first main aim of this paper: to explore how radiative stability constrains an effective field theory with a preferred time-like direction, \( n^\mu \), which explicitly breaks Lorentz boosts. The inclusion of \( n^\mu \) introduces a new set of operators into the EFT basis, for example explicit time derivatives \( n^\mu \partial_\mu \) may now appear. For a single scalar field,
we will consider the following power-counting scheme for the higher-order interactions,

$$S[\pi] = \int d^4x \sqrt{-Z} \left\{ \frac{1}{2} Z^{\mu\nu} \partial_\mu \pi \partial_\nu \pi + \mathcal{L}_{\text{LO}} + \frac{\epsilon_\pi \Lambda^4}{g_\pi^2} \mathcal{L}_{\text{HO}} \right\}(g^\mu_\pi A^\mu \Lambda, \frac{\partial_\mu}{\Lambda}, g_\pi n^\mu, Z^{\mu\nu}) \right\}.$$ (2.9)

where $\mathcal{L}_{\text{LO}}$ is a set of (leading order) interactions which are not renormalised by any interaction in $\mathcal{L}_{\text{HO}}$. The two key differences are the addition of a new coupling, $g_n$, which controls the symmetry-breaking (taking $g_n \to 0$ recovers invariance under boosts), and the introduction of a kinetic matrix $Z^{\mu\nu}$ (which can differ from the background metric $g^{\mu\nu}$ in the direction $n^\mu n^\nu$) which is implicitly used to perform all index contractions. Note that although the shift symmetry is softly broken by a small mass, $\delta Z = \epsilon_\pi g_\pi^2 \frac{\partial_\mu}{\Lambda}$ which is implicitly used to perform all index contractions. Note that although boosts are broken, we continue to treat spatial derivatives as suppressed (in particular this retains a linear dispersion relation $\omega \propto |p|$ at leading order). We will assume that $\pi$ has an approximate shift symmetry and the kinetic matrix is diagonal, $Z^{00} = -1$ and $Z^{ij} = c^2 \delta^{ij}$ in the frame $n_\mu = \delta^0_\mu$, where $c_\pi$ is the sound speed of $\pi$ and determines its dispersion relation,

$$Z^{\mu\nu} p_\mu p_\nu := -\omega^2 + \frac{\omega^2}{c_\pi^2} p^2 = -m_\pi^2.$$ (2.10)

Since this forms the free propagator of $\pi$, treating $Z^{\mu\nu}$ as an effective metric simplifies the power counting of the interactions.$^8$

There are now effectively three different scales in the power counting (2.9): in addition to the scale $\Lambda$ which suppresses derivative contractions $Z^{\mu\nu} \partial_\mu \partial_\nu$, and the scale $\Lambda_\pi = \Lambda/g_\pi$ which suppresses $\phi$ insertions, there is now a separate scale $\Lambda_n = \Lambda/g_n$ which suppresses time derivatives $n^\mu \partial_\mu$. However, since $n_\mu n_\nu Z^{\mu\nu} = -1$, loops can remove factors of $n_\mu$, e.g.

$$n^\mu n^\nu \int \frac{d^d k}{(2\pi)^d \sqrt{-Z}} \frac{k_\mu k_\nu}{Z^{\mu\nu} k^2 + m_\pi^2} = n^\mu n^\nu (Z^{\mu\nu})^{-1} \frac{\Gamma(1-d/2)}{d(4\pi)^{d/2}} m_\pi^d$$ (2.11)

and so radiative stability does not allow the scale $\Lambda/g_n$ to be made arbitrarily smaller than $\Lambda$. We can see this concretely by focusing on the interactions in $\mathcal{L}_{\text{HO}}$ with fewest derivatives per field (since including more derivatives will not change the conclusion),

$$O_{ab} = \epsilon_\pi \frac{\Lambda^4}{g_\pi^2} \left(\frac{g_\pi g_\pi \pi_\pi}{\Lambda^2}\right)^a \left(\frac{g_\pi^2 Z_{\pi\pi}}{\Lambda^4}\right)^b,$$ (2.12)

where we have introduced the notation $\pi = n^\mu \partial_\mu \pi$ and $Z_{\pi\pi} = Z^{\mu\nu} \partial_\mu \partial_\nu \pi$. An $L$-loop

---

$^7$The shift symmetry is softly broken by a small mass, $m_\pi$, which we take to be smaller than every other relevant scale in the problem—its only role is to regulate any potential IR divergences and allow for a convergent partial wave expansion in section 3.

$^8$Note that the effective volume element is $\sqrt{-Z} := \sqrt{\det (Z^{\mu\nu})^{-1}} = 1/c_\pi^2$. To avoid confusion, we will never use $Z_{\mu\nu}$ with lowered indices (since if indices are raised/lowered with the background $g_{\mu\nu}$, then $Z_{\mu\nu} \neq (Z^{\mu\nu})^{-1}$). On the other hand, since $g^{\mu\nu} n_\nu = Z^{\mu\nu} n_\nu$, we will use both $n^\mu$ and $n_\mu$. 

---
diagram with $V_{ab}$ insertions of each $O_{ab}$ leads to a counterterm,

$$\Delta O \sim \left( \epsilon_\pi \frac{\Lambda^4}{g_\pi^2} \right) \sum V_{ab} \left( \frac{g_\pi g_n}{\Lambda^2} \right)^n \sum aV_{ab} \left( \frac{g_\pi^2}{\Lambda^4} \right) \sum bV_{ab} \left( \frac{\partial^4}{16\pi^2} \right)^L (\partial\pi)^{2-2L+\sum(a+2b-2)V_{ab}}$$

(2.13)

where the factors of $n^\mu$ may either be contracted into the derivatives (producing $\dot{\pi}$) or into each other ($n^\mu n^\nu = -1$) as in (2.11), and all derivatives are implicitly contracted using $Z^\mu\nu$.

For radiative stability we require that this is smaller than the tree-level terms already in $L_{HO}$,

$$(\epsilon_\pi)^{-1+\sum V_{ab} \left( \frac{g_\pi^2}{16\pi^2} \right)^L (g_n^2)^N} \lesssim 1,$$

(2.14)

where $2N < \sum aV_{ab}$ is the number of $n^\mu n^\nu$ contractions. Since this must hold for all possible choices of $L$, $V_{ab}$ and $N$, we must have $g_\pi \lesssim 4\pi$, $\epsilon_\pi \lesssim 1$ and $g_n \lesssim 1$ separately.

Since loop corrections (in a scheme which preserves the EFT power counting, like dimensional regularisation) always introduce at least 4 additional derivatives (or a factor of $m_\pi^2/\Lambda^2$, which we assume to be negligible), one may have a radiatively stable hierarchy between interactions $\partial^p(\partial\pi)^n$ with $p < 4$ and $p \geq 4$ [19]. For instance, consider an action of the form (2.9) with,

$$L_{LO} = \sum_{a,b} C_{ab} \frac{\dot{\pi}^a Z^{b}}{f_\pi^{2a+4b-4}}$$

(2.15)

where $C_{ab}$ are a set of constant coefficients (and $f_\pi$ is an arbitrary scale). Although the $C_{ab}$ do not receive large renormalisation within the EFT (so can take any value from the point of view of radiative stability), they do generate terms in $L_{HO}$. Radiative stability then places lower bounds on the parameters appearing in $L_{HO}$, analogous to (2.4),

$$\frac{g_\pi^{a+2b-2}g_n^a}{\Lambda^{2a+4b-4}} \gtrsim \frac{C_{ab}}{f_\pi^{2a+4b-4}} \quad \text{and} \quad \frac{\epsilon_\phi \Lambda^4 g_\pi^{a+2b}g_n^a}{g_\pi^2} \lambda^{2a+4b} \gtrsim \frac{g_\pi^{2}}{16\pi^2} \frac{C_{ab}}{f_\pi^{2a+4b-4}}.$$

(2.16)

Compared with the “natural” value $C_{ab}/f_\pi^{2a+4b-4} \sim \epsilon_\phi g_\pi^{a+2b-2}g_n^a/\Lambda^{2a+4b-4}$ inferred from the counting in (2.9), the leading-order coefficients may be either a factor of $1/\epsilon_\phi$ larger or a factor of $16\pi^2/g_\pi^2$ larger without spoiling radiative stability.

We reiterate that the underlying UV physics may not produce a low-energy action of the form (2.9), but for the purposes of computing low-energy observables using $L_{LO}$ we must make some assumption about when $L_{HO}$ can be neglected, and here we have shown how to do this in a radiatively stable way. It may seem that, without independently measuring the corrections in $L_{HO}$, one cannot use the bound (2.16) to say anything about the size of the $C_{ab}$. However, we will now show that when boosts are only broken spontaneously there is an additional relation between $g_\pi/\Lambda^2$ and $g_n/f_\pi^2$ which translates (2.16) into a bound on $C_{ab}$. 

---

\[ \text{Page 12} \]
Spontaneously Broken Boosts

While (2.9) is a consistent power counting for any scalar field theory with broken Lorentz boosts, there is a special tuning of the Wilson coefficients which promotes the global shift symmetry, \( \pi(x) \to \pi(x) + c \), to a local one,

\[
\pi(x) \to \pi(x) + f_\pi^2 \xi(x) \quad \text{and} \quad n_\mu \to n_\mu + \partial_\mu \xi(x) \quad \text{with} \quad Z^{\mu\nu} + (1 - c_s^2) n^\mu n^\nu \quad \text{fixed}, \quad (2.17)
\]

where \( f_\pi^2 \) sets the scale of the associated Noether current (and is no longer arbitrary), \( c_s \) is a fixed constant, and the preferred time-like direction \( n_\mu \) is now allowed to be different at each spacetime point\(^9\). This symmetry corresponds to non-linearly realised Lorentz boosts which leave the speed \( c = c_\pi/c_s \) invariant—it mixes the time-like direction \( n_\mu \) with spatial coordinates, but locally preserves the metric \( c_s^2 g^{\mu\nu} = Z^{\mu\nu} + (1 - c_s^2) n^\mu n^\nu \) with associated \( ds^2 = -dt^2 + dx^2/c^2 \). The constant \( c_s \) in (2.17) describes the mismatch between the \( Z^{\mu\nu} \) cone and the cone which is preserved by the (non-linearly realised) boosts, i.e. the ratio of the scalar speed \( c_\pi \) at low energies to the invariant speed \( c \) at high energies.

This non-linearly realised symmetry has two important consequences. Firstly, it fixes higher-order Wilson coefficients in terms of lower-order ones. For instance, consider expanding \( \mathcal{L}_{LO} \) up to quartic order in fields, producing an action of the form (1.3). For (2.17) to be a symmetry of this action, it is enough for the variation proportional to \( \delta S/\delta Z_\xi \pi = Z_\xi \pi \partial_\mu \xi \partial_\nu \pi \) to vanish, since the variation proportional to \( n_\mu \partial_\mu \xi \) can always be removed by adding powers of \((1 + n_\mu n^\mu)\) to the action (which vanish once we fix the frame \( n_\mu = \delta_\mu^0 \)). Setting \( \delta S/\delta Z_\xi \pi = 0 \) fixes all but one Wilson coefficient at each order in \( \pi \),

\[
\alpha_2 = \frac{1 - c_s^2}{2c_s^2}, \quad \beta_2 + \frac{3}{2c_s^2} \alpha_1 = \frac{(1 - c_s^2)^2}{2c_s^2}, \quad \beta_3 = \frac{1 - c_s^2}{8c_s^4}. \quad (2.18)
\]

The second important consequence of a non-linear symmetry such as (2.17) is that the field coupling \( g_\pi \) is now fixed. Since the interactions must group together into invariant combinations, these must be compatible with the power counting. For instance, terms in \( \pi \) and \( Z_{\pi\pi} \) must now form the invariant,

\[
\hat{\pi} - \frac{\partial_\mu \pi \partial_\nu \pi}{2c_s^2 f^4_\pi} \left( Z^{\mu\nu} + (1 - c_s^2) n^\mu n^\nu \right) \quad \text{c.f.} \quad g_n \pi + \frac{g glimpse}{\Lambda^2} \left( Z_{\pi\pi} + g_{\pi\pi} \right) \quad (2.19)
\]

\(^9\)This allows \( n_\mu \) to “eat” the scalar \( \pi \), and in the unitary gauge \( \pi = 0 \) the dynamics is encoded entirely in \( n_\mu(x) \). When we connect with gravitational theories below, we will see that this is precisely the symmetry required if \( \pi \) is to represent the scalar fluctuations of a spacetime metric which has been foliated using \( n_\mu \)—it allows \( \pi \) to be traded for a geometric description in which the surface normal to \( n_\mu \) fluctuates.
and so we learn that a consistent power counting of the form (2.9) requires\(^\text{10}\),

\[
g_{\pi} \Lambda^2 \sim \frac{g_n}{c_s^2 f_{\pi}^2} \quad \text{and} \quad g_n^2 \gtrsim 1 - c_s^2. \tag{2.20}
\]

This is usually the case when \(\pi\) takes on the role of a Goldstone mode—previously the scale suppressing the field was arbitrary, \(g_{\pi} \pi / \Lambda\) could take any value (providing \(g_{\pi} \lesssim 4\pi\) for radiative stability), but now thanks to the shift symmetry there is a decay constant which naturally normalises \(\pi / f_{\pi}\). For pions, \(\pi\) non-linearly realises a simple shift symmetry, which results in the relation \(\Lambda_{\text{pion}} \sim g_{\pi}^{\text{pion}} f_{\pi}^{\text{pion}} \lesssim 4\pi f_{\pi}^{\text{pion}}\). For non-linearly realised boosts, we have,

\[
g_{\pi} \lesssim 4\pi \quad \Rightarrow \quad \Lambda^4 \lesssim \Lambda_{\text{max}}^4 = \frac{16\pi^2 c_s^4 f_{\pi}^4}{g_n^2} \tag{2.21}
\]

which involves the two additional parameters \(\{g_n, c_s\}\) which characterise the symmetry breaking. \(\Lambda\) is the scale at which the EFT breaks down, and the lowest possible scale at which the Lorentz boosts (2.17) can be restored. The bound (2.21) follows entirely from the symmetry breaking, and does not require any particular interaction to be large (for instance (2.21) still holds if all of the remaining coefficients vanish, \(\alpha_1 = 0, \beta_1 = 0\), etc.)

Since \(g_{\pi} / \Lambda^2\) is now fixed in terms of \(g_n\), the lower bound (2.16) from radiative stability becomes a lower bound on \(g_n\),

\[
g_n \gtrsim \left( c_s^{2a + 4b - 4} C_{ab} \right)^{\frac{1}{2a + 2b - 4}}. \tag{2.22}
\]

Having measured a \(C_{ab} \pi^a Z^{b}_{\pi\pi} / f_{\pi}^{2a + 4b - 4}\) interaction in a theory with spontaneously broken boosts (2.17), a power counting of the form (2.9) in which the symmetry breaking is controlled by a single parameter \(g_n\) is only radiatively stable providing,

\[
1 \gtrsim g_n^{2a + 2b - 2} \gtrsim c_s^{2a + 4b - 4} C_{ab} \quad \Rightarrow \quad C_{ab} \lesssim c_s^{4 - 2a - 4b}. \tag{2.23}
\]

For instance, for the quartic action in (1.3), this bound gives \(\alpha_i \lesssim 1 / c_s^2\) and \(\beta_i \lesssim 1 / c_s^4\)—this is consistent with (2.18) for those coefficients fixed by the symmetry, and additionally constrains the free coefficients \(\alpha_1\) and \(\beta_1\). So while radiative stability in a theory with explicitly broken Lorentz boosts can only relate the leading coefficients \(C_{ab}\) to lower bounds on the scale of higher order corrections, when boosts are broken spontaneously (non-linearly realised) then radiative stability requires that each \(C_{ab}\) be bounded in terms of \(c_s\) (the ratio of \(c_\pi\) to the invariant speed \(c\) which characterises the Lorentz boosts).

We will now show how this theory of a single scalar \(\pi\), with non-linearly realised boosts (2.17), emerges from the EFT of Inflation for metric perturbations about an FLRW background in the decoupling limit, and discuss how these bounds apply to the inflationary bis-

\(^{10}\)Note that \(g_n^2 \gtrsim 1 - c_s^2\) is enough for the \(\dot{\pi}^2\) term required in (2.19) to appear as a small correction of \(O((1 - c_s^2) / g_n^2)\) to the higher order terms in \(\dot{\pi}^n\). Meanwhile, since this is the only way to produce \(Z^a_{\pi\pi}\) interactions, one must have \(g_e \sim g_n \Lambda^2 / c_s^2 f_{\pi}^2\) (values much greater than this would not be acceptable).
2.3 EFT of Inflation

We will now show how the radiative stability constraints developed above can be applied to a theory of metric fluctuations about a background which spontaneously breaks boosts.

The action should be constructed from all local operators built from \( \delta g^{\mu\nu} \) which are invariant under spatial diffeomorphisms (since the background breaks temporal diffeomorphisms), as discussed in [4]. As above, we will separate the action into a leading-order piece, \( S_{\text{LO}} \), given in (1.1), plus higher-order corrections, \( S_{\text{HO}} \). In particular, while we allow for arbitrary scales \( \{M_n\} \) in (1.1), we will assume a particular power counting scheme for \( S_{\text{HO}} \),

\[
S_{\text{HO}}[\delta g^{\mu\nu}] = \int d^4x \sqrt{-g} \left\{ M^4 \mathcal{L}_{\text{diff}} \left[ \frac{R}{M^2}, \frac{\nabla_\mu}{M} \right] + \epsilon_* \frac{M^4}{g^*_2} \mathcal{L} \left[ \delta g^{\mu\nu}, \frac{\nabla_\mu}{M}, g_* n^\mu, \eta_* t \right] \right\},
\]

where \( R \) is the Ricci scalar of \( g^{\mu\nu} \), \( n^\mu \) is a constant time-like unit vector (which we take to be \( \delta^\mu_0 \)), and we have separated the terms according to whether they are invariant under the would-be diffeomorphism symmetry in the absence of the background. This split into \( \mathcal{L}_{\text{diff}} \) and \( \mathcal{L} \) parallels the massive vector power counting (2.5), with the addition of two new couplings: \( g_* \) controls the breaking of boosts, and \( \eta_* \) controls the breaking of time translations\(^{11}\). Since the canonically normalised metric fluctuations are \( 2 \delta g^{\mu\nu}/M_P \), it is the ratio \( 2M/M_P \) that plays the role of the field coupling.

If we were to compare the power counting parameters in \( S_{\text{HO}} \) with the scales in \( S_{\text{LO}} \), we might expect “natural” values of\(^{12}\),

\[
M^4 \sim M_P^2 H^2, \quad \epsilon_* \sim |\dot{H}|/H^2, \quad \eta_* \sim \ddot{H}/|\dot{H}|, \quad g^2_* \sim M_2^2/M_1^4,
\]

but in order to determine which range of these parameters are required for a consistent EFT we must turn to radiative stability arguments.

As in the massive vector case, analysing the radiative stability of (2.24) directly is difficult because the propagator of \( \delta g^{\mu\nu} \) no longer has a canonical \( 1/p^2 \) fall-off. It is more convenient to

---

\(^{11}\)Note that \( g_* \) should be at least as large as \( \eta_*/M \), since the breaking of time-translations automatically generates a preferred direction, \( \eta_*/M \nabla_\mu t \), but may be much larger (e.g. scale invariance sets \( \eta_* \to 0 \) but leaves \( g_* \) finite [20]).

\(^{12}\)For instance, for a canonical (dimensionless) scalar field which slowly-rolls down a potential,

\[
\mathcal{L} = M_P^2 (\partial \phi)^2 + M_P^2 H^2 V(\phi) + M^4 \mathcal{L} \left[ \phi, \frac{\nabla_\mu}{M} \right] \quad \text{with} \quad \partial^\mu \phi \sim (\sqrt{\epsilon} H)^n \phi
\]

fluctuations in unitary gauge have the form (2.24) with,

\[
\epsilon_* \sim \epsilon H^2/M^2, \quad g_* \sim \sqrt{\epsilon} H/M, \quad \eta_* \sim \sqrt{\epsilon} H,
\]

where \( M \gtrsim H \) controls the small derivative corrections to the potential. Tuning the potential to be flat corresponds to suppressing each \( \lambda_n \phi^n \) interaction in \( V(\phi) \) by a power of \( \epsilon^{n/2} \), thus lowering \( \eta_* \) to \( \epsilon H \), as described in [21].
perform the Stuckelberg procedure, restoring the broken time diffeomorphisms at the expense
of introducing an additional degrees of freedom \( \pi \). This is achieved by replacing the metric
by \( \tilde{g}^{\alpha \beta} = \partial_\mu \tilde{x}^\alpha \partial_\nu \tilde{x}^\beta \), as though performing the time diffeomorphism,

\[
x^0 = \tilde{x}^0 + \frac{\pi(\tilde{x})}{f_\pi^2}, \quad x^i = \tilde{x}^i.
\]  

(2.28)

This results in a new theory, with degrees of freedom \( \delta \tilde{g}^{\mu \nu} \) and \( \pi \), which now nonlinearly
realises time diffeomorphisms. The scale \( f_\pi \) is fixed so that \( \pi \) has a canonical kinetic term,

\[
f_\pi^4 = -2M_P^2 \mathcal{H} c_s \quad \text{and} \quad c_s^2 = \frac{1}{1 + \frac{2M_P^2}{M_T^2} f_\pi^2},
\]  

(2.29)

where \( c_s = c_\pi / c_T \) is the ratio of the scalar sound speed to the tensor sound speed. Equation
(2.29) ensures that expanding \( S_{\text{LO}} \) to quadratic order in \( \pi \) gives an action\(^{13}\),

\[
S^{(2)}_{\text{LO}}[\pi, \delta \tilde{g}^{\mu \nu}] \supset \int d^4 x \sqrt{-Z} \left\{ -\frac{1}{2} Z^{\mu \nu} \partial_\mu \pi \partial_\nu \pi - \frac{f_\pi^2}{M_P} \left( 1 + c_s^2 \right) \frac{M_P \delta g^{00}}{2} \right\},
\]  

(2.30)

in which the kinetic matrix \( Z^{\mu \nu} \) has canonical form \((Z^{00} = -1, Z^{ij} = c_s^2 / a^2 \delta^{ij})\). The mixing
between \( \pi \) and metric fluctuations becomes unimportant at energies \( \omega \gg f_\pi^2 (1 + c_s^2) / M_P \)—we
will work firmly in this decoupling limit (1.2), in which the Stuckelberged metric is simply,

\[
\delta g^{\mu \nu} = n^\mu n^\nu \left[ -\frac{2\pi}{f_\pi^2} + \frac{1 - c_s^2 \pi^2}{c_s^2} + \frac{Z_{\pi \pi}}{c_s^2 f_\pi^4} \right]
\]  

(2.31)

and the action \( S[\pi, \delta \tilde{g}^{\mu \nu}] \) becomes a functional of \( \pi \) only.

The higher-order corrections in \( \mathcal{L}_{\text{gauge}} \) do not contribute any \( \pi \) interactions (since they
are invariant under temporal diffeomorphisms), while the symmetry-breaking interactions can be written as,

\[
\int d^4 x \sqrt{-Z} \left\{ c_s^3 \frac{\epsilon_s M_4}{g_s^2} \left[ g_s^\alpha g_s^\beta \nabla_\mu n_\mu, \nabla_\nu Z^{\mu \nu} \right] \right\},
\]  

(2.32)

with \( \delta g^{00} \) given by (2.31). As in section 2.2, it is convenient to use the kinetic matrix \( Z^{\mu \nu} \) as
an effective metric for the interactions, but this requires rescaling \( M \) and \( g_s \) so that the power
counting in (2.32) matches that in (2.24) (which can be seen from \( g^{\mu \nu} \nabla_\mu \nabla_\nu \sim c_s^{-2} Z^{\mu \nu} \nabla_\mu \nabla_\nu \)
and \( g^{\mu \nu} n_\mu \nabla_\nu = Z^{\mu \nu} n_\mu \nabla_\nu \)). Just as with the massive vector, we can now see that it would
be incorrect to conclude that radiative stability requires \( g_s \lesssim 1 \) and \( \epsilon_s \lesssim 1 \). Explicitly, if we

\(^{13}\)The Stuckelberg field \( \pi \) also acquires a small mass from the explicit time dependence of the Wilson
coefficients, \( \sqrt{-g M_4^2 (t + \frac{\pi}{f_\pi})} \sim \sqrt{-Z} \mathcal{H} c_s \pi^2 \).
consider an $L$-loop diagram containing $V_{ab}$ vertices from (2.32) of the form,

$$O_{ab} = \epsilon_* c_s^2 M^4 \left( \frac{g_\pi^2}{f_\pi^2} \right)^a \left( \frac{g_\pi^2}{c_s f_\pi^2} Z_{\pi\pi} \right)^b,$$

(2.33)

where we have neglected insertions of the $\pi^2$ term from $\delta g^{00}$ because they will turn out to be subleading and focussed on the interactions with fewest derivatives (since adding more derivatives will not change the conclusion), we find that radiative stability requires

$$\left( \frac{\epsilon_* M^4}{M_p^2 \dot{H}} \right)^{-1 + \sum V_{ab}} \left( \frac{2 g_\pi^2 M_4^4}{16 \pi^2 f_\pi^2} \right)^L \left( c_s g_\pi^2 \right)^N \lesssim 1.$$

(2.34)

This can be made transparent by defining new power counting parameters

$$\Lambda = c_s M, \quad g_n = c_s g_* \quad \frac{g_\pi}{\Lambda^2} = \frac{1}{c_s f_\pi^2}, \quad \frac{\epsilon_\pi}{g_\pi^2} = \frac{1}{c_s g_\pi},$$

(2.35)

in terms of which (2.32) takes the form (2.9), and so radiative stability requires $\epsilon \lesssim 1, \ g_\pi \lesssim 4\pi$ and $g_n \lesssim 1$.

But the higher-order corrections cannot be made arbitrarily small, since they are also renormalised by loops from $S_{LO}$. Repeating the exercise of estimating loops of $\pi$ dimensional grounds, one arrives at the analogous relations for a $\Lambda_{\text{max}}$ (2.21) and for a $g_n^{\text{min}}$ (2.22), which in terms of the original power counting parameters in (2.24) read,

$$M_n^4 \lesssim M_{\text{max}}^4 = \frac{16 \pi^2 f_\pi^4}{c_s^2 g_*^2} \quad \text{and} \quad 1 \gtrsim (c_s g_*)^{2n-2} \gtrsim \frac{M_n^4}{f_\pi^4 c_s^{1-2n}},$$

(2.36)

for every $n$. Just as in (2.23), for a power counting scheme in which the spontaneous breaking of boosts is described by a single order parameter $g_n$ to be radiatively stable, one requires that the leading-order coefficients are bounded: $M_n^4 \lesssim f_\pi^4 c_s^{1-2n}$. This suggests the definition of dimensionless coefficients,

$$M_n^4 = c_n \times f_\pi^4 c_s^{1-2n},$$

(2.37)

in terms of which $g_\pi$ is radiatively stable providing $1 \gtrsim (c_s g_*)^{2n-2} \gtrsim c_n$. This result coincides with the naturalness arguments made in [22], but here follows from a consistency (radiative stability) of the EFT and assumes only that the UV physics produces higher-order corrections of the form (2.24). There is also an analogous bound to (2.16) on the hierarchy $\epsilon_*$,

$$1 \gtrsim \frac{\epsilon_* M^4}{M_p^2 |\dot{H}|} \gtrsim \frac{M^4}{M_{\text{max}}^4} \frac{c_n}{(c_s g_*)^{2n-2}},$$

(2.38)

for every $n$. Note that (2.36) and (2.38) imply that $g_*^2 \gtrsim M_n^4 / M_1^4$ and $\epsilon_* \lesssim |\dot{H}| / H^2$ when $M \sim M_p^2 H^2$, as anticipated in (2.26).
It is particularly worth emphasising the bounds which arise from the $M^4_n$ interactions, since they place bounds on $g_\pi$ and $\epsilon_\pi$ in terms of the sound speed ratio $c_s$. In terms of the $g_n$ and $\epsilon_\pi$ defined in (2.35),

$$1 \gtrsim g_n^2 \gtrsim 1 - c_s^2 \quad \text{and} \quad 1 \gtrsim \epsilon_\pi \gtrsim \frac{M^4}{f^4} \frac{1 - c_s^2}{16\pi^2}. \quad (2.39)$$

In the absence of the other $M_n$ interactions, as $c_s \to 1$ radiative stability allows for the higher-order corrections to be very small. In particular, the scale suppressing time derivatives can be as high as $\Lambda / \sqrt{1 - c_s^2}$ (in the next section we will see that this is naturally interpreted as a Lorentz-dilation of $\Lambda$ when a collection of $\pi$ particles are boosted to a velocity close to $c_T$). Also, we remark that any scaling in which $\epsilon_\pi g_n^2 \sim 1 - c_s^2$ is particularly useful because it ensures that the true sound speed of $\pi$ really does approach $c_T$ when we tune $M^4_n \to 0$. This is because higher derivative terms like $\partial^2 \dot{\pi}^2$ can contribute to the dispersion relation and redress the $c_s$ given by (2.29) to a new sound speed,

$$1 - \frac{\omega^2}{|p|^2} = 1 - c_s^2 + \epsilon_\pi g_n^2 \frac{\omega^2}{\Lambda^2} + \ldots \quad (2.40)$$

when $M^4_n$ is close to zero. In order to treat the sound speed as a constant (particularly when it is close to one) when scattering at energies $\omega \sim \Lambda$, then one must ensure that the combination $\epsilon_\pi g_n^2 \sim 1 - c_s^2$ in a way consistent with (2.39)\(^\dagger\).

**Connection with Observation**

To connect with observational constraints on the bispectrum (1.5), we can focus on the expansion of $S_{LO}$ up to quartic order in $\pi$ only (1.3). Note that the gauge symmetry we have introduced when performing the Stuckelberg procedure (2.28) is precisely the gauged shift symmetry (2.17) for $\pi$ discussed in section 2.2 (on identifying $n_\mu = \partial_\mu t$). The interaction coefficients $\alpha_j$ and $\beta_j$ are therefore fixed as in (2.18), with the remaining $\{\alpha_1, \beta_1\}$ given by,

$$\alpha_1 = -\frac{4}{3} c_3 - \frac{1}{2} \frac{(1 - c_s^2)^2}{c_s^2}, \quad \beta_1 = \frac{2}{3} c_4 + \frac{2}{c_s^2} c_3 - \frac{1}{8} \frac{1 - c_s^2}{c_s^4} \quad (2.41)$$

where the $c_n$ are given in (2.37), and are required to be $\lesssim 1$ (2.36), and consequently all $\alpha_j \lesssim 1/c_s^2$ and $\beta_j \lesssim 1/c_s^4$ are bounded by radiative stability.

Furthermore, although the EFT cutoff $\Lambda$ and symmetry breaking parameter parameter $g_n$ which suppress higher-order corrections are determined by the underlying UV physics, we have shown that they can be bounded in terms of the leading-order coefficients $c_s$, $\alpha_1$ and $\beta_1$.

\(^\dagger\)For this to be consistent with (2.36), we also need all of the $M_n \sim 1 - c_s^2$ to vanish in this limit as well—see also (3.62) from unitarity bounds.
using radiative stability (assuming the power counting (2.24)). In particular,

\[ 1 \gtrsim g_n^2 \gtrsim \max \left( \frac{1 - c_s^2}{\beta_1 c_s^3}, \frac{1}{12} \bigg( \frac{\alpha_1 c_s^2}{\beta_1} \bigg)^{1/2}, \frac{\beta_1 c_s^4}{\beta_1} \right) \]  

(2.42)
determines the maximum allowed cutoff, \( \Lambda_{\text{max}}^4 = 16 \pi^2 c_s^4 f_\pi^4 / g_n^2 \). In Figure 2, we plot the largest possible \( \Lambda_{\text{max}}^4 \) as a function of \( c_s \) and \( \alpha_1 \), and compare with the 2018 constraints from Planck on the (equilateral and orthogonal) bispectrum. For illustration, if we take the “best fit” values\(^{15} \), \( c_s \approx -0.031 \) and \( \alpha_1 \approx -2100 \), then the strength of the symmetry breaking is at least \( g_n^2 \sim 1.4 \) and the maximum cutoff is \( \Lambda_{\text{max}} \sim 0.093 f_\pi = 5.4 H \). Since these dimensional estimates are subject to order one corrections (from the explicit loop integrals, which we have not performed), one should not conclude from this that radiative stability has been violated—however it does motivate a more careful analysis of these higher order corrections in future, since observational constraints currently include regions of parameter space in which \( g_n \sim O(1) \) and \( \Lambda_{\text{max}} \sim O(H) \).

The central conclusion from this section is that the maximum energy \( \omega_{\text{max}} \) which we identify in section 3 from one-loop corrections to tree-level \( 2 \rightarrow 2 \) scattering is always below (or approximately) the \( \Lambda_{\text{max}}^4 \) estimated dimensionally from \( L \)-loop diagrams with \( V \) vertices, as shown in Figure 2. This is important because the unitarity bounds we develop below are only numerically precise \textit{up to higher order corrections from} \( S_{\text{HO}} \), and we have now established that it is always possible for these corrections to be small in a radiatively stable way. Schematically, when \( \epsilon_4 \) is tuned as low as possible (2.38) (but allowing \( g_n \) and \( \Lambda \) to be arbitrary), we can compare the expected size of the \( 2 \rightarrow 2 \) amplitude from \( S_{\text{LO}} \) (1.3) with that from \( S_{\text{HO}} \) (whose interactions start at \( \partial^4 (\partial \pi)^n \)),

\[ A_{2 \rightarrow 2}^{(S_{\text{LO}})} \sim (4\pi)^4 \frac{\omega^4}{\Lambda_{\text{max}}} \left[ \frac{c_2}{g_n^2} + g_2 \frac{c_2^2 + c_3}{g_n^4} + g_4 \frac{c_2^3 + c_2 c_3 + c_4}{g_n^6} + g_6 \frac{c_2^2 c_3 + c_3^2}{g_n^8} \right] \]  

(2.43)

\[ A_{2 \rightarrow 2}^{(S_{\text{HO}})} \sim (4\pi)^2 \frac{\omega^8}{\Lambda_{\text{max}}^8} \times \max \left( \frac{c_n}{g_n^{2n-2}} \right) \left[ 1 + O \left( \frac{\omega}{\Lambda_{\text{max}}} \right) \right] . \]  

(2.44)

Although very softly breaking boosts by making \( g_n \) as low as possible will raise \( \Lambda_{\text{max}} \) and suppress the corrections from \( S_{\text{HO}} \), it will also suppress the relative contributions of \( c_3 \) and \( c_4 \) from \( S_{\text{LO}} \)—for instance, if \( g_n^{2n-2} < \frac{1}{16 \pi^2} \omega^4 / \Lambda_{\text{max}}^4 \) then the \( c_n \) contribution from \( S_{\text{LO}} \) can no longer be reliably included. Instead, taking \( g_n = 1 \) and breaking boosts strongly ensures that all corrections are suppressed by \( (\omega/4\pi f_\pi)^4 \). Note that the cutoff \( \Lambda \), although it must be \( \gtrsim \omega \), does not need to be significantly larger if \( \omega \ll \Lambda_{\text{max}} \). We will now study the amplitudes from \( S_{\text{LO}} \) in more detail, neglecting \( S_{\text{HO}} \) hereafter.

\(^{15}\)Since \( f_{NL}^{\text{equil}} = -26 \pm 47 \) and \( f_{NL}^{\text{th}} = -38 \pm 24 \), the uncertainty in these values is order one.
\[ \Lambda_{\text{max}}^4 = 16\pi^2 c_s\rho_f^2/g_n^2 \] (where the minimum \( g_n \) is given in (2.42)), with dashed lines indicating where \( \Lambda_{\text{max}} = 3H, 10H \) and \( 30H \) respectively (though note that these are order-of-magnitude estimates of \( \Lambda_{\text{max}} \) and may be subject to order unity corrections). Right panel shows this minimum \( g_n \) (again subject to order unity corrections). The maximum energy identified in section 3 from unitary \( 2 \rightarrow 2 \) scattering,

\[ \omega_{\text{max}}^4/f^4 \rho_f = 480\pi c_s/\left(1 - c_s^2\right) \], is never larger than this \( \Lambda_{\text{max}} \) estimate by more than an \( \mathcal{O}(1) \) factor, so it possible to neglect higher order corrections in a radiatively stable way.

### 3 Unitarity Bounds

The \( \hat{S} \)-matrix is a cornerstone of scattering theory (see for instance [1, 2]). The central idea is to construct an operator \( \hat{S} \) which maps asymptotic in-states (in the far past) to asymptotic out-states (in the far future), such that \( \langle n'|\hat{S}|n \rangle \) gives the probability of transitioning from an initial \( n \)-particle state, \( |n \rangle \), into a final \( n' \)-particle state, \( |n' \rangle \). The scattering amplitude associated with this process, \( A_{n \rightarrow n'} \), is defined as the matrix element, \( \langle n'|\hat{T}|n \rangle = A_{n \rightarrow n'}(2\pi)^4\delta^4(\sum_i p_i) \), where \( \hat{S} = 1 - i\hat{T} \) has been separated into free and interacting parts, and an overall momentum conserving \( \delta \)-function has been extracted.

In order for states to remain properly normalised (necessary for consistent probabilities), \( \hat{S} \) must be a unitary operator, \( \hat{S}^\dagger \hat{S} = 1 \). This can be expressed as a condition on the scattering amplitudes by replacing \( \hat{S} \) with \( \hat{T} \) and inserting a complete set of states,

\[ -i\langle n'|\hat{T}|n \rangle + i\langle n'|\hat{T}^\dagger|n \rangle = \sum_j \langle n'|\hat{T}|j \rangle \langle j|\hat{T}^\dagger|n \rangle, \]

which in terms of momentum-space amplitudes can be written as,

\[ 2\text{Disc}_s A(p_1...p_n \rightarrow p'_1...p'_n) = \sum_j \int d\Pi_j(q) \ A(p_1...p_n \rightarrow q_1...q_j)A^*(p'_1...p'_n \rightarrow q_1...q_j), \]

where \( \int d\Pi_j(q) \) represents the phase space of each \( j \)-particle state (with momenta \( q_1, ..., q_j \)).
and the s-channel\textsuperscript{16} discontinuity is given by,

\[
\text{Disc}_s A_{n\rightarrow n'} = \frac{1}{2i} [A_{n\rightarrow n'} - A_{n'\rightarrow n}^*] \quad (3.3)
\]

and is equal to the imaginary part \(\text{Im} A_{n\rightarrow n'}\) in theories with time-reversal invariance.

The unitarity condition (3.2) will be the focus of this section. We will first review how the perturbative unitarity of scattering amplitudes constrains Lorentz-invariant EFTs and then extend this to EFTs in which boosts are broken, finding new bounds for every centre-of-mass velocity, and arrive at constraints on the EFT of Inflation’s Wilson coefficients \\{\(c_s, \alpha_1, \beta_1\)\}.

Throughout this section, we will denote the momentum of the \(j\)th particle by \(p_j^\mu = (\omega_j, p_j)\), and make use of the usual Mandelstam variables \(s_{ij} = -Z^{\mu\nu}(p_i + p_j)_\mu(p_i + p_j)_\nu\) (where the kinetic matrix \(Z^{\mu\nu}\) is used so that the propagator at momentum \(p_i + p_j\) is simply \(\sim 1/s_{ij}\)). We also define \(p_s^\mu = (\omega_s, p_s)\) as the total incoming momentum and \(\rho_s = c_\pi|p_s|/\omega_s\) as the velocity of the centre-of-mass (i.e. \(\rho_s = v_{\text{CoM}}/c_\pi\), where \(v_{\text{CoM}}\) is the speed measured with respect to\textsuperscript{17} the kinetic matrix \(Z^{\mu\nu}\)).

### 3.1 EFTs with Lorentz Invariance

In Lorentz-invariant theories, the amplitude can be written solely in terms of relativistic invariants (e.g. \(A_{2\rightarrow 2}(p_1p_2 \rightarrow p_3p_4) = A_{2\rightarrow 2}(s_{12}, s_{13}, s_{23})\)) and the boost symmetry can be used to replace arbitrary kinematics with \(p_s = 0\) (transforming to the “centre-of-mass frame”). These two features allow simple unitarity constraints to be placed on \(A_{n\rightarrow n'}\) from (3.2), which we now demonstrate. These results are well-known, but will highlight what needs to be changed when we develop new bounds for EFTs with broken boosts in section 3.2.

First we sketch schematic bounds on \(n\rightarrow n\) scattering (in the forward limit) from a simple dimensional analysis, and then give numerically precise bounds on \(2\rightarrow 2\) scattering using an expansion in partial waves.

#### Elastic \(n\rightarrow n\) Scattering

An elastic process (in which the particles in \(|n'\rangle\) are the same as those in \(|n\rangle\)) has a forward limit in which the outgoing momenta are equal to the ingoing momenta, \(p_i' = p_i\) for all \(i\). In this limit, the initial and final states are identical, and the unitarity condition (3.2) becomes,

\[
2 \text{Im} A(p_1...p_n \rightarrow p_1...p_n) \geq \int d\Pi_n(q) |A(p_1...p_n \rightarrow q_1...q_n)|^2. \quad (3.4)
\]

\textsuperscript{16}We will refer to the \(n \rightarrow n'\) process as the s-channel. The amplitude \(A_{n\rightarrow n'}\) is related via crossing to other channels in which some outgoing and ingoing particles have been exchanged—however these will not be needed since we will work entirely within the s-channel process.

\textsuperscript{17}If one instead used the metric \(g^{\mu\nu}\) to define a \(v_{\text{CoM}}\), then \(\rho_s = c_s v_{\text{CoM}}/c_T\). Consequently, \(p_s^\mu\) satisfies the \(\tau\) dispersion relation when \(\rho_s \rightarrow 1\) (\(c_s|p_s| \rightarrow \omega_s\)), and satisfies the tensor dispersion relation when \(\rho_s \rightarrow c_s\) (\(c_T|p_s| \rightarrow \omega_s\)). Since we work in the decoupling limit, there is nothing special about the point \(\rho_s = c_s\) in our (purely scalar) amplitudes.
Since $2|\mathcal{A}_{n \rightarrow n}| \geq 2\text{Im} \mathcal{A}_{n \rightarrow n} \geq \int d\Pi_n |\mathcal{A}_{n \rightarrow n}|^2$, by estimating the size of $d\Pi_n$ one can use unitarity to place an upper bound on $|\mathcal{A}_{n \rightarrow n}|$. The $n$-particle phase space element is given explicitly by integrating over all future-pointing, on-shell 4-momenta which conserve the total ingoing momenta,

$$d\Pi_n(q) = (2\pi)^4 \delta^4 \left(p_s - \sum_i^n q_i\right) \prod_{i=1}^n \frac{d^4 q_i}{(2\pi)^4} 2\pi \delta \left(q_i^2 + m^2\right) \Theta(q_i^0), \quad (3.5)$$

and is Lorentz invariant. On purely dimensional grounds, the phase space volume scales with the total energy as $\int d\Pi_n \sim (4\pi)^{-1} (s/16\pi^2)^{n-2}$, and so (3.4) becomes a simple bound,

$$|\mathcal{A}_{n \rightarrow n}| \lesssim 4\pi \left(\frac{16\pi^2}{s}\right)^{n-2}, \quad (3.6)$$

where $\sim$ denotes the various order unity numerical factors which we have neglected.

When applied to a perturbative loop expansion of $\mathcal{A}_{n \rightarrow n}$, unitarity (3.4) provides a lower bound on the size of the non-analytic (imaginary) part of the amplitude generated at one-loop in terms of the tree-level amplitude—this is shown graphically in figure 3 for $n = 2$. A violation of the bound (3.6) applied to $\mathcal{A}_{n \rightarrow n}^{\text{tree}}$ signals that the one-loop contribution to the amplitude has exceeded the tree-level contribution, i.e. the loop expansion has broken down\(^\text{18}\).

For example, for the interaction $C_{ab} \partial^2 a \phi b / f_{\phi}^{2a+b-4}$, the tree level scattering amplitude with $V$ vertices is,

$$\mathcal{A}_{n \rightarrow n}^{\text{tree}} \sim \left(\frac{C_{ab}s^a}{f_{\phi}^{2a+b-4}}\right) V \frac{1}{s^{V-1}} \quad (3.7)$$

and the number of external legs is $2n = (b - 2)V + 2$. This tree-level amplitude only satisfies the inequality (3.6) providing,

$$s^{a-2+b/2} \lesssim (4\pi)^{b-2-1/V} f_{\phi}^{2a-4+b} / |C_{ab}|, \quad (3.8)$$

which defines a maximum energy, $s_{\text{max}}$, above which the theory is strongly coupled and loops must be resummed (or new physics beyond the EFT must be included) if one is to restore unitarity. In terms of naive factors of 4\(\pi\), the lowest $s_{\text{max}}$ comes from scattering with the fewest vertices—i.e. the contact diagram with $V = 1$ if $b = 2n$ is even, or the single-exchange diagram with $V = 2$ if $b = 2n + 1$ is odd—but in reality the unitarity bound which gives the lowest numerical $s_{\text{max}}$ will depend on the details of the phase space integration.

When there is a scale of interest in the problem which the EFT must resolve, such as a light mass (e.g. $m_\phi$) or background scale (e.g. $H$ when we discuss inflation), then

---

\(^{18}\)This does not necessarily correspond to new physics since a non-perturbative (all-loop) calculation of $\mathcal{A}_{n \rightarrow n}$ may still satisfy the unitarity condition (3.2).
demanding that \( s_{\text{max}} \gtrsim \) this scale allows (3.8) to be written as an upper bound on the size of \( |C_{ab}| \). For example, the EFT is only consistent if \( s_{\text{max}} > 4m^2 \), and consequently \( |C_{ab}| < (4\pi)^{b-3}(f_\phi/2m_\phi)^{2a-b-4} \) if \( b \) is even or \( |C_{ab}| < (4\pi)^{b-3/2}(f_\phi/2m_\phi)^{2a-b-4} \) if \( b \) is odd.

It is only possible to trust such unitarity bounds on \( |C_{ab}| \) providing this \( s_{\text{max}} \) scale at which we are computing the amplitude is \( \lesssim \Lambda \), the EFT cutoff. Otherwise, higher-order EFT corrections will also contribute to \( A_{n\to n} \) and give large corrections to the bounds. From our discussion of radiative stability in section 2, in the presence of \( C_{ab} \partial^{2a} \phi^b / f_\phi^{2a+b-4} \) the EFT cutoff can be at most \( \Lambda^{a+b-4} = g_\pi^{b-2} f_\pi^{2a+b-4} / |C_{ab}| \), with \( g_\pi \lesssim 4\pi \). This shows that there is at least one radiatively stable tuning of higher-order corrections (namely (2.9) with \( g_\phi \sim 4\pi \)) which ensures that they are no more than an \( \mathcal{O}(1) \) correction\(^{19} \) to the \( A_{n\to n} \) bound (3.6) for any \( n \).

The bound (3.6) is only schematic— there is a (roughly order one) numerical prefactor from the phase space integral which cannot be obtained using simple dimensional arguments. We will now demonstrate how this precise factor can be calculated by focussing on the unitarity bound from \( A_{2\to 2} \) scattering.

**Elastic 2 \( \to \) 2 Scattering**

Consider the transition between two-particle states, \( |p_1, p_2\rangle \) and \( |p_3, p_4\rangle \). In Lorentz-invariant theories, since the amplitude does not depend on the choice of inertial frame, we can without loss of generality use three Lorentz boosts to set \( p_1 + p_2 = 0 \). Two spatial rotations can be used to orient the \( z \)-axis along \( p_1 \), and the remaining azimuthal rotation ensures the amplitude can depend only on \( \cos \theta = \hat{p}_1 \cdot \hat{p}_3 \). Only one of \( |p_1| \) and \( |p_3| \) is independent (thanks to the mass-shell condition of \( p_4^\mu = -p_1^\mu - p_2^\mu - p_3^\mu \)), and can be written in terms of the total incoming energy \( \omega_s = \omega_1 + \omega_2 \). The amplitude for a general scattering process in a Lorentz-invariant theory can therefore be expressed in terms of just two variables, \( A(\omega_s, \theta) \), for instance by taking,

\[
\omega_1 = \omega_2 = -\omega_3 = -\omega_4 = \frac{\omega_s}{2}, \quad s_{12} = \omega_s^2, \quad s_{13} = -\frac{\omega_s^2}{2} (1 - \cos \theta).
\]

For two-particle intermediate states, the phase space integral \( d\Omega_2 \) can be performed exactly, yielding unitarity bounds with precise numerical coefficients. Performing the integrals over the six \( \delta \)-functions in \( d\Omega_2 \) leaves an integral over just two angular variables, which we

---

\(^{19}\) Radiative stability requires that the scale of the higher-order corrections, \( 1/\Lambda_{\text{max}} \), may not be tuned smaller than the floor set by loops of the \( C_{ab} \) interaction. Since \( s_{\text{max}} \) is the scale at which the loop contributions from \( C_{ab} \) exceed its tree-level contribution to the amplitude, one might imagine that \( s_{\text{max}} \sim \Lambda_{\text{max}}^2 \) is inevitable. However, since it is only the real part of the one-loop amplitude which determines the running and the radiative stability floor, the imaginary part of the amplitude could be much larger and violate unitarity at a scale \( s_{\text{max}} \ll \Lambda_{\text{max}} \). In this simple example, there is an additional factor of \( 4\pi \) in our estimate of the imaginary part (in practice this would come from the taking the discontinuity of the polylogarithmic branch cut in the \( L \)-loop amplitude), which for instance allows for \( s_{\text{max}}^2 \) as low as \( \Lambda_{\text{max}}^4/4\pi \) for the interaction \( \partial \phi^3 \).
Figure 3. While the non-linear unitarity relation (3.2) holds non-perturbatively, it can be applied in perturbation theory to relate the one-loop discontinuity (shown left) to the square of the tree-level amplitude (shown right), where the inequality holds whenever the kinematic states of particles 1 and 2 coincide with those of particles 3 and 4. Perturbative unitarity breaks down if the $|A^{\text{loop}}_{2\rightarrow 2}|$ required by this (perturbative) form of the unitarity condition is larger than the $|A^{\text{tree}}_{2\rightarrow 2}|$ contribution.

choose to be the solid angle $d^2\hat{q}_1$. In the CoM frame, this becomes,

$$
\int d\Pi_2 = \int \frac{d^2\hat{q}_1}{(4\pi)^2} \frac{|\hat{q}_1|}{\omega_s} = \int \frac{d^2\hat{q}_1}{4\pi} \mathcal{N}_2
$$

(3.10)

where $|\hat{q}_1|=\sqrt{(\omega_s/2)^2 - m^2}$ is fixed by momentum conservation, and we define the constant factor arising from this phase space volume as $\mathcal{N}_2 = |\hat{q}_1|/4\pi \omega_s$. The unitarity condition (3.2), which contains this angular integral, therefore relates $\mathcal{A}$ at different kinematics (different scattering angles), making it a functional relation which is difficult to analyse.

However, while we have accounted for translational invariance (factoring out an overall $\delta^4(\sum_j p_j)$) and Lorentz boosts (setting $p_s = 0$), we have yet to exploit rotational invariance. This is achieved by transforming the angle $\theta$ to an angular momentum $\ell$ (which is conserved by the interactions in $\hat{T}$), resulting in the the well-known partial wave expansion,

$$
\mathcal{A}(\omega_s, \theta) = \mathcal{N}_a \sum_{\ell} a_\ell(\omega_s) f_\ell(\theta)
$$

(3.11)

where $\mathcal{N}_a$ is an overall normalisation, and the mode functions $f_\ell(\theta) = (2\ell + 1) P_\ell(\cos \theta)$ are the longitudinal analogue of Fourier modes\textsuperscript{20}. Using (3.11) in (3.2) gives,

$$
2 \text{Im} a_\ell(\omega_s) = \mathcal{N}_a \mathcal{N}_2 |a_\ell(\omega_s)|^2.
$$

(3.14)

\textsuperscript{20}The mode functions $f_\ell(\theta)$ form a complete orthonormal basis,

$$
\frac{1}{2} \int_0^{\pi} d\theta \sin \theta f^*_\ell(\theta)f_\ell'\theta = (2\ell + 1) \delta_{\ell_1 \ell_2}.
$$

(3.12)

and satisfy an addition formula,

$$
\int \frac{d^2\hat{q}}{4\pi} f_\ell(\hat{p}_1 \cdot \hat{q}) f^*_\ell(\hat{p}_3 \cdot \hat{q}) = \delta_{\ell_1 \ell_2} f_\ell(\hat{p}_1 \cdot \hat{p}_3).
$$

(3.13)
Thanks to rotational invariance, the \( \hat{S} \)-matrix is diagonal when written in \( \ell \)-space, and this has removed the angular integral appearing in the unitarity condition (3.2). Choosing the overall normalisation to be \( N_a = 2/N_2 \) gives a particularly simple condition on the \( a_\ell \) coefficients: since \( |a_\ell| \geq \text{Im} a_\ell \),

\[
|a_\ell(\omega_s)| \leq 1 \Rightarrow |\text{Re} a_\ell(\omega_s)| \leq 1/2 \, ,
\]

and so every partial wave coefficient is bounded by unity.

For example, for the interaction \( C_{24} (\partial \phi)^4/f_\phi^4 \), only two partial wave coefficients are non-zero at tree-level,

\[
a_0(\omega_s) = \frac{5C_{24}}{6} \frac{\omega_s^4}{4\pi f_\phi^4} \, , \quad a_2(\omega_s) = \frac{C_{24}}{30} \frac{\omega_s^4}{4\pi f_\phi^4} \, .
\]

The unitarity condition (3.15) applied to these tree-level partial waves defines a maximum energy above which the perturbative unitarity is lost (when the loop shown in Figure 3 becomes large). Although we have worked in a particular frame, with \( |p_s| = 0 \), Lorentz symmetry allows us to replace \( \omega_s^2 \to s \) and translate these bounds to a general frame,

\[
s^2 \lesssim \frac{3}{5} \frac{4\pi f_\phi^4}{|C_{24}|} \, .
\]

This is the numerically precise version of (3.8) when \( n = 2 \).

In summary, in a Lorentz-invariant theory one is free to work in the centre-of-mass frame with \( p_s = 0 \), in which the partial wave expansion,

\[
A(\omega_s, \theta) = \frac{16\pi \omega_s}{\sqrt{\omega_s^2 - 4m^2}} \sum_\ell a_\ell(\omega_s)(2\ell + 1)P_\ell(\cos \theta) \, ,
\]

contains coefficients \( a_\ell(\omega_s) \) which are bounded by unitarity (3.15). Using (3.12) we can extract each of these coefficients, for instance when the masses are negligible,

\[
a_\ell(\omega_s) = \frac{1}{32\pi} \int_{-1}^{1} d\cos \theta \, P_\ell(\cos \theta) \, A(\omega_s, \theta) \, ,
\]

and this allows one to place unitarity constraints on parameters appearing in the EFT.

### 3.2 EFTs with Broken Boosts

In this section we discuss how to impose unitarity in a scalar field theory with broken boosts. Due to the lack of boost symmetry, the \( n \to n \) scattering amplitude now depends on the center-of-mass velocity \( \rho_s \), and in general there is both an \( s_{\text{max}} \) and a \( |p_s|_{\text{max}} \) above which perturbative unitarity is lost. In particular, since boosts are broken by some preferred time-like direction \( n^\mu \), the total momentum \( p_s \) can no longer be freely set to zero—this introduces a preferred spatial direction, as shown in figure 1, and for \( 2 \to 2 \) a more general partial
wave expansion is required since the angular momentum of individual particles is no longer conserved.

For computation, it is often most convenient to express amplitudes as an overall power of \( \omega_s \) (the total energy) multiplying a dimensionless function of \( \rho_s \) (the centre-of-mass velocity) or \( \gamma_s = 1/\sqrt{1 - p_s^2} \) (the centre-of-mass Lorentz factor). However, since it is not particularly transparent physically what a “maximum velocity” means for an EFT\(^{21}\), we will quote our final bounds in terms of \( s \) (the centre-of-mass-frame energy) and \(|p_s|\) (the centre-of-mass momentum), where

\[
\omega_s^2 = s + c_r^2 |p_s|^2 = \gamma_s^2 s. \tag{3.20}
\]

\( s \) is the internal energy that the system has (in its “rest frame”, when \( p_s = 0 \)), while \( c_r |p_s| \) is the energy the system has as a result of its overall motion. Unitarity will place maximum thresholds on both \( s \) and \(|p_s|\), which are readily interpreted as a bound on how strongly the particles can interact with each other (their relative energy, \( s \)) and a bound on how strongly the particles can interact with the background (their \(|p_s|\) relative to \( n^\mu \)).

**Elastic \( n \to n \) Scattering**

The \( n \)-particle phase space is again given by integrating over all future-pointing, on-shell 4-momenta which conserve the total ingoing momenta\(^{22}\),

\[
d\Pi_n = (2\pi)^4 \sqrt{-Z} \delta^4(p_s - \sum_i q_i) \prod_{i=1}^n \frac{d^4q_i}{(2\pi)^4 \sqrt{-Z}} 2\pi \delta(\mathcal{E}(q_i)) \Theta(q_i^0) \tag{3.21}
\]

where \( \mathcal{E}(q) = Z^{\mu\nu} q_\mu q_\nu + m^2 \) is the classical equation of motion. Since this is invariant under rotations of \( q_\mu \) which preserve the kinetic matrix \( Z^{\mu\nu} \), we can again estimate the phase space volume, \( d\Pi_n \sim (4\pi)^{-1} (s/16\pi^2)^{n-2} \), where now \( s = -Z^{\mu\nu} p_s \mu p_s \nu \). Then the unitarity condition (3.2) implies the same upper bound (3.6) on the amplitude, providing \( A_{2\to2} \) is computed using \( Z^{\mu\nu} \) as an effective metric.

For the general interaction \( C_{abc}(n^\mu \partial_\mu)\gamma^2(Z^{\mu\nu} \partial_\mu \partial_\nu)^a \pi^b / f^{2a+b+c-4} \), the coefficient \( C_{abc} \) can be constrained by estimating the tree-level amplitude from \( V \) such vertices,

\[
|A_{n\to n}^{\text{tree}}| \sim \left( \frac{C_{abc} \omega^2 s^q}{f^{2a+b+c-4}} \right)^V \frac{1}{8^{V-1}} \tag{3.22}
\]

where the number of external legs is \( 2n = (b - 2)V + 2 \) and we have neglected masses since we are interested in high energy \( (s \gg 4m_s^2) \) behaviour of this amplitude. The schematic

\(^{21}\)One might argue that the natural maximum for \( \rho_s \) should be \( c_s \), since an intermediate state with \( \rho_s > c_s \) can be transformed using a \( c_T \)-preserving Lorentz boost to one with a negative invariant mass squared (which should be unstable). However, since the \( c_T \)-preserving Lorentz boosts are only a symmetry of theory in the UV, where \( \pi \) is no longer the correct degree of freedom, it seems premature to conclude that there is any issue with having \( \rho_s > c_s \) in the EFT. We will come back to this at the end of section 3.3.

\(^{22}\)The constant factors of \( \sqrt{-Z} \) arise from the Fourier transform, \( f(p) = \int d^4x \sqrt{-Z} e^{ip_x x^\mu} f(x) \)
unitarity bound (3.6) then requires that,
\[
s^{(2a+b+c-4)/2} \left(1 + \frac{|p_s|^2}{s}\right)^{c/2} \lesssim (4\pi)^{b-2-1/V} \frac{f_{\pi}^{2a+b+c-4}}{|C_{abc}|}. \tag{3.23}
\]
When the number of time derivatives $c \neq 0$ the range of allowed interaction energies $s$ depends on the centre-of-mass motion $|p_s|$. The largest possible $s$ is achieved when $|p_s| = 0$, and coincides with the $s_{\text{max}}$ result expected from the Lorentz invariant case (because $n^\mu \partial_\mu$ is now also $\propto \sqrt{s}$ in this frame). Now, at lower values of $s$, there is always a maximum $|p_s|$ at which unitarity is violated: there is a separate strong coupling scale for the centre-of-mass motion (which can also be viewed as a maximum velocity) above which perturbative unitarity is violated.

A complementary way to view this scattering is to change to coordinates in which $p_s = 0$, but since the system is no longer Lorentz-invariant this change affects the interaction strengths. In particular, the transformation to centre-of-mass coordinates results in a change in $n_\mu$,
\[
p_{s\mu} = \begin{pmatrix} \omega_s \\ p_s \end{pmatrix} \rightarrow p_{s'\mu} = \begin{pmatrix} \sqrt{s} \\ 0 \end{pmatrix} \Rightarrow n_\mu = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \rightarrow n_{\mu}' = \begin{pmatrix} \omega_s/\sqrt{s} \\ -p_s/\sqrt{s} \end{pmatrix}, \tag{3.24}
\]
shown in Figure 1. Intuitively, the faster the whole system is moving with respect to the background, the larger the couplings induced by the background (since $|p_s| \gg s$ corresponds to $n_\mu$ having large components). The system cannot move at arbitrarily high centre-of-mass velocities because the interaction strengths cannot be arbitrarily large without violating unitarity.

Whenever $2a + b \geq 4$, the largest possible $|p_s|$ is achieved when $s \sim 4m_\pi^2$ is a minimum,
\[
(|p_s|_{\text{max}})^c \sim f_\pi^c \frac{(4\pi)^{b-2-1/V}}{|C_{abc}|} \left(\frac{f_\pi}{m_\pi}\right)^{2a+b-4}, \tag{3.25}
\]
with $V = 1$ when $b$ is even and $V = 2$ when $b$ is odd, and where typically $f_\pi/m_\pi \gg 1$ (since $f_\pi$ controls the EFT cutoff $\Lambda_{\text{max}}$). However, something interesting happens when $a = 0$, $b = 3$: then unitarity demands that $|p_s|$ be small at both large and small $s$, reaching a maximum only at an intermediate value $m_\pi^2 \ll s \ll f_\pi^2$ (this is shown in Figure 5). For interactions of the form $(n^\mu \partial_\mu)^c \pi^3$, at a fixed $|p_s|$ there is always a minimum $s$ at which unitarity is violated, given by (3.23)—for example, when $|p_s| \lesssim f_\pi$ this is approximately,
\[
s \gtrsim f_\pi^2 \left(\frac{|p_s|}{f_\pi}\right)^{2c} \frac{|C_{03c}|^2}{4\pi}. \tag{3.26}
\]
Unless all $|C_{03c}| \lesssim m_\pi/f_\pi$ are suppressed (bringing this minimum $s$ below $4m_\pi^2$), scattering with low $s$ will violate perturbative unitarity. Physically, this is because a large $|p_s|$ generates a large $\pi^3$ vertex, which allows a propagating $\pi$ particle to emit a single off-shell $\pi$. As
s is decreased this emission becomes longer lived and contributes more to the scattering amplitude, until eventually (perturbative) unitarity is lost. The reason this effect is not observed for higher-point interactions is that their \( n > 2 \) particle phase space volume shrinks as \( s \) decreases, and so the two competing effects (longer-lived emissions versus smaller available phase space) cancel out and unitarity is not violated. However, we stress that we have only considered forward limit scattering in which there are no hierarchies between the incoming momenta, i.e. all \( n^\mu \partial_\mu \sim \omega_s \) and \( Z^{\mu
u} \partial_\mu \partial_\nu \sim s \). A more detailed study of higher-\( n \)-point unitarity bounds, in which say the total \( s \) is held fixed (so the phase space volume is fixed) but one of the channels \( s_{ij} \ll \vert \mathbf{p}_s \vert^2 \) (allowing for production of a long-lived intermediate), may lead to a similar IR cutoff on \( s \) for any interaction of the form \( (n^\mu \partial_\mu)^n \pi^b \). We will not pursue that further here, and instead focus on the \( 2 \to 2 \) unitarity bound and its consequences for a \( \pi^3 \) interaction.

**Elastic 2 \to 2 Scattering**

A scattering process between four on-shell momenta with \( \sum_j p_{j\mu} = 0 \) in a theory with broken boosts can depend on at most five independent kinematic variables: the two Lorentz invariant variables \( \{ s, t \} \), plus now we have \( n^\mu \) with which to extract three energies, \( \{ n^\mu p_{1\mu}, n^\mu p_{2\mu}, n^\mu p_{3\mu} \} \).

The analogue of the \( (\omega_s, \theta) \) variables from the Lorentz-invariant problem are \( (\omega_s, \rho_s, \theta_1, \theta_3, \phi) \) where \( \cos \theta_j \) is the angle between \( \mathbf{p}_j \) and \( \mathbf{p}_s \), and \( \phi_{13} = \phi_1 - \phi_3 \) where \( \phi_j \) is the azimuthal angle about \( \mathbf{p}_s \). To convert an amplitude \( \mathcal{A}(p_{1\mu}, p_{2\mu}, p_{3\mu}, p_{4\mu}) \) to these variables, first use momentum conservation to set \( p_{3\mu}^s = -p_{1\mu}^s - p_{2\mu}^s \) and \( p_{4\mu}^s = -p_{3\mu}^s - p_{2\mu}^s \), then the remaining momenta (using two spatial rotations to align the \( \hat{z} \) axis along \( \mathbf{p}_s \)) are,

\[
\mathbf{p}_s = |\mathbf{p}_s| \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \quad \mathbf{p}_1 = |\mathbf{p}_1| \begin{pmatrix} \sin \theta_1 \cos \phi_1 \\ \sin \theta_1 \sin \phi_1 \\ \cos \theta_1 \end{pmatrix}, \quad \mathbf{p}_3 = |\mathbf{p}_3| \begin{pmatrix} \sin \theta_3 \cos \phi_3 \\ \sin \theta_3 \sin \phi_3 \\ \cos \theta_3 \end{pmatrix},
\]

where the energies are fixed by the four mass-shell conditions as\(^{23}\),

\[
\omega_j = c_s |\mathbf{p}_j| = \frac{\omega_s}{2} \frac{1 - \rho_s^2 \cos \theta_j}{1 - \rho_s \cos \theta_j},
\]

\(^{23}\)We have neglected the small mass in (3.29). With finite mass, \( \omega_j = \sqrt{c_s^2 |\mathbf{p}_j|^2 + m_s^2} \), with

\[
c_s |\mathbf{p}_j| = \frac{\omega_s}{2} \frac{1 - \rho_s^2}{1 - \rho_s \cos \theta_j} \frac{\sqrt{1 - 4m_s^2} \frac{1 - \rho_s^2 \cos^2 \theta_j}{(1 - \rho_s^2)^2} + \rho_s \cos \theta_j}{1 + \rho_s \cos \theta_j}.
\]

The mass is unimportant providing the scattering energy is sufficiently large, \( s > 4\gamma_s^2 m_s^2 \), where \( \gamma_s = 1/\sqrt{1 - \rho_s^2} \) (else the mass forbids large angles, where the square root becomes imaginary). We will always work with \( s \gg m_s^2 \), and hereafter any \( \gamma_s \gg 1 \) limit is always taken with the understanding that it remains smaller than \( s/4m_s^2 \), i.e. \( 1 + \vert \mathbf{p}_j \vert^2/s < 4m_s^2 \gg 1 \). This was not needed for our discussion of \( n \to n \) scattering because that focussed solely on \( \theta_j \approx 0 \) (so that \( \omega_j \approx \omega_s \)).
and finally use the remaining spatial rotation (around \( \mathbf{p}_s \)) to set \( \phi_1 + \phi_3 = 0 \). Note that in the CoM frame, \( \rho_s \to 0 \), all particles have the same energy, \( c_s |\mathbf{p}_j| = \omega_s / 2 \), and in this limit the three angles can appear only in the combination, \( \hat{\mathbf{p}}_1 \cdot \hat{\mathbf{p}}_3 = \cos \theta_1 \cos \theta_3 + \sin \theta_1 \sin \theta_3 \sin \phi_3 \equiv \cos \theta \), which is the usual relativistic scattering angle appearing in the Mandelstam invariants,

\[
s_{12} = \omega_s^2 (1 - \rho_s^2) , \quad s_{13} = 2 \omega_1 \omega_3 (1 - \cos \theta) . \tag{3.30}
\]

Performing integrals over the six \( \delta \)-functions in \( d\Pi_2 \) leaves an integral over just two angular variables, which we choose to be the solid angle \( \cos \theta \) and the energy \( E_{q_1} \) fixed as in (3.29) in terms of the the angle between \( \hat{\mathbf{q}}_1 \) and \( \mathbf{p}_s \), and we have defined the angle-dependent factor \( N_2(\mathbf{q}_1) = 2E_{q_1}^2 / 4\pi s \).

We now seek an angular momentum expansion of the non-relativistic scattering amplitude analogous to the partial-wave expansion (3.18) adapted to a general \( \mathbf{p}_1 + \mathbf{p}_2 \neq 0 \) frame. Since the amplitude now depends on both longitudinal and azimuthal angles the expansion should naturally contain both the total angular momentum \( L^2 \) (eigenvalues \( \ell_j (\ell_j + 1) \)) as well as its projection \( L_z \) (eigenvalues \( m_j \)) for each particle,

\[
\mathcal{A}(p_1 p_2 \to p_3 p_4) = N_a \sum_{\ell_1, \ell_3, m_1, m_3} a^{n_1 m_3}_{\ell_1 \ell_3}(\omega_s, \rho_s) Y^{m_1}_{\ell_1}(\hat{\mathbf{p}}_1) Y^{m_3*}_{\ell_3}(\mathbf{p}_3) \tag{3.33}
\]

where \( N_a \) is again an overall normalisation, and the spherical harmonics \( Y^{m}_{\ell} \) are a complete set of orthonormal mode functions on the sphere,

\[
\int d^2 \hat{\mathbf{q}} Y^{m_1}_{\ell_1}(\hat{\mathbf{q}}) Y^{m_2*}_{\ell_2}(\hat{\mathbf{q}}) = \delta_{\ell_1 \ell_2} \delta_{m_1 m_2} . \tag{3.34}
\]

As shown in Figure 1, even once \( n^\mu \) is fixed (which uses all three boosts) and the direction of \( \mathbf{p}_s \) is fixed (which uses two rotations), there is still one rotational freedom left – this imposes a selection rule on the spherical wave coefficients, \( a^{m_1 m_3}_{\ell_1 \ell_3} \sim \delta_{m_1 m_3} \) (since angular momentum along \( \mathbf{p}_s \) is conserved). We will write,

\[
a^{m_1 m_3}_{\ell_1 \ell_3} = \delta_{m_1 m_3} (a^{m_1}_{\ell_1 \ell_3}) \tag{3.35}
\]

and refer to \( a^{m_1} \) as a matrix with indices \( \{ \ell_1, \ell_3 \} \) running from \( m_1 + 1 \) onwards. Finally, our

\[\text{We note for completeness that for general masses the phase space element in a frame with } \rho_s \neq 0 \text{ is given by,}
\]

\[
d\Pi_2 = \frac{d^2 \hat{\mathbf{q}}_1}{(4\pi)^2} \frac{c_s |\mathbf{q}_1|}{\omega_s} \left( 1 - \frac{\rho_s E_{q_1}}{c_s |\mathbf{q}_1|} \hat{\mathbf{p}}_s \cdot \hat{\mathbf{q}}_1 \right)^{-1} = \frac{d^2 \hat{\mathbf{q}}_1}{4\pi} N_2(\mathbf{q}_1) , \tag{3.31}
\]

where \( E_{q_1} = \sqrt{c_s^2 |\mathbf{q}_1|^2 - m^2} \) and \( |\mathbf{q}_1| \) is fixed as in (3.28) in terms of the angle \( \hat{\mathbf{p}}_s \cdot \hat{\mathbf{q}}_1 \).
spherical wave coefficients \( a_{\ell_1 \ell_3}^{m_1 m_3}(\omega_s, \rho_s) \) in a general frame are related to the usual centre-of-mass coefficients \( a_\ell(\omega_s) \) by\(^{25}\),

\[
a_\ell(\omega_s) = \frac{1}{2\ell + 1} \sum_{m=-\ell}^{+\ell} a_{\ell \ell}^m(\omega_s, 0), \tag{3.36}
\]

which is effectively averaging over all possible \( L_z \) projections when particles 1 and 3 have fixed total angular momentum \( \ell (\ell + 1) \).

Now that we are equipped with the spherical-wave expansion (3.33), we can apply the unitarity condition (3.2) to scattering amplitudes with no boost-invariance. Explicit substitution of (3.33) into (3.2) yields the condition,

\[
\text{Disc}_s a_{\ell_1 \ell_3}^{m_1 m_3}(\omega_s, \rho_s) \geq \sum_{k,j} a_{\ell_1 j}^{m_1 k}(\omega_s, \rho_s) a_{\ell_3 j}^{m_3 k*}(\omega_s, \rho_s), \tag{3.37}
\]

where we have taken care to fix the overall normalisation, \( N_a = 8\pi/\sqrt{N_2(\hat{p}_1) N_2(\hat{p}_3)} \), such that the angular integral from \( \int d\Pi_2 \) becomes simply (3.34). In terms of the matrix \( a^m \),

\[
\frac{1}{2\ell} (a^m - a^{m\dagger}) \geq a^m a^{m\dagger}, \tag{3.38}
\]

and our unitarity condition (3.37) can be seen as taking a particular matrix element of (3.1) (in a basis of 2-particle spherical wave states rather than a 2-particle plane wave state [23]).

Focussing on a particular diagonal element of \( a^m \), we have,

\[
\text{Im} a_{\ell \ell}^m \geq \sum_j |a_{\ell j}^m|^2 \geq |a_{\ell \ell}^m|^2 \tag{3.39}
\]

and hence an analogous boundedness to (3.15), namely \( |\text{Re} a_{\ell \ell}^m(\omega_s, \rho_s)| < 1/2 \), which now holds at any centre-of-mass velocity. In fact, we show in Appendix A.1 that this extends to any minor of \( a^m \),

\[
|\det a_{\text{sub}}^m| < 1 \tag{3.40}
\]

where \( a_{\text{sub}}^m \) is any (finite) submatrix of \( a^m \). We will focus on the simplest bound (3.39), and leave further exploration of the (infinite number of) unitarity bounds (3.40) for the future.

In Appendix A.2, we show how to systematically evaluate the angular integrals in (3.51) for any \( A_{2\rightarrow 2}^{\text{tree}} \) (which is analytic up to simple poles in \( s, t \) and \( u \)). There are three particular features which seem to apply quite generally,

\[\text{This follows from the Legendre addition formula (3.13), which can be written as:}\]

\[
P_\ell(\hat{p}_1 \cdot \hat{p}_3) = \frac{4\pi}{2\ell + 1} \sum_m Y_\ell^m(\hat{p}_1) Y_{\ell^*}^{m*}(\hat{p}_3).\]
At small $\rho_s$, the strongest bound is from low $\ell$ modes.

At large $\rho_s \to 1$, the strongest bound is from high $\ell$ modes.

There is an apparent (at most dilogarithmic) branch cut at $\rho_s = \pm 1$, which arises from an (anti)collinear singularity and would be resolved by finite mass effects. Since we assume that $s \gg 4m^2_\pi$, we never approach this cut.

For example, consider the single interaction $\beta_1 \pi^4 / f^4_\pi$. The corresponding scattering amplitude is $A = 24\beta\omega_1 \omega_2 \omega_3 \omega_4 / f^4_\pi$, and since this does not depend on $\phi_{13}$ the only non-zero spherical-wave amplitudes are those with $m_1 = m_3 = 0$. Furthermore, since $A$ can be written in the form $f(\omega_1) f(\omega_3)$, it has the convenient property that the spherical-wave amplitudes factorise into a product of two angular integrals,

$$a^{00}_{\ell_1 \ell_3}(\omega_s, \rho_s) = \beta_1 \frac{\omega^4_1}{f^4_\pi} \hat{I}_{\ell_1}(\rho_s) \hat{I}^*_{\ell_3}(\rho_s)$$

which are given in (A.32) and plotted in Figure 4. Perturbative unitarity then requires that,

$$\beta_1 \frac{\omega^4_s}{f^4_\pi} \sum_\ell |\hat{I}_{\ell}(\rho_s)|^2 \leq 1 .$$

At small $\rho_s$ the constraining power comes mostly from $\ell = 0$, but as $\rho_s$ is increased the higher $\ell$ modes begin to dominate the sum. In particular, when $\rho_s \to 0$ only the first spherical-wave coefficient,

$$a^{00}_{00}(\omega_s, \rho_s) = \frac{3\beta_1}{32\pi} \frac{\omega^4_s}{f^4_\pi} (1 - \rho^2_s)$$

is non-zero, as required by (3.36). But as $\rho_s$ approaches 1 the higher order spherical-waves become important and the integrals take the form,

$$\hat{I}_{\ell}(\rho_s) = \sqrt{\frac{2\ell+1}{4\pi}} \frac{\sqrt{3}}{2\gamma_s} \left[ 1 + \frac{2\ell^2}{\gamma_s^2} \log \left( \frac{\ell}{2\gamma_s} \right) + \mathcal{O} \left( \frac{\ell}{\gamma_s^2}, \frac{\ell^2}{\gamma_s^3} \right) \right] ,$$

where $\gamma_s = 1/\sqrt{1 - \rho^2_s}$ and we have assumed that $1 \ll \ell < \gamma_s$ in order to perform an expansion of (A.14). By summing only up to $\ell \approx \gamma_s / \sqrt{2}$, we can discard all but the first term of (3.44) and write a simple approximation for the sum when $\gamma_s$ is large, $\sum_\ell |\hat{I}_{\ell}|^2 \approx 3/64\pi$. Unitarity (3.42) is therefore violated if $s$ exceeds the $s_{\text{max}}$ threshold defined by (3.43) at small $|p_s|$, or if $|p_s|$ exceeds the $|p_s|_{\text{max}}$ threshold defined using (3.44) at small $s$,

$$s_{\text{max}} = 4f^2_\pi \left( \frac{2\pi}{3\beta_1} \right)^{1/2} c_{\pi} |p_s|_{\text{max}} = 2f^2_\pi \left( \frac{4\pi}{3\beta_1} \right)^{1/4} .$$

The full region of $\{ s, |p_s| \}$ compatible with perturbative unitarity is shown in Figure 5.
Figure 4. Properties of the $\pi^4$ spherical-wave amplitudes. Left: The ratio $a_{\ell\ell}^{00}/a_{00}^{00}$ is plotted against $\gamma_s = 1/\sqrt{1 - \rho_s^2}$ for $L = 3, 10$ and $30$, and gray grid lines show $\gamma_s = 3, 10$ and $30$. At small $\gamma_s$ it is $a_{00}^{00}$ which dominates, while at large $\gamma_s$ the $a_{\ell\ell}^{00}$ amplitude is larger by a factor of $2\ell + 1$. The $\gamma_s$ at which $a_{\ell\ell}^{00}$ first exceeds $a_{00}^{00}$ scales $\propto \ell$, so terms with $\ell \gg \gamma_s$ can always be neglected from the sum in (3.42). Right: The angular integral $I_\ell$ is plotted against $\ell$ at fixed $\gamma_s = 3, 10, 30, 100$ and $300$. In general they display a large positive maximum at a finite $\ell$ near to $\gamma_s$, which is followed by a smaller negative minimum before they approach zero from below as $\ell \to \infty$ (sufficiently fast for the spherical wave expansion (3.50) to converge). Black dashed lines show the approximation (3.44) for $\gamma_s = 100$ and $\gamma_s = 300$, agreeing well with $I_\ell$ at sufficiently small $\ell < \gamma_s/\sqrt{2}$.

Now consider the cubic interaction, $\alpha_1 \pi^3/f_\pi^4$. The $t$- and $u$-channel exchange poles in $A$ lead to more complicated expressions for the $a_{\ell_1\ell_3}^{m_1m_3}$—in particular, they are non-zero for all $m_1 = m_3$ and do not factorise like (3.41)—however they still exhibit the same three features as the $\beta_1$ amplitude above (and can be found in the Appendix). In particular, in the low $\rho_s$ regime only the first spherical-wave coefficient,

$$a_{00}^{00}(\omega_s, \rho_s) = -\frac{3\alpha_1^2 \omega_s^4}{64\pi f_\pi^4} (3 + 2\rho_s^2)$$

is non-zero, again consistent with (3.36). In the high $\rho_s$ regime,

$$a_{\ell_1\ell_3}^{m_1m_3}(\omega_s, \rho_s) = -\frac{15\alpha_1^2 \omega_s^4}{64\pi f_\pi^4} \sqrt{2\ell_1 + 1} \sqrt{2\ell_3 + 1} \left( \frac{\delta_{m_1}}{\delta_{0}} \frac{\delta_{m_3}}{\delta_{0}} + O \left( \frac{\ell_1}{\gamma_s}, \frac{\ell_3}{\gamma_s} \right) \right).$$

When $\ell_1 = \ell_3$, the first subleading term is $\frac{12}{n} \ell^2/\gamma_s^2 \log (\ell/\gamma_s)$, so once again we can approximate (3.37) by truncating the sum at $\ell^2 \approx \frac{5}{12} \gamma_s^2$ and using only the leading term in (3.47), which gives $\sum_j a_{\ell j} a_{\ell j}^* / a_{\ell \ell} \approx -\frac{15\alpha_1^2 \omega_s^4}{64\pi f_\pi^4} \frac{5\gamma_s^2}{12}$ when $\gamma_s$ is large. Unlike for the quartic interaction, as $s$ is made small at fixed $|p_s|$ perturbative unitarity becomes a stronger constraint: arbitrarily small $s$ are not allowed, since demanding $\frac{15\alpha_1^2 \omega_s^4}{64\pi f_\pi^4} \frac{5\gamma_s^2}{12} < 1$ requires,

$$s \geq \frac{25}{64} \frac{f_\pi^2}{|p_s|} \left( \frac{|p_s|}{f_\pi} \right)^6 \frac{\alpha_1^2}{4\pi}$$

(3.48)
Figure 5. The values of $s$ and $|p_s|$ consistent with perturbative unitarity for a $\pi^4/f_\pi^4$ interaction (left) and a $\pi^3/f_\pi^2$ interaction (right). Gray grid lines show $\gamma_s = 3, 10$ and 30—the spherical waves with $\ell \approx \gamma_s$ are most responsible for the constraints. The cubic vertex has the special feature that at fixed $|p_s|$ there is a minimum $s$, and the black dashed line shows the $s^{1/6}$ scaling expected from (3.48). The maximum $s_{\text{max}}$ and $|p_s|_{\text{max}}$ are given in (3.45) with $\alpha_1 = 1$ and in (3.49) with $\beta_1 = 1$ respectively—note that the actual maximum $|p_s|$ for the cubic vertex is set by $\ell \approx 1$ spherical-waves and is lower than the upper bound in (3.49) from the $\ell = 0$ wave.

when $\gamma_s \approx |p_s|^2/s \gg 1$. This agrees with (3.26), anticipated earlier using dimensional analysis, now with a numerical prefactor which has been determined by performing the $d\Pi_2$ integral explicitly using the spherical-wave expansion. The full region of $\{s,|p_s|\}$ compatible with perturbative unitarity is shown in Figure 5. The maximum $|p_s|$ is no longer reached at small $s$, but rather at some intermediate scale, and cannot be larger than that allowed by\(^{26}\) (3.46), which can be used to infer both the maximum allowed energy and a maximum possible momentum,

$$s_{\text{max}} = 4f_\pi^2 \left( \frac{2\pi}{9\alpha_1^2} \right)^{1/2} |p_s|_{\text{max}} \leq 2f_\pi \left( \frac{2\pi}{15\alpha_1^2} \right)^{1/4}. \quad (3.49)$$

In summary, in a general frame in which $p_s \neq 0$, the spherical wave expansion\(^{27}\),

$$A(\omega_s, \rho_s, \theta_1, \theta_3, \phi_{13}) = \frac{16\pi^2 \omega_s^2 (1 - \rho_s^2)}{\omega_1 \omega_3} \sum_{\ell_1, \ell_3} \sum_{-\ell_3 < m_3 < \ell_3} \sum_{-\ell_1 < m_1 < \ell_1} a_{\ell_1 \ell_3}^{m_1 m_3}(\omega_s, \rho_s) Y_{\ell_1}^{m_1}(\hat{p}_1) Y_{\ell_3}^{m_3}(\hat{p}_3),$$

\(^{26}\)The $a_{000}^{00}$ mode can provide this upper bound on $|p_s|_{\text{max}}$ at small $s$ since it is a monotonic function of $\rho_s$, while the higher $\ell$ modes are not.

\(^{27}\)(3.50) has used the small mass expression for $N_2$ given in (3.32), but more generally if (3.31) is used in (3.33) with $N_a = 8\pi/\sqrt{N_2(\hat{p}_1)N_2(\hat{p}_3)}$ then the unitarity bound (3.37) holds for any finite mass.
contains coefficients $a_{t_1t_3}^{m_1m_3} = \delta_{m_1m_2}(a^m)_{t_1t_3}$ in which all minors of $a^m$ are bounded as in (3.40). Using (3.34), we can extract each of these coefficients,

$$a_{t_1t_3}^{m_1m_3}(\omega_s, \rho_s) = \frac{1}{1 - \rho_s^2} \int \frac{d^2 p_1}{4\pi} \frac{d^2 p_3}{4\pi} Y_{m_1}^{*}(p_1) Y_{m_3}^{*}(p_4) \frac{\omega_1}{\omega_s} \frac{\omega_3}{\omega_s} A(p_1p_2 \to p_3p_4),$$  

(3.51)

and thus translate the unitarity condition (3.40) into bounds on EFT coefficients. In Appendix A.2, we show how to systematically evaluate the angular integrals in (3.51) for any $A_{2 \to 2}^{\text{tree}}$ (which is analytic up to simple poles in $s, t$ and $u$). We will now apply these new spherical-wave unitarity bounds to the EFT of Inflation.

3.3 EFT of Inflation

Since in the decoupling limit the EFT of Inflation becomes a simple theory of a single scalar field with a preferred direction, $n^\mu$, we can use the unitarity constraints developed above to constrain the Wilson coefficients $\alpha_1, \beta_1$ of (1.3). In particular, the momentum space scattering amplitude mediated by $S[\pi]$ is,

$$f_\pi^4 A(p_1p_2 \to p_3p_4) = 24\beta_1 \omega_1 \omega_2 \omega_3 \omega_4 + 2\beta_2 \big( \omega_1^2 s_{12}^3 + \omega_2^3 s_{12} + \omega_3^2 s_{13} + \omega_4^2 s_{23} \big) + 2\beta_3 \big( s_{12}^2 + s_{13}^2 + s_{23}^2 \big)$$

$$+ \frac{1}{s_{12}} \left( 6\alpha_1 \omega_1^2 \omega_2 \omega_{12} + 2\alpha_2 \omega_2 s_{12} \right) \left( 6\alpha_1 \omega_3 \omega_4 \omega_{34} + 2\alpha_2 \omega_{34} s_{34} \right)$$

$$+ \frac{1}{s_{13}} \left( 6\alpha_1 \omega_1 \omega_3 \omega_{13} + 2\alpha_2 \omega_3 s_{13} \right) \left( 6\alpha_1 \omega_2 \omega_4 \omega_{24} + 2\alpha_2 \omega_{24} s_{24} \right)$$

$$+ \frac{1}{s_{23}} \left( 6\alpha_1 \omega_1 \omega_4 \omega_{14} + 2\alpha_2 \omega_4 s_{14} \right) \left( 6\alpha_1 \omega_2 \omega_3 \omega_{23} + 2\alpha_2 \omega_{23} s_{23} \right)$$

(3.52)

where $s_{ij} = -Z^\mu(p_i + p_j)_\mu(p_i + p_j)_\nu, \omega_{ij} = \omega_i + \omega_j$ and $\omega_{ij}^{kl} = \omega_i \omega_j + \omega_k \omega_l$, and we have used momentum conservation and mass-shell conditions (neglecting all terms in $m_\pi$ to be consistent with neglecting $\eta_*$ terms in the effective action).

Scattering in the Inflationary Rest Frame

Let us recall how this amplitude was used in [22, 24]. By enforcing the restriction $p_* = 0$ by hand (i.e. focussing on a particular choice of kinematics), then one may use the partial wave expansion (3.18). Taking the amplitude (3.52) in the frame (3.9), one finds that the only non-zero partial waves are:

$$a_0(\omega_s) = \frac{1}{192\pi} \left( -3(3\alpha_1 + 4\alpha_2)^2 + 18\beta_1 + 24\beta_2 + 40\beta_3 \right) \frac{\omega_4^4}{f_\pi^4}$$

(3.53)

$$a_2(\omega_s) = \frac{\beta_3}{120\pi} \frac{\omega_4^4}{f_\pi^4}.$$  

(3.54)

Since perturbative unitarity is lost when $|\text{Re} a_\ell(\omega_s)|$ first reaches $1/2$, which by definition does not happen at energies below the strong coupling scale, $\omega_s \lesssim 2\Lambda_\phi$, each of these partial waves gives an upper bound on $\Lambda_\phi$. Since $a_2$ depends only on $c_s$, it gives a bound which is
independent of the bispectrum and trispectrum,

\[ s_{\text{max}}^2 = f_\pi^4 \times \frac{480\pi c_s^4}{1 - c_s^2}. \] (3.55)

As described in [22], since the unitarity bound from \(a_0\) mixes \(c_s\) and \(\alpha_1\) with \(\beta_1\) (which is far less constrained observationally), it is essentially always possible to tune \(\beta_1\) such that the bound is satisfied. For example, when \(\beta_1 \gg \alpha_1 \gg 1 - c_s^2\), we can focus on the contributions from \(\alpha_1^2\) and \(\beta_1\) only,

\[ a_0(\omega_s) = \frac{3}{64\pi} \frac{\omega_s^4}{f_\pi^4} \left[ (-3 + 2\rho_s^2)\alpha_1^2 + (2 - 2\rho_s^2)\beta_1 \right], \] (3.56)

and so if \(2\beta_1 \approx 3\alpha_1^2\) then this bound does not lower \(\Lambda_o\) even if \(\alpha_1\) is very large.

However, since the amplitude (3.52) is not Lorentz invariant, focussing on the CoM frame restricts the possible kinematics of the scattering particles unnecessarily. Satisfying (3.15) is therefore necessary but not sufficient for the \(2 \to 2\) scattering processes to respect perturbative unitarity, since we must also demand the unitarity condition (3.2) in other frames. One immediate consequence of this is that although a tuning (such as \(3\alpha_1^2 \approx 2\beta_1\)) always seemed possible to guarantee perturbative unitarity, this is an artefact of CoM scattering – we will now use scattering in a more general frame to show that \(\alpha_1\) and \(\beta_1\) are separately bounded.

**Beyond the Inflationary Rest Frame**

Using the spherical wave expansion (3.50) at an arbitrary centre-of-mass velocity, there is an upper bound on the strong coupling scale for every \(a_{m\ell\ell}^{m}(\omega_s, \rho_s)\) – for every choice of \(m\) and \(\ell\), there is a whole function of \(\rho_s\) which must be bounded. A systematic approach for computing these spherical wave coefficients is described in Appendix A.2. For example, when \(\beta_1 \gg \alpha_1 \gg 1 - c_s^2\), the first spherical wave coefficient is simply,

\[ a_{00}^0(\omega_s, \rho_s) = \frac{3}{64\pi} \frac{\omega_s^4}{f_\pi^4} \left[ (-3 + 2\rho_s^2)\alpha_1^2 + (2 - 2\rho_s^2)\beta_1 \right]. \] (3.57)

Now, a tuning like \(2\beta_1 \approx 3\alpha_1^2\) is no longer enough to guarantee perturbative unitarity of the \(2 \to 2\) scattering amplitude for all centre-of-mass velocities. In fact, since each spherical-wave amplitude gives a separate bound of the form,

\[ |a_{m\ell\ell}^m(\omega_s, \rho_s)| = \frac{1}{4\pi} \frac{\omega_s^4}{f_\pi^4} \left| F_{\ell}^m(\rho_s)\alpha_1^2 + G_{\ell}^m(\rho_s)\beta_1 \right| < \frac{1}{2}, \] (3.58)

where \(F_{\ell}^m\) and \(G_{\ell}^m\) are known functions of \(\rho_s\), unitarity requires that both \(\alpha_1\) and \(\beta_1\) be separately bounded (since there is no tuning for \(\beta_1\) which simultaneously satisfies all bounds).

Since the observational constraints on \(\beta_1\) are much weaker than those on \(\alpha_1\) and \(c_s\), our goal is now to remove \(\beta_1\) from these unitarity bounds. We do this by noting that, for a fixed \(\{\alpha_1, c_s\}\), each spherical wave amplitude implies there is a maximum \(\beta_1^{\max}\) (at which \(a_{m\ell\ell}^m = 1/2\)
and a minimum $\beta_1^{\text{min}}$ (at which $a_{\ell}^n = -1/2$). For example, from (3.57), these maximum and minimum values occur at,

$$\beta_1 = \frac{1}{2(1-\rho_s^2)} \left[ \pm \frac{32\pi}{3} f_\pi^4 \frac{\alpha_1^2}{\omega_s^4} + (3 - 2\rho_s^2)\alpha_1^2 \right]. \quad (3.59)$$

When $\rho_s = 0$, this restricts $\beta_1$ to the range $2\beta_1 = 3\alpha_1^2 \pm \frac{32\pi}{3} f_\pi^4 / 3\omega_s^4$. As $\rho_s$ is increased (at fixed $\omega_s$), this allowed range of $\beta_1$ shifts, until eventually there are no real values of $\beta_1$ which can simultaneously satisfy both bounds. This happens when,

$$\beta_1^{\text{min}}(\omega_s, \rho_s) \geq \beta_1^{\text{max}}(\omega_s, 0) \Rightarrow \omega_s^4 \frac{\rho_s^2}{2 - \rho_s^2} \geq f_\pi^4 \frac{32\pi}{15\alpha_1^2} \quad (3.60)$$

irrespective of $\beta_1$. In particular, when $s$ is made very small this implies that $|p_s|_{\text{max}} \leq 2 f_\pi^4 (2\pi/15\alpha_1^2)^{1/4}$, which agrees with (3.49) since at small $s$ it is the $\alpha_1^2$ interaction which dominates the spherical-wave amplitudes. Comparing $\beta_1^{\text{max}}$ with $\beta_1^{\text{max}}$ in this way for every $a_{\ell}^n$ then provides a set of constraints on $\alpha_1$ independently of $\beta_1$. We also mention in passing that another way to remove $\beta_1$ is to look for bounds (3.58) where its coefficient $G_\ell^n(\rho_s)$ vanishes. As discussed in section 2.2, this happens whenever $m \neq 0$, and also when $m = 0$ for particular values of $\rho_s$ (which depend on $\ell$)—both of these have the potential to constrain $\alpha_1$ independently of $\beta_1$, and could be explored further in future.

While (3.59) and (3.60) are useful illustrations, it is not necessary to assume that $\beta_1 \gg \alpha_1 \gg 1 - c_s^2$. In general, each spherical-wave amplitude is quite lengthy, and has contributions from $\{\alpha_1^2, \alpha_1\alpha_2, \alpha_2^2, \beta_1, \beta_2, \beta_3\}$. But our strategy is the same: from each $a_{\ell}^n$, first identify the maximum/minimum $\beta_1$ at $\rho_s = 0$, and then increase $\rho_s$ at fixed $\omega_s$ until there is no longer any value of $\beta_1$ which can all simultaneously satisfy all bounds. The resulting range of momenta and energies which are consistent with perturbative unitarity is shown in Figure 6 for a few fixed values of $\{c_s, \alpha_1\}$. In general there are maxima at,

$$s_{\text{max}}^2 = 16 f_\pi^4 30\pi c_s^4 \quad \text{and} \quad c_\pi |p_s|_{\text{max}} \leq 2 f_\pi \left( \frac{2\pi}{15\alpha_1^2} \right)^{1/4}, \quad (3.61)$$

as described in (3.55) and (3.60).

Using these spherical-wave bounds, we can now ask precise questions about which regions in $\{c_s, \alpha_1\}$ parameter space can describe subhorizon physics (1.7) (i.e. when does the tree-level scattering mediated by those interactions respect perturbative unitarity). For instance, if we find that the scattering of subhorizon scales (at $c_\pi |p_s|/H \gg |\nu^2 - \frac{1}{4}| \approx 2$) is only unitary if $c_\pi |p_s| < 2H$, then we can conclude that it is never possible for the EFT to describe

\[\text{For } \ell > 2 \text{ there is always at least one zero, e.g., for } \ell = 3 \text{ this is at } \rho_s = 0.715931, \ldots, \text{ and at } \ell = 4 \text{ this is at } \rho_s = 0.842835, \ldots \text{ In fact Figure 4 shows that for any fixed } \rho_s \text{ there is always exactly one value of } \ell \text{ (not necessarily an integer) for which } I_\ell \text{ vanishes.} \]
this regime. Analogously, since both incoming particles have \( \omega_1 > 2H \) and \( \omega_2 > 2H \) (and we must allow for any relative scattering angle), then the total energy \( \omega_s \) must be larger than \( 6H \) to be in the subhorizon regime (1.7).

This can be translated into a theoretical prior: demanding that inflation was approximately single-field (and weakly coupled) on subhorizon scales (1.7) restricts us to the green region shown in Figure 7. We can then translate our allowed region of \( \{c_s, \alpha_1\} \) into an allowed region for bispectrum shapes \( \{f_{NL}^{\text{equil}}, f_{NL}^{\text{orth}}\} \), and multiply a simple Gaussian likelihood (using Planck’s mean and variance) by a prior distribution which is uniform for the allowed \( \{f_{NL}^{\text{equil}}, f_{NL}^{\text{orth}}\} \) (and zero otherwise)—this is described in more detail in Appendix B. This produces the posterior contours shown in Figure 7.

Finally, we note that there are several other ways to investigate perturbative unitarity. We have focussed on the most conservative possible application, identifying the range of parameter space in which subhorizon scattering (approximated by plane waves) could possibly be unitary in perturbation theory. One could further demand that there are no new states (or non-perturbative effects) until at least some scale, say \( 10H \), above the background—this would give an even stronger theoretical prior. Also, rather than focussing on energy/momentum cutoffs relative to the background scale \( H \), one could instead compare the maximum velocity cutoff \( \rho_s^\text{max} \) with the speed of tensors modes: in fact, if the theory is to be perturbatively unitary at all centre-of-mass speeds up to \( c_T \) (which corresponds to \( \rho_s = c_s \)), then,

\[
\frac{\alpha_1^2 \omega_s^4}{4 \pi f^4} \leq \frac{64}{25} \left(1 - c_s^2 \right) \quad \text{when} \quad 1 - c_s^2 \ll 1.
\]

Interestingly, taking \( c_s \to 1 \) and demanding unitarity up to the speed \( c_T \) forces \( \alpha_1 \) to vanish at least as fast as \( 1 - c_s^2 \).

Going beyond the CoM frame has therefore furnished us with a number of new unitarity bounds with interesting consequences for inflation. These give tighter constraints on the Wilson coefficients \( \alpha_3 \) and \( \beta_1 \), and are just the tip of the unitarity iceberg—there are further bounds from higher \( \ell \) coefficients, as well as non-linear constraints coming from the determinant condition (3.40), which we leave for future investigation.

4 Discussion

Demanding consistent scattering amplitudes has proven to be a valuable tool in constraining low-energy Effective Field Theories (EFTs). However, due to their reliance on Lorentz invariance, to date existing techniques have not been fully exploited in cosmology (where the background spacetime spontaneously breaks Lorentz invariance). Here, we have taken the first steps towards extending EFT constraints from radiative stability and unitarity to theories in which boosts are broken. This has allowed us to constrain the dynamics of subhorizon

\footnote{This is the most conservative possible requirement, since in practice if unitarity required \( c_s |p_s| < 2H + \delta \) for some small \( \delta \) then one might still conclude that the subhorizon regime (1.7) is never unitarity—this simply leads to even stronger bounds than those we present in Figure 7.}
modes produced during inflation, and identify the region of parameter space in which these subhorizon scales are approximately single-field and weakly coupled.

We began by showing that the higher-order EFT corrections can be parameterised according to a simple power counting scheme in terms of a single heavy scale \( \Lambda \), a field coupling \( g_\pi \), and an order parameter \( g_n \) which controls the breaking of boosts. Radiative stability, the requirement that quantum corrections are at most order unity, then places bounds on these power counting parameters in terms of the leading-order interaction coefficients (which are not renormalised on dimensional grounds and so we allow them to take independent values). In particular, when boosts are broken spontaneously the field coupling is fixed in terms of \( f_\pi / \Lambda \) (the ratio of the decay constant to the EFT cutoff), and then radiative stability can place constraints directly on the leading-order coefficients. When applied to the EFT of Inflation, this shows that the leading-order interactions \( S_{\text{LO}}[\delta g^{\mu \nu}] \) in (1.1) can only be accompanied by higher-order corrections of the form (2.24) in a radiatively stable way providing \( M_n^4 \lesssim f_\pi^4 c_s^{1-2n} \) (or \( \alpha_j \lesssim 1/c_s^2 \) and \( \beta_j \lesssim 1/c_s^4 \)). The the EFT cutoff in \( Z^{\mu \nu} \partial_\mu \partial_\nu \) is \( \Lambda^2 = 4\pi c_s^2 f_\pi^2 / g_n \), while the cutoff in time derivatives \( n^{\mu \nu} \partial_\mu \partial_\nu \) is \( \Lambda / g_n \), and the lowest value that \( g_n \) can take is given by whichever \( (M_n^4 / f_\pi^4 c_s^{1-2n})^{1/(2n-2)} \) is largest. In particular, \( g_n \) may not be lower than \( \sqrt{1 - c_s^2} \).

We then focussed on the scattering amplitude mediated by \( S_{\text{LO}} \). By extending the usual

---

**Figure 6.** Left panel shows the unitary values of \( \omega_s \) and \( |p_s| / f_\pi \) for four different values of \((c_s, \alpha_1)\)—namely \((0.16, -250)\) in light blue, \((0.16, 1600)\) in dark blue, \((0.032, -250)\) in light red, \((0.032, 1600)\) in dark red. The hatched region \( c_\pi |p_s| > \omega_s \) cannot be accessed kinematically (no real time-like \( p_1 \) and \( p_2 \) give such a \( p_s \)). The maximum energy scale, \( \omega_{\text{max}} \), is independent of \( \alpha_1 \), while the maximum spatial scale, \( |p|_{\text{max}} \), is independent of \( c_s \). Right panel shows the \((c_s, \alpha_1)\) parameter space, with corresponding values of \((\omega_{\text{max}}, |p|_{\text{max}})\). Grid lines indicate \( \alpha_1 = \pm 250, \pm 1600 \) and \( c_s = 0.032, 0.16 \) (crosses correspond to the points depicted on the left). The gray contours show the Planck 2018 68%, 95% and 99.7% confidence intervals from \( f_{NL}^{\text{equil}} \) and \( f_{NL}^{\text{orth}} \). Finally, the region in which there is at least some \( \omega_s > 6H \) for which \( |p_s| \) can be \( > 2H \) is shown is green. Although when \( \alpha_1 = -1600 \) both \( \omega_{\text{max}} > 6H \) and \( c_\pi |p|_{\text{max}} > 2H \), they cannot both be subhorizon.
partial wave expansion of the $2 \to 2$ scattering amplitude to include preferred-frame effects (a breaking of boosts due to a constant time-like $n^\mu$),

$$\mathcal{A}(\omega_s, \rho_s, \hat{p}_1, \hat{p}_3) = \frac{16\pi^2 \omega_s^2 (1 - \rho_s^2)}{\omega_1 \omega_3} \sum_{\ell_1, \ell_3, m_1, m_3} a_{\ell_1\ell_3, m_1} \delta_{m_1, m_3} Y_{\ell_1}^{m_1*}(\hat{p}_1) Y_{\ell_3}^{m_3}(\hat{p}_3), \quad (4.1)$$

we have shown that unitarity requires $|\det a_{\omega_s}^{\mu m}(\omega_s, \rho_s)| < 1$ for any finite submatrix of $a^{\mu m}$ and for any centre-of-mass velocity, $\rho_s$. Applying these new constraints to the EFT of Inflation, we have shown how the leading order Wilson coefficients $\{c_s, \alpha_1, \beta_1\}$ are related to the strong coupling scales (at which perturbation theory breaks down) for $s$ (the internal interaction energy) and $c_\pi |p_s|$ (the energy in the centre-of-mass motion). Combining Planck observational constraints on the equilateral and orthogonal bispectrum shapes with the prior that inflation is approximately single-field and weakly coupled on subhorizon scales (i.e. the regime (1.7) in which mode functions behave as plane waves) results in an improvement of the 68% confidence interval by a factor of $\approx 3$. For comparison, Simons Observatory (due to begin taking data in early 2020s) has a goal sensitivity [26] which would improve this interval by a factor of $\approx 3.6$, while the later CMB-S4 experiment [27, 28] forecasts an improvement by a factor of $\approx 6.0$.

One of our main observations is that the EFT cutoffs in energy and momentum need not be the same in EFTs with broken boosts. In fact, we have shown that they must necessarily differ in the presence of Lorentz-violating cubic interactions (like $\dot{\pi}^3$). It is worth pointing
out that, if the new physics which UV completes the EFT (1.3) was Lorentz-invariant, one
might expect that the energy and momentum cutoffs should be the same, particularly in the
limit $c_s \to 1$ (since then at the level of particle propagation there is no Lorentz-violation
either from $\pi$ or from the heavy physics). We certainly have not proved this. But it might
be pointing in the same direction as the $c_s = 1$ conjecture made in [24] that the existence of
a Lorentz-invariant UV completion requires $\alpha_1 \to 0$ as $c_s \to 1$.

Looking ahead, there are many other avenues for future work. For instance, here we
have deliberately removed the $\beta_1$ dependence from the spherical-wave unitarity bounds in
order to provide robust constraints on the bispectrum alone, but in future, as bispectrum
measurements improve, there is the possibility of using these unitarity bounds to instead
constrain $\beta_1$. Exploiting this unitarity connection between bispectrum and trispectrum may
allow us to use future measurements of $f_{NL}$ to bootstrap information about higher order
non-Gaussianities.

There are also other systems to which our unitarity bounds may be applied. The most
closely related is dark energy fluctuations in the late Universe, which can be described by
a functionally similar EFT [29]. Further afield are condensed matter systems (with approxi-
mately linear dispersion relations) which share the same symmetry breaking pattern, for
instance those studied in [30, 31].

Furthermore, there is the potential to constrain the EFT of Inflation even further by
extending other amplitude techniques to finite centre-of-mass velocities. One example is the
positivity bounds put forward in [32] and applied to the EFT of Inflation in [24], which we
have also extended to $\rho_s \neq 0$ and will present in a following work [33].

Also, since we have worked entirely within the decoupling limit, our amplitudes are not
sensitive to the propagation of any degrees of freedom at $c_T$ (rather $c_T$ enters only as the
invariant-speed preserved by the non-linear boosts). This means that we do not encounter
any particular effects when $\rho_s$ crosses $c_s$ (i.e. when the $\pi$ centre-of-mass speed exceeds $c_T$),
in particular there is no Cerenkov radiation in the decoupling limit. In fact, the graviton
exchange contribution to $A_{\text{tree}}^{2 \to 2}$ can be estimated by assuming that the propagator receives
$O(H \omega)$ corrections as it is taken on-shell, giving $\sim O(\omega^3/M_P^2 H)$, which is indeed negligible.
It would be interesting to go beyond the decoupling limit to explore the consistency of the
theory at speeds $\rho_s > c_s$ further in future.

And finally, developing a more thorough understanding of how radiative stability and
unitarity manifest directly in the in-in correlators (the natural observables from inflation),
removing the need for our restriction to strictly subhorizon scales, is particularly important.
One possible in-road may be the recent connection uncovered in [34–37] that cosmological
correlators contain the corresponding flat space scattering amplitudes as the residue of the
total momentum pole.

To sum up, we have considered how EFTs with broken boosts can be constrained using
radiative stability and perturbative unitarity and used these to improve constraints on the
primordial bispectrum produced during inflation. In this work we have made a step forward
in importing powerful Lorentz-invariant EFT techniques to cosmology, a programme which
is particularly important in light of next generation experiments.

Acknowledgments

We thank Sadra Jazayeri, Daan Meerburg, Johannes Noller and David Stefanyszyn for useful discussions, and in particular Enrico Pajer for comments on a draft. TG is supported by the Cambridge Trust. SM is supported by an Emmanuel College Research Fellowship. This work is supported in part by STFC under grants ST/L000385, ST/L000636 and ST/P000681/1.

A The Spherical Wave Expansion

A.1 Unitarity Bound on the Minors of $a^m$

Strictly speaking, $(a^m)_{\ell_1 \ell_2}$ is an infinite dimensional matrix, and operations like $\det(\cdot)$ must be handled with care. However, if we focus on any (finite) submatrix, $a^m_{\text{sub}}$, in which both $\ell_i$ indices run over a (finite) set of integers $N$, then we can write the unitarity condition as,

$$\text{Im} (a^m_{\text{sub}})_{\ell_1 \ell_2} = \sum_{\ell \in N} (a^m_{\text{sub}})_{\ell_1 \ell} (a^{m\dagger}_{\text{sub}})_{\ell \ell_2} + \sum_{L \not\in N} a^m_{\ell_1 L} a^{m\dagger}_{L \ell_2} =: A_{\ell_1 \ell_2} + B_{\ell_1 \ell_2} \quad (A.1)$$

where $A$ and $B$ are both positive-definite matrices.

Since $A$ and $B$ are both positive-definite, then for any vector $v$,

$$v^T (A + B) v \geq v^T A v \geq 0 \quad (A.2)$$

This implies $\det(A + B) \geq \det(A)$, since we can use

$$e^{-v^T (A+B) v} \leq e^{-v^T A v} \quad (A.3)$$

to show that,

$$\int dv \ e^{-v^T (A+B) v} \leq \int dv \ e^{-v^T A v} \Rightarrow \frac{1}{\sqrt{\det(A + B)}} \leq \frac{1}{\sqrt{\det(A)}}. \quad (A.4)$$

The unitarity condition can therefore be written as,

$$\det(\text{Im} a^m_{\text{sub}}) \geq \det(a^m_{\text{sub}} a^{m\dagger}_{\text{sub}}) = |\det(a^m_{\text{sub}})|^2. \quad (A.5)$$

Finally, note that if we perform the Cartesian decomposition, $a^m = \text{Re} a^m + i \text{Im} a^m$, and $\text{Im} a^m$ is positive-definite, then $\det(a^m) \geq \det(\text{Im} a^m)$. This allows us to write the following unitarity condition for any finite submatrix $a^m_{\text{sub}}$ of $a^m_{\ell_1 \ell_2}$,

$$\det(a^m_{\text{sub}}) \leq 1. \quad (A.6)$$
as quoted in the main text.

A.2 Evaluating the Spherical Wave Coefficients

In the main text we focused only on $\ell_1 = \ell_3 \leq 2$, but here we provide general formulae which can be used to generate any desired $a_{\ell_1 \ell_3}^{m_1 m_3}$. For convenience, we split the partial wave coefficient into three pieces,

$$a_{\ell_1 \ell_3}^{m m} = [a_{\ell_1 \ell_3}^{m m}]_s + [a_{\ell_1 \ell_3}^{m m}]_t + [a_{\ell_1 \ell_3}^{m m}]_u \tag{A.7}$$

which correspond to the parts of $A$ with no angular poles (i.e. the contact and $s$-channel exchange diagrams, as well as the analytic parts of the $t$- and $u$-channel exchange diagrams), a pole in $t$ (from the $t$-channel exchange diagram) and a pole in $u$ (from the $u$-channel exchange diagram). We will discuss the computation of each piece in turn.

We will use our single remaining rotation (around $p_s$) to fix $\phi_1 = \pi/2$, and focus on the remaining integral over $\phi_3$, $\theta_1$ and $\theta_3$. The overall strategy is to first fix an $m$ and do the $\phi_3$ integral explicitly, and then fix $\ell_1$ and $\ell_3$ and do the $\theta$ integrals. In the limit $\rho_s \to 1$, it is possible to write down simple closed form expressions for $a_{\ell_1 \ell_3}^{m m}$ for any $m$ and $\ell$.

A.2.1 The $s$-channel pole

The $\phi$ integrals

The integral over $\phi$ can be done immediately, since there are no angular poles in $[a_{\ell_1 \ell_3}^{m m}]_s$ by construction. For any positive integer $n$, we have that\(^{30}\),

$$\int_0^{2\pi} d\phi_3 \frac{t^n}{\omega_1^n \omega_3^n} e^{-im\phi_3} \frac{n! i^{n-|m|}}{|m|! (n-|m|)!} \frac{(1-xy)^{n-|m|} (1-x^2)^{|m|/2} (1-y^2)^{|m|/2}}{(1-x^2)(1-y^2)} \times {}_2F_1 \left( \frac{|m|-n}{2}, \frac{1+|m|-n}{2}; 1+|m|; \frac{(1-x^2)(1-y^2)}{(1-xy)^2} \right), \tag{A.8}$$

where we have used that $t = 2\omega_1 \omega_3 (1 - \cos \theta_1 \cos \theta_3 - \sin \theta_1 \sin \theta_3 \sin \phi_3)$ when $\phi_1 = \pi/2$ is fixed, and introduced $x = \cos \theta_1, y = \cos \theta_3$. Note that the integral vanishes unless $n \geq |m|$, and so when combined with the $(1-x^2)^{|m|/2}$ from the spherical harmonic this integral always produces an analytic polynomial in $x^p y^q$. Using $u = 4m^2 - s - t$, all of the integrals appearing in $[a_{\ell_1 \ell_3}^{m m}]_s$ can be written in this form (multiplied by $\phi$-independent factors).

\(^{30}\)One simple way to derive (A.8) is to expand $t^n$ as a power series in $\sin \phi_3$, do the integration term by term, and then resum into a standard hypergeometric form.
For the EFT amplitude (3.52) and up to $\ell = 2$, we require only the first few,

$$\int_0^{2\pi} \frac{d\phi_3}{2\pi} t = 2\omega_1\omega_3(1-xy)$$

$$\int_0^{2\pi} \frac{d\phi_3}{2\pi} t^2 = 2\omega_1^2\omega_3^2 (2(1-xy)^2 + (1-x^2)(1-y^2))$$

$$\int_0^{2\pi} \frac{d\phi_3}{2\pi} e^{\pm i\phi_3} t = \pm i\omega_1\omega_3\sqrt{1-x^2}\sqrt{1-y^2}$$

$$\int_0^{2\pi} \frac{d\phi_3}{2\pi} e^{\pm i\phi_3} t^2 = \pm 4i\omega_1^2\omega_3^2 (1-xy)\sqrt{1-x^2}\sqrt{1-y^2}$$

$$\int_0^{2\pi} \frac{d\phi_3}{2\pi} e^{\pm i\phi_3} t^2 = -\omega_1^2\omega_3^2 (1-x^2)(1-y^2).$$

(A.9)

The $\theta$ integrals

After carrying out the $\phi$ integral, this leaves a sum over terms in $\omega_1^a\omega_3^b x^p y^q$ with integer powers. We can then replace every $x$ and $y$ with inverse powers of $\omega_i$,

$$\rho_s x = 1 - \frac{1 - \rho_s^2}{2} \omega_s \omega_1, \quad \rho_s y = 1 + \frac{1 - \rho_s^2}{2} \omega_s \omega_3$$

(A.10)

and express $[a_{\ell_1,\ell_3}^{mm}]_s$ as a sum over integrals of $\omega_1^a\omega_3^b$. We then write this as a sum over $I^a I^b$, where $I^a$ is a set of elementary master integrals,

$$I^a(\rho_s) = \frac{1}{2} \int_{-1}^{1} dx \ (1 - \rho_s x)^a,$$

(A.11)

and the index $a$ runs from $-\max_1$ to $\ell_1 - \min_1$, where $\min_1$ ($\max_1$) is related to the minimum (maximum) power of $\omega_1$ appearing in $[A]_s$ (and similarly $b$ runs from $-\max_3$ to $\ell_3 - \min_3$).

For the EFT amplitude (3.52), the ranges are $-3 \leq a \leq \ell_1 - 1$ and $-3 \leq b \leq \ell_3 - 1$ for the sum over $I^a I^b$. These integrals can be evaluated straightforwardly (e.g. by changing the integration variable to $u = 1 - \rho_s x$),

$$I^a(\rho_s) = \begin{cases} 
\frac{1}{2\rho_s} \log (\rho_+/\rho_-) & \text{if } a = -1, \\
\frac{\rho_{s+a+1} - \rho_{s+a+1}}{2\rho_s (a+1)} & \text{otherwise},
\end{cases}$$

(A.12)

where $\rho_- = 1 - \rho_s$ and $\rho_+ = 1 + \rho_s$.

We also note that Rodriguez’s formula for the Legendre polynomial can be used to provide a closed form expression directly for,

$$I^a_\ell(\rho_s) := \frac{1}{2} \int_{-1}^{1} dx \ \left( \frac{\omega_1}{\omega_s} \right)^a P_\ell(x)$$

(A.13)
in terms of Gauss hypergeometric functions,
\[
I^a_\ell(\rho_s) = \frac{\sqrt{\pi} \Gamma(\ell + 1) \Gamma(a + \ell)}{\ell! 2^{a+\ell+1} \Gamma(a) \Gamma(\ell + \frac{3}{2})} (1 - \rho_s^2)^a \rho_s^\ell \ {}_2F_1 \left( \frac{a + \ell}{2}, \frac{a + \ell + 1}{2}; \ell + \frac{3}{2}; \rho_s^2 \right),
\]  
(A.14)
which can speed up numerical implementation.

**Small** $1 - \rho_s$

When $\rho_s \to 1$, $\omega_1$ (and $\omega_3$) has support only at $x = 1$ (and $y = 1$), and can be expressed in terms of Dirac $\delta$ functions and their derivatives. For instance,
\[
\frac{1 - \rho_s}{1 - \rho_s x} \approx \rho_- \log \left( \frac{\rho_-}{2} \right) \delta(x - 1) + \rho_- \sum_{k=1}^{\infty} \frac{\rho_-^k}{k!} \delta^{(k)}(x - 1)
\]  
(A.15)
where $\rho_- = 1 - \rho_s$, and we have used $\approx$ to indicate that this is how the right-hand side behaves when integrated against a regular function of $x$. For finite $\rho_-$, the infinite sum indicates finite support in $x$, but when truncated to leading order $\omega_1$ behaves as a single insertion of $\delta(x - 1)$. For arbitrary powers,
\[
\left( \frac{1 - \rho_s}{1 - \rho_s x} \right)^n \approx \sum_{j=1}^{n-1} \rho_-^j \frac{\delta^{(j-1)}(x - 1)}{(n - 1)_j} - \rho_-^n \log \left( \frac{\rho_-}{2} \right) \delta^{(n-1)}(x - 1) + \mathcal{O}(\rho_-^n)
\]  
(A.16)
where $(n - 1)_j = (n - 1)(n - 2) ...(n - j + 1)$ is the falling Pochhammer symbol, and the $\mathcal{O}(\rho_-^n)$ correction is generally an infinite sum over $\delta^{(k)}(x - 1)$ (and so has finite support in $x$).

Since derivatives of the Legendre polynomials obey $P^{(n)}_\ell(1) = (\ell + n)2n/2^n/n!$, where $a_n = a(a - 1) ...(a - n + 1)$ is the falling Pochhammer symbol, we have that $\delta^{(k)}(x - 1)P_\ell(x) \approx \ell^{2k}$, and therefore the series expansion (A.16) of $\omega_1$ can be viewed as an expansion in powers of $(1 - \rho_s)\ell^2$.

So when $1 - \rho_s \ll \ell^{-2}$, we can replace the $\theta$ integrals over $P_\ell$ with simply\footnote{Note that the whole series (A.16) can be resummed, yielding a hypergeometric expression which agrees with (3.41).},
\[
I^a_\ell(\rho_s) = \rho_- \frac{P_1(1)}{a - 1} + \mathcal{O}(\rho_-^2 \ell^2).
\]  
(A.17)
Note that when $m \neq 0$, the $\phi$ integral vanishes when $x \to 1$ and $y \to 1$, and therefore in the $\rho_s \to 1$ limit the dominant partial waves are the $a_{\ell_1 \ell_3}^{b_0}$ modes.
A.2.2 The $t$-channel pole

The $\phi$ integral

First, notice that for a scalar field the residue of the \( t \)-channel pole is always independent of \( s \) and \( u \). The $\phi$ integral over the $t$-channel pole therefore always takes the form,

\[
I_m(c) := \int_0^{2\pi} \frac{d\phi}{2\pi} \frac{e^{-im\phi}}{1 - c \sin \phi} = \int_{-\infty}^{\infty} \frac{dz}{\pi} \frac{1}{1 + z^2 - 2cz} \left( \frac{i + z}{i - z} \right)^m
\]  

(A.18)

where we have used the substitution \( z = \tan(\phi/2) \). For \( m > 0 \), this has two poles in the upper half-plane (at \( z = i \) and \( z = c + i\sqrt{1 - c^2} \)) and one pole in the lower half-plane (at \( z = c - i\sqrt{1 - c^2} \)). For \( m < 0 \), there are two poles in the lower half-plane (now at \( z = -i \)) and one in the upper half-plane. Closing the contour in either direction gives the same answer,

\[
I_m(c) = \frac{i^m}{\sqrt{1 - c^2}} \left( -1 + \sqrt{1 - c^2} \right)^{|m|}
\]  

(A.19)

Consequently,

\[
\int_0^{2\pi} \frac{d\phi_3}{2\pi} e^{-im\phi_3} \frac{\omega_1\omega_s}{s_{13}} = \frac{\rho_s}{1 - \rho_s^2} \frac{\cos \theta_1 - \cos \theta_3}{1 - \cos \theta_1 \cos \theta_3} \frac{I_m}{1 - \cos \theta_1 \cos \theta_3} \left( \frac{\sin \theta_1 \sin \theta_3}{1 - \cos \theta_1 \cos \theta_3} \right)
\]

which leaves the integrand invariant. The first three such integrals can be written in terms of \( x = \cos \theta_1 \) and \( y = \cos \theta_3 \) as,

\[
\int_0^{2\pi} \frac{d\phi_3}{2\pi} \frac{\omega_1\omega_s}{s_{13}} = \frac{\rho_s}{1 - \rho_s^2} \text{sgn}(x - y)
\]  

(A.21)

\[
\int_0^{2\pi} \frac{d\phi_3}{2\pi} e^{i\phi} \frac{\omega_1\omega_s}{s_{13}} = \pm \frac{i\rho_s}{1 - \rho_s^2} \frac{(x - y) + \text{sgn}(x - y)(-1 + xy)}{\sqrt{1 - x^2}\sqrt{1 - y^2}}
\]

\[
\int_0^{2\pi} \frac{d\phi_3}{2\pi} e^{2i\phi} \frac{\omega_1\omega_s}{s_{13}} = \frac{-\rho_s}{1 - \rho_s^2} \frac{2(-1 + xy)(x - y) + \text{sgn}(x - y)\left(1 + x^2 - 4xy + y^2 + x^2y^2\right)}{(1 - x^2)(1 - y^2)}
\]

Note that the denominators always cancel with the factor of \((1 - x^2)^{|m|/2}\) appearing from the \( P_\ell^m(x) \) in the spherical harmonics, leaving an analytic polynomial in \( xy^q \).

The $\theta$ integrals

After carrying out the $\phi$ integral, this leaves a sum over terms in \( \omega_1^p \omega_3^q x^p y^q \) with integer powers. We can then replace every \( x \) and \( y \) with inverse powers of \( \omega_i \) using (A.10) and
express \( \alpha_{\ell_1 \ell_2}^{m} \) as a sum over the elementary master integrals,

\[
J_{ab}(\rho_s) = \frac{1}{4} \int_{-1}^{1} dx \int_{-1}^{1} dy \, (1 - \rho_s x)^a (1 - \rho_s y)^b \, \text{sgn}(x - y)
\]

(A.22)

where \( a \) runs from \(-\max_1 \) to \( \ell_1 - \min_1 \), where \( \min_1 \) (\( \max_1 \)) is related to the minimum (maximum) power of \( \omega_1 \) appearing in \([A]_t\)—for the EFT amplitude (3.52), the ranges are \(-4 \leq a \leq \ell_1 - 2\) and \(-4 \leq b \leq \ell_3 - 2\).

These integrals may be straightforwardly evaluated (for instance by changing integration variables, \( u = 1 - \rho_s x \) and \( v = 1 - \rho_s y \)),

\[
J_{ab}(\rho_s) = \begin{cases} 
\frac{\rho_+^{b+1} - \rho_-^{b+1}}{2\rho_+^2(b+1)} & \text{if } a = -1 \\
\frac{1}{4\rho_+^2(a+1)} \left[ \frac{\rho_+^{1+a}}{\rho_-^{1+a}} \left( 1 - \left( \frac{\rho_-}{\rho_+} \right)^{2a+2} + 2(a+1) \left( \frac{\rho_-}{\rho_+} \right)^{a+1} \log \left( \frac{\rho_-}{\rho_+} \right) \right) \right] & \text{if } a + b = -2 \\
\frac{(b-a)(\rho_+^{a+b+2} - \rho_-^{a+b+2}) + (a+b+2)(\rho_+^{a+1} - \rho_-^{a+1})}{4\rho_+^2(a+1)(b+1)(a+b+2)} & \text{otherwise}
\end{cases}
\]

(A.23)

where \( J_{ab} = -J_{ba} \) is antisymmetric.

**Small \( 1 - \rho_s \)**

Since the \( t \)-channel pole contains a \( 1/|x - y| \) collinear divergence, the expression (A.16) is no longer valid. Nonetheless, the rationale that as \( \rho_s \to 1 \) quantities like \( \omega_1^a \omega_3^b \) become infinitesimally supported is still useful, and indeed,

\[
J_{0 \ell_1 \ell_3}^{ab}(\rho_s) := \frac{1}{4} \int_{-1}^{1} dx \int_{-1}^{1} dy \, \left( \frac{\omega_1}{\omega_s} \right)^a \left( \frac{\omega_3}{\omega_s} \right)^b \, P_{\ell_1}(x) P_{\ell_3}(y) \, \text{sgn}(x - y)
\]

\[= \rho_s^2 \frac{(a-b)P_{\ell_1}(1)P_{\ell_3}(1)}{(a-1)(b-1)(a+b-2)} + \mathcal{O}(\rho_s^3 \ell^2)^2 .
\]

(A.24)

Note that when \( m \neq 0 \), the \( \phi \) integral vanishes when \( x \to 1 \) and \( y \to 1 \), and therefore in the \( \rho_s \to 1 \) limit the dominant partial waves are the \( a_{\ell_1 \ell_3}^{00} \) modes.

**A.2.3 The \( u \)-channel pole**

**The \( \phi \) integral**

Since \( \phi_4 = \pi + \phi_3 \), we can also use the result (A.19) to write,

\[
\int_0^{2\pi} \frac{d\phi_3}{2\pi} e^{-i m \phi_3} \frac{\omega_1 \omega_3}{s_{14}} = \frac{(-i)^m \rho_s}{1 - \rho_s^2} \text{sgn} \left( \cos \theta_1 - \cos \theta_4 \right) \frac{\left( -1 + \cos \theta_1 \cos \theta_4 + |\cos \theta_1 - \cos \theta_4| \right)^{|m|}}{(\sin \theta_1 \sin \theta_4)^{|m|}}
\]

(A.25)
and then use,
\[ \omega_3 \sin \theta_3 = \omega_4 \sin \theta_4 \quad \text{and} \quad \omega_3 \cos \theta_3 + \omega_4 \cos \theta_4 = -\rho_s \omega_s \]  
(A.26)
to write,
\[ \int_0^{2\pi} d\phi_3 \frac{e^{-im\phi_3 \omega_4 \omega_s}}{s_{14}} = \frac{(-i)^m \rho_s}{1 - \rho_s^2} \text{sgn}(\omega_{14}) \left( \frac{-\omega_4 - x(\rho_s \omega_s + \omega_3 y) + \text{sgn}(\omega_{14})(\omega_4 x + \omega_3 y + \rho_s \omega_s))}{\omega_3 \sqrt{1 - x^2} \sqrt{1 - y^2}} \right)^{|m|}. \]
(A.27)

Finally, use \( \omega_4 = -\omega_s - \omega_3 \) to express this integral in terms of \( \omega_1, \omega_3, x \) and \( y \). Note that \( \text{sgn}(\omega_{14}) = \text{sgn}(\frac{x+y}{1+xy} - \frac{2\rho_s}{1+\rho_s^2}) \).

The \( \theta \) integral

This leaves a sum over terms in \( \omega_1^a \omega_3^b x^p y^q \) with integer powers. We can then replace every \( x \) and \( y \) with inverse powers of \( \omega_i \), and express \([r_{\ell_1 \ell_3}^{mm}]_u \) as a sum over the elementary integrals,
\[ K^{ab}(\rho_s) := \frac{1}{4} \int_{-1}^{1} dx \int_{-1}^{1} dy (1 - \rho_s x)^a (1 - \rho_s y)^b \text{sgn}(\frac{x+y}{1+xy} - \frac{2\rho_s}{1+\rho_s^2}) \]
(A.28)

where \( a \) runs from \(-\max_1 \) to \( \ell_1 - \min_1 \), where \( \min_1 \) is related to the minimum power of \( \omega_1 \) appearing in \([A]_u \). For the EFT of Inflation amplitude, the \( K^{ab} \) integrals range from \(-4 \leq a \leq \ell_1 - 2 \) and \(-4 \leq b \leq \ell_3 - 2 \).

The \( K^{ab} \) integrals can also be performed explicitly, for instance by substituting,
\[ 1 - \rho_s x = \frac{1 - \rho_s^2}{1 - \rho_s x'}, \quad 1 - \rho_s y = \frac{1 - \rho_s^2}{1 - \rho_s y'} \]
(A.29)

which sends \( \text{sgn}(\omega_{14}) \) to \(-\text{sgn}(x' + y') \). This gives,
\[ K^{ab}(\rho_s) = \begin{cases} \frac{1}{\rho_s^2} \left[ \frac{\pi^2}{12} - \frac{1}{2} \log \left( \frac{\rho_s}{2} \right)^2 + \frac{1}{4} \log \left( \frac{\rho_s}{\rho_\ast} \right)^2 - \text{Li}_2 \left( \frac{\rho_s}{2} \right) \right] & \text{if } a = b = -1, \text{ else:} \\ -\frac{(1-\rho_s^2)^{a+b+2}}{4\rho_s(b+1)} \left[ (\rho_\ast^{-b-1} + \rho_\ast^{-b-1}) I^{-a-2}(\rho_s) - \frac{2}{\rho_s} \sum_{n} \left( \frac{a-b+1}{2n} \right) B_{\rho_s^2} (n + \frac{1}{2}, -b - 1) \right] \end{cases} \]
(A.30)

where \( \text{Li}_2(z) \) is the dilogarithm and \( B_z(a, b) \) is the incomplete Beta function. Note that since \( K^{ab} = K^{ba} \) is symmetric, one can always choose \( b \neq -1 \) unless \( a = b = -1 \). For small values of \( a \) and \( b \) (i.e. small values of \( \ell_1 \) and \( \ell_3 \)), (A.30) produces relatively simple polynomials in \( \rho_s \).
Small $1 - \rho_s$

Since the $u$-channel pole also contains a collinear divergence, the expression (A.16) is no longer valid. Nonetheless, the rationale that as $\rho_s \to 1$ quantities like $\omega_1^a \omega_3^b$ become infinitesimally supported is still useful, and indeed,

$$K^{ab}_{\ell_1 \ell_3}(\rho_s) := \frac{1}{4} \int_{-1}^{1} dx \int_{-1}^{1} dy \left( \frac{\omega_1}{\omega_s} \right)^a \left( \frac{\omega_3}{\omega_s} \right)^b P_{\ell_1}(x) P_{\ell_3}(y) \text{sgn} \left( \frac{x + y}{1 + xy} - \frac{2\rho_s}{1 + \rho_s^2} \right)$$

$$= \rho_s^2 \left[ \frac{1}{(a - 1)(b - 1)} - \frac{2(a - 2)! (b - 2)!}{(a + b - 2)!} \right] P_{\ell_1}(1) P_{\ell_3}(1) + O(\rho_s^3 \ell^2). \quad (A.31)$$

Note that when $m \neq 0$, the $\phi$ integral vanishes when $x \to 1$ and $y \to 1$, and therefore in the $\rho_s \to 1$ limit the dominant partial waves are the $a_0^{00} \ell_{\ell_3}$ modes.

A.3 Some Explicit Examples

$\hat{\pi}^4$ Interaction

For example, the simple $\hat{\pi}^4$ amplitude (3.43) can be written in terms of (A.14),

$$\hat{I}_\ell(\rho_s) = \sqrt{\frac{4! (2\ell + 1)}{4\pi(1 - \rho_s^2)}} \left( I^2_\ell(\rho_s) - I^3_\ell(\rho_s) \right). \quad (A.32)$$

In particular, the first subleading term in the small $1 - \rho_s$ expansion of (3.41) is,

$$\hat{I}_\ell(\rho_s) = \sqrt{\frac{3(2\ell + 1)}{16\pi}} \sqrt{1 - \rho_s} \left[ 1 - \frac{\rho_s}{4} (1 + 2\ell(5 + \ell) - 4\ell(1 + \ell)(H_{\ell/2} + H_{(\ell+1)/2} + \log(2 - 2\rho_s))) + O((1 - \rho_s)^2) \right] \quad (A.33)$$

where $H_n$ is the $n^{th}$ harmonic number, which grows logarithmically at large $n$. When $\ell \gg 1$, it is this logarithmic growth which dominates, and leads to the simple expression (3.44) which we used to estimate at which $\ell$ the unitarity sum should be truncated.

All leading order cubic/quartic interactions

Although (1.3) contains five Wilson coefficients, the first spherical-wave coefficient depends on only four independent combinations of them, which we can denote,

$$C_0 = -27\alpha_1^2 - 72\alpha_1\alpha_2 - 48\alpha_2^2 + 18\beta_1 + 24\beta_2 + 40\beta_3,$$  \quad (A.34)

$$C_1 = \frac{3}{2} \left( -27\alpha_2^2 + 108\alpha_1\alpha_2 + 112\alpha_2^2 - 27\beta_1 - 56\beta_2 - 140\beta_3 \right), \quad (A.35)$$

$$C_2 = \frac{2\alpha_2^2 - \beta_2 - 4\beta_3}{300}, \quad C_3 = \frac{-2\alpha_2^2 + \beta_2 - 4\beta_3}{300} \quad (A.36)$$
I express the Legendre-weighted integrals \( \omega^a_{\ell m} \) and \( \omega^a_{\ell m} \) zero otherwise. The resulting posterior is used to generate contours within which 68\%, 95\% mean and variance) by a prior distribution which is uniform for the allowed spectrum shapes.

This concludes this section on evaluating the partial wave coefficients.

\[ a^0_{00} = \frac{1}{192\pi} [C_0 f_0(\rho_s) + C_1 f_1(\rho_s) + C_2 f_2(\rho_s) + C_3 f_3(\rho_s)] \]  

where the functions \( f_j(\rho_s) \) have the convenient property that \( f_j(\rho_s) = \rho_s^{2j} + \ldots \) at small \( \rho_s \), and are given explicitly by,

\[
\begin{align*}
  f_0 &= 1 , \\
  f_1 &= \rho_s^2 , \\
  f_2 &= \rho_s^4 (3F(1 - \rho_s^2)^2 + 5(1 + \rho_s^2))(3F(1 - \rho_s^2)^2 (-3 + 7\rho_s^2) + 5(-91 + 4\rho_s^2 + 7\rho_s^4)) , \\
  f_3 &= \rho_s^4 (-1 + \rho_s^2)(30F(-1 + \rho_s^2)(9 + 20\rho_s^2 + 7\rho_s^4) + 25(43 + 34\rho_s^2 + 7\rho_s^4) + 9F^2(3 - 2\rho_s^4 - 8\rho_s^6 + 7\rho_s^8))
\end{align*}
\]

where \( F = -(5(2\rho_s(3 + \rho_s^2) + 3 \log \left(\frac{1 - \rho_s}{1 + \rho_s}\right) / 6\rho_s^2 \) has the limit \( F(0) = 1 \).

The small \( \rho_s \) expansion of the \( \ell = 1 \) and \( \ell = 2 \) diagonal entries of \( a^m \) are,

\[
\begin{align*}
  a^0_{11} &= -\frac{7\rho_s^2}{150\pi} \beta_3 + \frac{\rho_s^2}{192\pi} [ -(3\alpha_1 + 4\alpha_2)^2 + 6\beta_1 + 8\beta_2 + 24\beta_3] + \ldots \\
  a^\pm_{11} &= \frac{\rho_s^2}{150\pi} \beta_3 + \ldots \\
  a^0_{22} &= \beta_3 + \frac{\rho_s^2}{120\pi} (135\alpha_1^2 + 112\alpha_2^2 - 56\beta_2 - 452\beta_3) + \ldots \\
  a^\pm_{22} &= \frac{\beta_3}{120\pi} - \frac{\rho_s^2}{3360\pi} (45\alpha_1^2 + 92\beta_3) + \ldots \\
\end{align*}
\]

which indeed satisfy (3.36) and reduce to the \( a_\ell \) of [24] when \( \rho_s \to 0 \).

This concludes this section on evaluating the partial wave coefficients \( a^m_{\ell m} \). We have successfully broken a general partial wave up into a sum over simple master integrals, \( I^a I^b J^{ab} \) and \( K^{ab} \), and given explicit closed expressions for them. Furthermore, when \( \rho_s \to 1 \), the \( \omega_1 \) and \( \omega_3 \) act as \( \delta(x - 1) \) and \( \delta(y - 1) \) and their derivatives, allowing us to write particularly compact expressions for the Legendre-weighted integrals \( I^a_{\ell_1} I^b_{\ell_3}, J^{ab}_{\ell_1} \ell_3 \) and \( K^{ab}_{\ell_1} \ell_3 \) which make up \( a^0_{\ell_1} \ell_3 \).

## B Unitarity Priors

In this short appendix we describe how our unitarity bounds were used to produce Figure 7. In simple terms, we first translate our allowed region of \{\( c_s, \alpha_1 \)\} into an allowed region for bispectrum shapes \( \{f_{NL}^{\text{equil}}, f_{NL}^{\text{orth}} \} \), and then multiply a simple Gaussian likelihood (using Planck’s mean and variance) by a prior distribution which is uniform for the allowed \( \{f_{NL}^{\text{equil}}, f_{NL}^{\text{orth}} \}\) (and zero otherwise). The resulting posterior is used to generate contours within which 68\%, 95\% and 99.7\% of the total probability lie (preferentially filling the most likely bins first).
The observed limits on the bispectrum reported by the Planck Collaboration [25] on equilateral and orthogonal shapes are:

\[ f_{NL}^{\text{eqil}} = -26 \pm 47, \quad f_{NL}^{\text{orth}} = -38 \pm 24, \]  

(B.1)

from \( T + E \) correlations once lensing is subtracted (uncertainties correspond to 68% confidence). From the EFT of Inflation’s scalar sector (1.3), the expected non-Gaussianity with model parameters \( c_s \) and \( \alpha_1 \) are (using the Planck 2018 Fisher matrix),

\[ f_{NL}^{\text{eqil}} = -0.156\alpha_1 - 0.353 \frac{1 - c_s^2}{c_s^2}, \]  

(B.2)

\[ f_{NL}^{\text{orth}} = 0.0334\alpha_1 - 0.00081 \frac{1 - c_s^2}{c_s^2} + 0.0334 \left( 1 - c_s^2 \right)^2. \]  

(B.3)

We will assume that the uncertainties are Gaussian, such that the likelihood of measuring \((f_{NL}^{\text{eqil}}, f_{NL}^{\text{orth}})\) when the true values are \((-26, -38)\) is simply, \( L = e^{-\chi^2/2/(2\pi \det C)} \), where,

\[ \chi^2(f_{NL}^{\text{eqil}}, f_{NL}^{\text{orth}}) = \left( f_{NL}^{\text{eqil}} + 26, f_{NL}^{\text{orth}} + 38 \right) C^{-1} \left( f_{NL}^{\text{eqil}} + 26, f_{NL}^{\text{orth}} + 38 \right). \]  

(B.4)

coincides with the \( \chi^2 \) used in the analysis of [25], with a covariance matrix given by \( C \approx \text{diag}(47^2, 24^2) \). Contours of \( \chi^2(f_{NL}^{\text{eqil}}, f_{NL}^{\text{orth}}) \leq 2.28, 5.99, \) and 11.62 correspond to an enclosed probability of 68%, 95% and 99.7% respectively, and are plotted in Figure 7 as “Planck 2018”, giving constraints on the parameters \((c_s, \alpha_1)\) from Planck’s observations alone.

The advantage of assuming a simple likelihood, \( L \), is that we can use Bayesian inference to incorporate our unitarity priors: once we identify the “theoretically allowed” region (in which, say, \( \Lambda \geq 10H \)), we use a uniform prior \( P(f_{NL}^{\text{eqil}}, f_{NL}^{\text{orth}}) \) which is constant in the allowed region and zero elsewhere. The analogous 68%, 95% and 99.7% contours of the posterior \( \propto L \times P(f_{NL}^{\text{eqil}}, f_{NL}^{\text{orth}}) \) are then plotted in Figure 7. Demanding that perturbative unitarity is not violated in the subhorizon regime (1.7) improves the volume of the 68% confidence interval in \((f_{NL}^{\text{eqil}}, f_{NL}^{\text{orth}})\) by a factor of \( \approx 3 \).

For comparison, Simons Observatory (due to begin taking data in early 2020s) has a goal sensitivity of \( \sigma(f_{NL}^{\text{eqil}}) = \pm 24 \) and \( \sigma(f_{NL}^{\text{orth}}) = \pm 13 \) (compared with Planck’s \( \pm 47 \) and \( \pm 24 \)) [26]. This corresponds to improving the volume of the 68% confidence interval by a factor of \( \approx 3.6 \), versus Planck 2018 data alone. Similarly, the later CMB-S4 experiment aims to reduce the uncertainty to \( \sigma(f_{NL}^{\text{eqil}}) = \pm 21 \) and \( \sigma(f_{NL}^{\text{orth}}) = \pm 9 \) [27, 28], reducing Planck’s 68% confidence interval by a factor of \( \approx 6.0 \). The provides context for the improvement (by up to a factor of \( \approx 14 \)) that we have achieved here using the Planck 2018 data supplemented with theoretical priors. Going forwards, the theoretical priors identified here can also be combined with future data sets to further improve our understanding of primordial non-Gaussianity.
References

[1] G. F. Chew, *S-matrix theory of strong interactions*. Benjamin, New York, 1961.

[2] R. J. Eden, P. V. Landshoff, D. I. Olive, and J. C. Polkinghorne, *S-matrix theory of strong interactions*. Cambridge University Press, 1966.

[3] B. W. Lee, C. Quigg and H. Thacker, *Weak Interactions at Very High-Energies: The Role of the Higgs Boson Mass*, *Phys. Rev. D* **16** (1977) 1519.

[4] C. Cheung, P. Creminelli, A. L. Fitzpatrick, J. Kaplan and L. Senatore, *The Effective Field Theory of Inflation*, *JHEP* **03** (2008) 014, [0709.0293].

[5] X. Chen, M.-x. Huang, S. Kachru and G. Shiu, *Observational signatures and non-Gaussianities of general single field inflation*, *JCAP* **01** (2007) 002, [hep-th/0605045].

[6] C. Cheung, A. L. Fitzpatrick, J. Kaplan and L. Senatore, *On the consistency relation of the 3-point function in single field inflation*, *JCAP* **0802** (2008) 021, [0709.0295].

[7] X. Chen, M.-x. Huang and G. Shiu, *The Inflationary Trispectrum for Models with Large Non-Gaussianities*, *Phys. Rev. D* **74** (2006) 121301, [hep-th/0610235].

[8] PLANCK collaboration, N. Aghanim et al., *Planck 2018 results. VI. Cosmological parameters*, 1807.06209.

[9] N. Birrell and P. Davies, *Quantum Fields in Curved Space*. Cambridge Monographs on Mathematical Physics. Cambridge Univ. Press, Cambridge, UK, 2, 1984, 10.1017/CBO9780511622632.

[10] H. Georgi, *Vector Realization of Chiral Symmetry*, *Nucl. Phys. B331* (1990) 311–330.

[11] A. V. Manohar, *Introduction to Effective Field Theories*, in *Les Houches summer school: EFT in Particle Physics and Cosmology*, 4, 2018. 1804.05863.

[12] G. F. Giudice, C. Grojean, A. Pomarol and R. Rattazzi, *The Strongly-Interacting Light Higgs*, *JHEP* **06** (2007) 045, [hep-ph/0703164].

[13] A. Manohar and H. Georgi, *Chiral Quarks and the Nonrelativistic Quark Model*, *Nucl. Phys. B234* (1984) 189–212.

[14] A. G. Cohen, D. B. Kaplan and A. E. Nelson, *Counting 4 pis in strongly coupled supersymmetry*, *Phys. Lett. B412* (1997) 301–308, [hep-ph/9706275].

[15] E. E. Jenkins, A. V. Manohar and M. Trott, *Naive Dimensional Analysis Counting of Gauge Theory Amplitudes and Anomalous Dimensions*, *Phys. Lett. B726* (2013) 697–702, [1309.0819].

[16] G. Buchalla, O. Cat and C. Krause, *On the Power Counting inEffective Field Theories*, *Phys. Lett. B731* (2014) 80–86, [1312.5624].

[17] S. Weinberg, *Phenomenological Lagrangians*, *Physica A96* (1979) 327–340.

[18] C. P. Burgess, *Introduction to Effective Field Theory*, *Ann. Rev. Nucl. Part. Sci.* **57** (2007) 329–362, [hep-th/0701053].

[19] G. Goon, K. Hinterbichler, A. Joyce and M. Trodden, *Aspects of Galileon Non-Renormalization*, *JHEP* **11** (2016) 100, [1606.02295].
[20] E. Pajer, G. L. Pimentel and J. V. S. Van Wijck, *The Conformal Limit of Inflation in the Era of CMB Polarimetry*, *JCAP* 06 (2017) 009, [1609.06993].

[21] P. Adshead, C. P. Burgess, R. Holman and S. Shandera, *Power-counting during single-field slow-roll inflation*, *JCAP* 1802 (2018) 016, [1708.07443].

[22] D. Baumann, D. Green and R. A. Porto, *B-modes and the Nature of Inflation*, *JCAP* 1501 (2015) 016, [1407.2621].

[23] J. D. Richman, *An Experimenter’s Guide to the Helicity Formalism*, .

[24] D. Baumann, D. Green, H. Lee and R. A. Porto, *Signs of Analyticity in Single-Field Inflation*, *Phys. Rev.* D93 (2016) 023523, [1502.07304].

[25] PLANck collaboration, Y. Akrami et al., *Planck 2018 results. IX. Constraints on primordial non-Gaussianity*, 1905.05697.

[26] SIMons OBServatory collaboration, P. Ade et al., *The Simons Observatory: Science goals and forecasts*, *JCAP* 1902 (2019) 056, [1808.07445].

[27] CMB-S4 collaboration, K. N. Abazajian et al., *CMB-S4 Science Book, First Edition*, 1610.02743.

[28] K. Abazajian et al., *CMB-S4 Science Case, Reference Design, and Project Plan*, 1907.04473.

[29] G. Gubitosi, F. Piazza and F. Vernizzi, *The Effective Field Theory of Dark Energy*, *JCAP* 02 (2013) 032, [1210.03845].

[30] A. Nicolis, R. Penco, F. Piazza and R. Rattazzi, *Zoology of condensed matter: Framids, ordinary stuff, extra-ordinary stuff*, *JHEP* 06 (2015) 155, [1501.03845].

[31] T. Grall, S. Jazayeri and D. Stefanyszyn, *Symmetries and Amplitudes for a Single Phonon*. In preparation.

[32] A. Adams, N. Arkani-Hamed, S. Dubovsky, A. Nicolis and R. Rattazzi, *Causality, analyticity and an IR obstruction to UV completion*, *JHEP* 10 (2006) 014, [hep-th/0602178].

[33] T. Grall and S. Melville, *Positivity Bounds with Spontaneously Broken Boosts*. To appear.

[34] J. M. Maldacena and G. L. Pimentel, *On graviton non-Gaussianities during inflation*, *JHEP* 09 (2011) 045, [1104.2846].

[35] S. Raju, *New Recursion Relations and a Flat Space Limit for AdS/CFT Correlators*, *Phys. Rev.* D85 (2012) 126009, [1201.6449].

[36] N. Arkani-Hamed and J. Maldacena, *Cosmological Collider Physics*, 1503.08043.

[37] N. Arkani-Hamed, D. Baumann, H. Lee and G. L. Pimentel, *The Cosmological Bootstrap: Inflationary Correlators from Symmetries and Singularities*, 1811.00024.