Causality in Classical Field Theory

Randy S

Abstract In special relativity, the **causality** principle says that the speed at which information propagates from one place to another cannot exceed a finite maximum speed, usually called the speed of light. In classical field theory, this means that if two solutions of the field equations have initial data that differ only within a bounded region of space, then any differences between the two solutions should remain contained within a bounded region that does not grow any faster than a finite maximum speed. This article reviews proofs of causality in simple cases, namely free scalar fields and the free electromagnetic field, and also describes some intuition suggesting that it still holds when the equations of motion are nonlinear.

Contents

© 2018-2024 Randy S
For noncommercial use only

1 Causality: definitions

The principle of causality says that information cannot propagate faster than a finite maximum speed.[1] Article 48968 introduced the principle of causality as a constraint on an object's worldline: the worldline representing an object's journey through spacetime cannot be spacelike anywhere.[2] In this article, that version of the principle of causality will be called **worldline causality**.

For fields, the principle of causality is expressed differently, because the dynamics of a field is not expressed in terms of worldlines. The equations of motion for a classical field are partial differential equations (PDEs). For classical fields, the principle of causality is a condition on those PDEs. Given a system of PDEs and a hypersurface[3] H on which initial data[4] is specified, the **domain of influence**[5] of a subset $h \subset H$ is the region of dimensional space(time) in which the solution of the PDE can be affected by a change of its initial data in h.[6] The principle of causality for classical fields[7] says that the part of spacetime that can be reached by causal worldlines through h should contain the domain of influence of h. This version of the principle of causality will be called **field causality**.

To determine whether a system of PDEs has this property, we need to understand something about the set of solutions to those PDEs. This makes the study of field causality – the subject of this article – more challenging than the study of worldline causality. For the rest of this article, the unqualified word **causality** refers to field causality.

[1] In the physics literature, the concept of **group velocity** is sometimes used in discussions about causality, but the concept of group velocity cannot generally be equated to the speed at which information propagates. Robinett (1978) highlights a counterexample.

[2] A worldline that is not spacelike anywhere is called a **causal** worldline.

[3] A **hypersurface** is a submanifold with one less dimension than the ambient manifold (Berger (2003), section 4.1.3.1). Here, the ambient manifold is spacetime, and we can take the initial hypersurface to be space at time $t = 0$.

[4] Section 2 explains what *initial data* means.

[5] Klainerman (2006), page 27

[6] Example: for equation (2) with $V = 0$ (the **wave equation**), when h is a single point, the boundary of its domain of influence is known to be what physicists call the **light cone** whose apex is at that point.

[7] Article 21916 introduces a version of causality that makes sense for *quantum* fields.

2 The family of models

In this article, $\phi(t, \mathbf{x})$ denotes a real-valued scalar field in Minkowski spacetime with one "time" coordinate t and D space coordinates $\mathbf{x} = (x_1, ..., x_D)$. The coordinate system is such that the equation for proper time τ is[8]

$$d\tau^2 = dt^2 - d\mathbf{x}^2. \tag{1}$$

Let $V(r)$ be a real-valued function of a single real variable r, and let $V'(r)$ denote its derivative with respect to r. This article is mostly about equations of motion of the form[9]

$$\ddot{\phi} - \nabla^2\phi + V'(\phi) = 0. \tag{2}$$

Each overhead dot is a derivative with respect to t, and ∇ is the gradient with respect to the spatial coordinates $\mathbf{x}$. Equation (1) implies that (2) has Lorentz symmetry.[9] Equation (2) implies that total energy[9]

$$\mathcal{E} \equiv \int d^Dx \; \left(\frac{\dot{\phi}^2(t, \mathbf{x}) + \big(\nabla\phi(t, \mathbf{x})\big)^2}{2} + V\big(\phi(t, \mathbf{x})\big) \right) \tag{3}$$

is independent of time.

In the context of equation (2), specifying **initial data** means specifying the values of $\phi(t, \mathbf{x})$ and $\dot{\phi}(t, \mathbf{x})$ at any one time (called the **initial time**).[10] Beware that even if the initial data is finite and smooth, the corresponding solution may not be defined everywhere in spacetime. Example: with $V = -\phi^4/4$, equation (2) has a solution $\phi(t, \mathbf{x}) = \sqrt{2}/(a - t)$, which is undefined at $t = a$. General questions about the existence, uniqueness, and domain of definition of solutions with the given initial data won't be addressed in this article.[11]

[8] Article 48968

[9] Article 49705. Equation (2) is sometimes called a **nonlinear Klein-Gordon equation**.

[10] More generally, initial data can be specified on a spacelike hypersurface, but see section 7.

[11] Examples of research in this area include Cazenave *et al* (2019), Bilgin and Kalantarov (2018), Chatzikaleas and Donninger (2017), and Keel and Tao (1999).

3 Causality for free scalar fields

A partial differential equation is called **linear** if any linear combination of solutions is also a solution. If $V \propto \phi^2$, then equation (2) is linear, and in this case the scalar field is called a **free**.[12] This section proves[13] that the free scalar field satisfies causality as defined in section 1 if $V \geq 0$.[14]

When (2) is nonlinear (example: when $V = \phi^4$), the calculation in this section doesn't quite prove causality, but it does prove a weaker property that I'll call **partial causality**: if a solution's initial data is nonzero only in a given region of space, then the region in which the solution is nonzero cannot grow any faster than a finite maximum speed if $V \geq 0$. If the equation of motion is linear, then partial causality implies causality, because the difference of two solutions is itself a solution, and because the difference is zero wherever the two original solutions are equal to each other.

Let H be the hypersurface $t = 0$, and let $h(0)$ be a subset of H. Define the **future causal completion**[15] of $h(0)$ to be the set of points in spacetime with $t \geq 0$ that cannot be reached by any causal worldline unless it also intersets $h(0)$.[16] Let $h(t)$ be the intersection of the causal completion of $h(0)$ with the constant-time hyperplane at the specified time t. Define

$$\mathcal{E}(t) \equiv \int_{\mathbf{x}\in h(t)} \left(\frac{\dot{\phi}^2 + (\nabla\phi)^2}{2} + V(\phi) \right). \tag{4}$$

The condition $V \geq 0$ implies that this is nonnegative and that it is zero if and only if ϕ and its time-derivative $\dot{\phi} \equiv \partial_t \phi$ are both zero in $h(t)$ at the specified time. Consider $t > 0$, so that $h(t)$ shrinks as t increases, and take the time-derivative of

[12] Article 30983 explains the reason for this name.

[13] Wang (2015) describes a generalization of this proof.

[14] Section 4 considers the "tachyonic" free scalar field, which has $V < 0$.

[15] Witten (2018)

[16] Example: if spacetime is $2+1$-dimensional and $h(0)$ is a disc, then its future causal completion is a cone whose based is $h(0)$ and whose apex is a point in the future of $h(0)$.

(4) to get

$$\begin{aligned}
\dot{\mathcal{E}} &= \int_{\mathbf{x}\in h(t)} \Big(\ddot{\phi}\dot{\phi} + (\nabla\phi)\cdot(\nabla\dot{\phi}) + V'(\phi)\dot{\phi}\Big) \\
&\qquad - \int_{\mathbf{x}\in\partial h(t)} \left(\frac{\dot{\phi}^2 + (\nabla\phi)^2}{2} + V(\phi)\right) \\
&= \int_{\mathbf{x}\in h(t)} \Big(\ddot{\phi} - \nabla^2\phi + V'(\phi)\Big)\dot{\phi} + \int_{\mathbf{x}\in h(t)} \nabla\cdot\big(\dot{\phi}\nabla\phi\big) \\
&\qquad - \int_{\mathbf{x}\in\partial h(t)} \left(\frac{\dot{\phi}^2 + (\nabla\phi)^2}{2} + V(\phi)\right),
\end{aligned}$$

where $\partial h(t)$ denotes the boundary of $h(t)$. The second integral in the first equation comes from the t-dependence of the integration domain $h(t)$, and this is where the derivation uses the fact that $h(t)$ is the causal completion of $h(0)$: as a function of t, the boundary of $h(t)$ moves inward with speed 1. If ϕ satisfies the equation of motion (2), then the first term in the second equation is zero. The second term in the second equation may be written as an integral over the boundary $\partial h(t)$, with unit normal $\mathbf{n}$. This gives

$$\begin{aligned}
\dot{\mathcal{E}} &= \int_{\mathbf{x}\in\partial h(t)} \mathbf{n}\cdot\big(\dot{\phi}\nabla\phi\big) - \int_{\mathbf{x}\in\partial h(t)} \left(\frac{\dot{\phi}^2 + (\nabla\phi)^2}{2} + V(\phi)\right) \\
&= -\int_{\mathbf{x}\in\partial h(t)} \left(\frac{\big(\mathbf{n}\dot{\phi} - \nabla\phi\big)^2}{2} + V(\phi)\right).
\end{aligned}$$

This is nonpositive, so $\mathcal{E}(t)$ must be a decreasing function of time for $t > 0$. But $\mathcal{E}(t)$ is nonnegative, so if it's zero initially, then it's zero for all $t > 0$, because it cannot decrease any further.

This proves partial causality: if the initial data is zero in a given spatial region $h(0)$ at time $t = 0$, then it remains zero at all points in spacetime with $t \geq 0$ that cannot be reached by any causal worldline that doesn't intersect $h(0)$. When the equation of motion is linear, this implies full causality.

4 Why require nonnegative energy?

The analysis in section 3 assumes $V \geq 0$, but causality may hold even if V does not have a lower bound. The real reason for requiring $V \geq 0$ in physics is **stability**. Causality doesn't require V to have a lower bound, but stability does.

Consider the case $V(\phi) = -(m^2/2)\phi^2$, where m is real-valued. The equation of motion (2) in this case is[17]

$$\ddot{\phi} - \nabla^2\phi - m^2\phi = 0. \tag{5}$$

Robinett (1978) shows that equation (5) satisfies causality,[18] so causality doesn't require V to have a lower bound.

Stability is a different issue. Define

$$\omega(\mathbf{p}) \equiv \begin{cases} \sqrt{\mathbf{p}^2 - m^2} & \text{if } \mathbf{p}^2 > m^2 \\ \sqrt{m^2 - \mathbf{p}^2} & \text{otherwise.} \end{cases}$$

Then equation (5) is satisfied by any function of the form

$$\begin{aligned} \phi(t,\mathbf{x}) = & \int_{\mathbf{p}^2>m^2} d^Dp\; e^{i\mathbf{p}\cdot\mathbf{x}} \left(f_1(\mathbf{p})e^{i\omega(\mathbf{p})t} + f_2(\mathbf{p})e^{-i\omega(\mathbf{p})t} \right) \\ & + \int_{\mathbf{p}^2<m^2} d^Dp\; e^{i\mathbf{p}\cdot\mathbf{x}} \left(f_3(\mathbf{p})e^{\omega(\mathbf{p})t} + f_4(\mathbf{p})e^{-\omega(\mathbf{p})t} \right). \end{aligned}$$

If the term involving $e^{\omega t}$ isn't zero,[19] then the solution's magnitude grows exponentially. In this sense, equation (5) is unstable.[20]

[17] This example has what is sometimes called "a mass term with the wrong sign." This example is often associated with literature about **tachyons**, a concept that is more relevant to the entertainment industry than it is to physics.

[18] This result is reviewed briefly in Garbarz and Palau (2021), section 5.4.

[19] If the solution's initial data is nonzero only in a finite region $h(0)$ of space at $t = 0$, then the terms involving $e^{\omega t}$ and $e^{-\omega t}$ must both be nonzero.

[20] The energy (3) is still conserved: the individual terms have different signs, so their exponentially growing contributions to the energy cancel each other.

5 Causality for the free electromagnetic field

For simplicity, most of this article focuses on scalar fields. Those are toy models: they share some features (like Lorentz symmetry) with more realistic models, but they are not intended to have realistic applications by themselves. For an example that does have realistic applications, this section shows that the free[21] electromagnetic field satisfies causality, using the same approach that was used for scalar fields in section 3.

Both for scalar fields and for the electromagnetic field, the first part of the proof can be expressed in terms of the stress-energy tensor T^{ab}. Explicit expressions for T^{ab} are given in articles 49705 and 78463 for scalar fields and the electromagnetic field, respectively. The indices a, b take values in $\{0, 1, 2, ..., D\}$. The component T^{00} is the energy density, and T^{0j} with $j \in \{1, 2, ..., D\}$ are the components of the momentum density. The equations of motion[22] imply the local conservation law

$$\partial_a T^{ab} = 0, \tag{6}$$

with an implied sum over the index $a \in \{0, 1, 2, ..., D\}$. To derive causality, define the time-dependent region $h(t)$ as in section 3, and consider the energy in that region:

$$\mathcal{E}(t) = \int_{h(t)} T^{00}. \tag{7}$$

Take the derivative of this with respect to t to get

$$\dot{\mathcal{E}} = \int_{h(t)} \partial_0 T^{00} - \int_{\partial h(t)} T^{00},$$

use (6) to get

$$\dot{\mathcal{E}} = -\int_{h(t)} \partial_j T^{j0} - \int_{\partial h(t)} T^{00},$$

[21] *Free* means that the field doesn't interact with anything else. This is a model in which charges and currents don't exist.

[22] For the scalar field, the equation of motion is equation (2). For the free electromagnetic field, the equations of motion are Maxwell's equations with no charges or currents (article 31738).

with an implied sum over the index $j \in \{1, 2, ..., D\}$. Now use integration-by-parts to get

$$\dot{\mathcal{E}} = \int_{\partial h(t)} n_j T^{j0} - \int_{\partial h(t)} T^{00}, \tag{8}$$

where n_j are the components of a unit vector field normal to $\partial h(t)$.

For a scalar field, the energy and momentum densities are (article 49705)

$$T^{00} = \frac{\dot{\phi}^2 + (\nabla\phi)^2}{2} + V(\phi) \qquad\qquad T^{0j} = T^{j0} = -\dot{\phi}\nabla_j\phi.$$

We can use these in (8) to reproduce the calculation in section 3. For the electromagnetic field, the energy and momentum densities are (article 78463)

$$\begin{aligned} T^{00} &= \frac{1}{2}\sum_k (E_k)^2 + \frac{1}{2}\sum_{j<k}(B_{jk})^2 \\ T^{0j} &= T^{j0} = \sum_k B_{jk}E_k, \end{aligned} \tag{9}$$

where E_k and $B_{jk} = -B_{kj}$ are the electric and magnetic components of the field. This expression for T^{00} implies

$$\mathcal{E} \geq 0. \tag{10}$$

To streamline the rest of the equations, use this matrix notation:

- n is a column matrix with components n_k.
- E is a column matrix with components E_k.
- B is a square matrix with components B_{jk}.
- The transpose of a matrix M is denoted M^T.

The fact that n is a unit vector implies that the square matrix $P \equiv nn^T$ is a projection matrix ($P^2 = P$ and $P^T = P$). Let I denote the identity matrix, and use the abbreviation $\overline{P} \equiv I - P$. With this notation, the terms in equations (9) are

$$\begin{aligned}
\sum_k (E_k)^2 &= E^T E = E^T P E + E^T \overline{P} E \\
\sum_{j<k} (B_{jk})^2 &= \frac{1}{2}\sum_{j,k}(B_{jk})^2 = \frac{1}{2}\langle BB^T\rangle \\
\sum_j n_j T^{j0} &= n^T B E
\end{aligned}$$

where $\langle M\rangle$ denotes the trace of M. The fact that B is antisymmetric ($B^T = -B$) implies $n^T B n = 0$, which may also be written $PBP = 0$. Use this to get

$$n^T B E = n^T B \overline{P} E$$

and

$$\begin{aligned}
\langle BB^T\rangle &= -\langle BB\rangle && \text{(use } B^T = -B\text{)} \\
&= -\big\langle (P+\overline{P})B(P+\overline{P})B\big\rangle && \text{(use } P+\overline{P} = I\text{)} \\
&= -\langle \overline{P}B\overline{P}B\rangle - \langle \overline{P}BPB\rangle - \langle PB\overline{P}B\rangle && \text{(use } PBP = 0\text{)} \\
&= -\langle \overline{P}B\overline{P}B\rangle - 2\,\langle \overline{P}BPB\rangle && \text{(use } \langle XY\rangle = \langle YX\rangle\text{)} \\
&= -\langle \overline{P}B\overline{P}B\rangle - 2\,\big\langle (P+\overline{P})BPB\big\rangle && \text{(use } PBP = 0\text{)} \\
&= -\langle \overline{P}B\overline{P}B\rangle - 2\,\langle BPB\rangle && \text{(use } P+\overline{P} = I\text{)} \\
&= \big\langle (\overline{P}B\overline{P})(\overline{P}B\overline{P})^T\big\rangle + 2\,(Bn)^T(Bn).
\end{aligned}$$

Use these equation (8) to get

$$\dot{\mathcal{E}} = -\frac{1}{2}E^T P E - \frac{1}{2}(\overline{P}E - Bn)^T(\overline{P}E - Bn) - \frac{1}{4}\big\langle (\overline{P}B\overline{P})(\overline{P}B\overline{P})^T\big\rangle < 0.$$

Together with (10), this shows that $\mathcal{E}$ is nonnegative and $\dot{\mathcal{E}}$ is nonpositive, which proves that $\dot{\mathcal{E}}$ must be zero if $\mathcal{E}$ is zero initially. As explained in section 3, this proves that the free electromagnetic field satisfies causality.

6 Nonlinear equations of motion and causality

When the equation of motion (2) is nonlinear, the calculation in section 3 proves what that section called partial causality, but not full causality, and only when $V(\phi)$ has a lower bound. When $V(\phi)$ doesn't have a lower bound, the calculation in section 3 doesn't prove anything at all.

Does causality hold for every PDE of the form (2), regardless of the function $V(\phi)$? This is a simple question, but I'm not aware of any theorem that answers it in complete generality.[23,24] In lieu of a general proof, the rest of this article describes two sources of intuition, both suggesting that the causality property probably does hold for arbitrary $V(\phi)$:

- Sections 7-13 use the concept of a **characteristic (hyper)surface** as a source of intuition suggesting that equation (2) satisfies causality.

- Sections 14-20 use a discretized (lattice) version of equation (2) as a source of intuition suggesting that the continuum version satisfies causality.

These sources of intuition complement the result that was already derived in section 3.

[23] All of the proofs I've found are either limited to the linear case or else prove only what section 3 calls *partial causality*, but maybe I just haven't looked in the right places. (The study of PDEs is not my specialty.) If you know of a general proof of causality for equation (2) with arbitrary V, you could post it as an answer to this question: https://math.stackexchange.com/questions/3746464.

[24] In the literature about PDEs, the name **finite speed of propagation** sometimes seems to refer to what I'm calling *causality* (Klainerman (2006), page 27), but more often it seems to be used as a synonym for what section 3 calls *partial causality* (Klainerman (2006), page 23, footnote 12, and Keel and Tao (1999), section 1).

7 Characteristic hypersurfaces: definition

Causality is a statement about the dependence of a solution on its initial data. Questions about *solutions* of nonlinear PDEs tend to be difficult, but we can gain some insight by asking this easier question about *initial data*: which hypersurfaces allow both the field and its first derivative to be specified without any constraints? To make the question more precise, let H be a hypersurface in $D+1$-dimensional spacetime, and choose a coordinate system in which one of the coordinates ξ is zero everywhere on H. The question is whether the field ϕ and its derivative $\partial_\xi\phi$ can both be specified on H without any constraints. This will be true only if equation (2) involves the second derivative $\partial_\xi^2\phi$. If it doesn't, then H is called a **characteristic (hyper)surface**.[25]

A relationship between characteristic hypersurfaces and causal completions[26] will be established in sections 12-13: if H is the hyperplane $t=0$, then the boundary of the future causal completion of a region $h\subset H$ is a characteristic hypersurface for equation (2). A connection with field causality – the main subject of this article – is suggested on page 28 in Klainerman (2006), which asserts for a hyperbolic PDE like equation (2),

> The boundaries of domains of dependence[27] ...are characteristic hypersurfaces...

The scope of that assertion isn't clear,[28] and no proof was given, but it is plausible because the relationship between a solution of equation (2) and its initial data on a characteristic hypersurface does involve some degree of arbitrariness. This will be illustrated in sections 8-10.

[25] Vitagliano (2014), section 1.2.1

[26] This was defined in section 3.

[27] The *domain of dependence* of a given region R consists of all points in the region's past to which the solution in R is sensitive. This is complementary to the *domain of influence*.

[28] The text says "This is a general fact," but it doesn't say how general.

8 Example: $1+1$-dimensional spacetime, part 1

The geometry is simplest in $1+1$-dimensional spacetime, so let's start there. In this case, the equation of motion (2) is

$$\partial_t^2\phi - \partial_x^2\phi + V'(\phi) = 0 \tag{11}$$

where ∂_t and ∂_x are the partial derivatives with respect to t and x. If we define

$$r \equiv (t-x)/2 \qquad\qquad s \equiv (t+x)/2,$$

then the partial derivatives with respect to these new variables are

$$\partial_r = \partial_t - \partial_x \qquad\qquad \partial_s = \partial_t + \partial_x,$$

so the equation of motion (11) may also be written[29]

$$\partial_r\partial_s\varphi + V'(\varphi) = 0 \qquad\qquad \text{with } \varphi(r,s) \equiv \phi(r+s,\, s-r). \tag{12}$$

If the V' term were absent, then this equation would imply that $\partial_r\varphi$ is independent of s and that $\partial_s\varphi$ is independent of r. That, in turn, would imply that every solution can be written as a sum of two functions, one depending only on r and one only on s. With a generic V' term, such an explicit description of the general solution is not available, but writing the equation of motion in the form (12) still leads to some insights: section 9 shows that the hypersurface defined by $r = 0$ is a characteristic hypersurface, and section 11 explains why this is related to the subject of causality.

[29] I'm using different symbols for ϕ and φ because they're expressed in different coordinate systems, so they are different functions of their respective arguments even though they're the same function of spacetime (article 09894).

9 Example: $1+1$-dimensional spacetime, part 2

Equation (11) involves the second derivative of ϕ with respect to t. As a consequence, when initial data is specified on the spacelike hypersurface defined by $t = 0$, the values of the functions ϕ and $\partial_t\phi$ can both be specified independently of each other at $t = 0$,[30] and any such choice determines a unique solution for $t \neq 0$.[31]

In contrast, equation (12) involves only the first derivative of φ with respect to r. As a consequence, when initial data is specified on the hypersurface $r = 0$, the values of φ and $\partial_r\varphi$ cannot be specified independently of each other on H: given the values of φ at $r = 0$, equation (12) is a condition that the values of $\partial_r\varphi$ must satisfy on H in order to be consistent with the specified values of φ.[32] (In the case $V' = 0$, the condition is that $\partial_r\varphi$ must be the same everywhere on H.) This shows that the hypersurface $r = 0$ is an example of a characteristic hypersurface.[33]

The next section generalizes this example to any number of dimensions.

[30] To make this precise, we should work within some appropriate class of functions, like the class of smooth functions. I won't try to be that precise here, but beware that the class of *analytic* functions is not sufficient for addressing causality, because the difference between two analytic functions (on the initial hypersurface) cannot be contained in any finite region of space. This follows from the fact that an analytic function has a unique analytic continuation (Hadamard (1923), chapter 1, page 11).

[31] To make this precise, we would need to account for the possibility that a solution may eventually become singular even if its initial data is nonsingular (section 7), but I won't try to be that precise here.

[32] To see this, use the identity $\partial_r\partial_s\varphi = \partial_s(\partial_r\varphi)$.

[33] By taking n derivatives of equation (12) with respect to r, we also get conditions that the values of $\partial_r^{n+1}\varphi$ must satisfy on H.

10 A generalization to arbitrary dimensions

The preceding example of a characteristic hypersurface can be generalized to $D+1$-dimensional spacetime for any D. To do this, choose one of the spatial coordinates and call it x. Write the other $D-1$ spatial coordinates collectively as $\mathbf{x}_\perp$, and write the field as

$$\phi\big(t,\, x,\, \mathbf{x}_\perp\big).$$

Then equation (2) may be written

$$\partial_t^2\phi - \partial_x^2\phi - \nabla_\perp^2\phi + V'(\phi) = 0.$$

With r and s defined as before, we can also write the equation of motion (2) as

$$\partial_r\partial_s\varphi = \nabla_\perp^2\varphi - V'(\varphi) \tag{13}$$

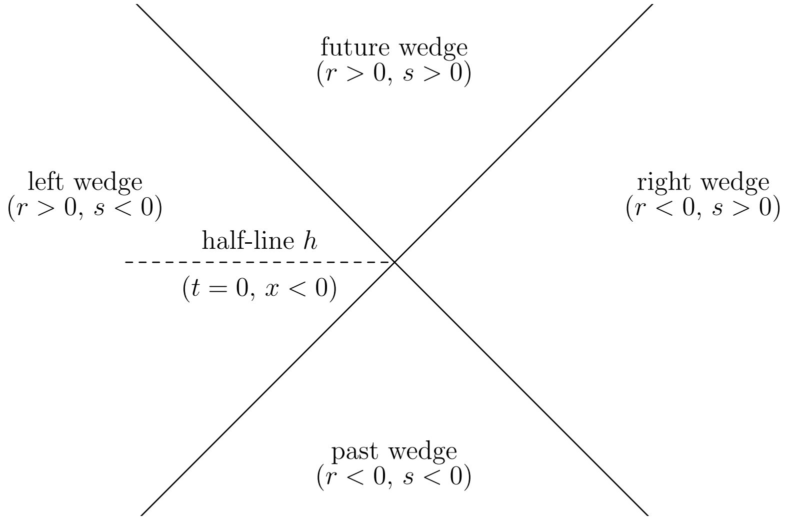
with

$$\varphi\big(r, s, \nabla_\perp\big) \equiv \phi\big(r+s,\, s-r,\, \mathbf{x}_\perp\big).$$

The reasoning that was used in the $1+1$-dimensional case (section 9) can be used again here to show that the hypersurface defined by $r=0$ is a characteristic hypersurface.

11 Characteristic hypersurfaces and causality

Sections 8-10 described an example of a characteristic hypersurface for equation (2). This sections shows that the relationship between a solution and its initial data on these characteristic hypersurfaces involves a degree of arbitrariness that suggests a connection with causality. For simplicity, this section considers only $1+1$-dimensional spacetime. This picture summarizes some of the terminology that will be used:[34]



The preceding sections showed that on the hypersurface $r=0$, the field φ and its derivative $\partial_r\varphi$ cannot be independently specified. On the other hand, specifying only φ on that hypersurface doesn't determine a unique solution of (12). This becomes obvious when the derivatives in equation (12) are replaced by finite differences:

$$\frac{1}{\epsilon_r}\left(\frac{\varphi(r+\epsilon_r,s+\epsilon_s)-\varphi(r+\epsilon_r,s)}{\epsilon_s}-\frac{\varphi(r,s+\epsilon_s)-\varphi(r,s)}{\epsilon_s}\right)+V'(\varphi)=0,$$

where ϵ_r and ϵ_s are the step-sizes in the r and s directions. Re-arrange this to get

$$\varphi(r+\epsilon_r,s+\epsilon_s)-\varphi(r+\epsilon_r,s)=\varphi(r,s+\epsilon_s)-\varphi(r,s)-\epsilon_r\epsilon_s V'(\varphi). \tag{14}$$

[34] In this picture, t increases upward and x increases to the right.

Suppose that a solution of (14) is specified at all points in the discretized spacetime with $r < 0$. In the original coordinate system, this is the region of spacetime defined by $x > t$, which covers the right and past wedges in the picture. Given a solution in that region, equation (14) almost tells us what the values of φ at the next step $r + \epsilon_r$ must be, but not quite: it allows us to add an arbitrary s-independent constant to all of the values at $r + \epsilon_r$. Equation (14) also allows us to add such an arbitrary s-independent constant at every subsequent step in the r-direction (every step toward the upper-left).[35] We can use this freedom to assign arbitrary values to the field along the half-line h defined by $t = 0$ and $x < 0$ (the dashed line in the picture). This means that the given solution in the region $x > t$ is consistent with *arbitrary* values of the field on h.

To establish causality, we need to show that a given solution in the right wedge is consistent with arbitrary values of both ϕ and $\partial_t\phi$ on h. With that motive, suppose that a solution is specified only in the right wedge (the intersection of $r < 0$ and $s > 0$). In this case, we can use equation (14) to determine the solution in the future wedge modulo an arbitrary s-independent constant at each step in the positive r-direction, and after rearranging equation (14), we can also use it to determine the solution in the past wedge modulo an arbitrary r-independent constant at each step in the negative s-direction. Those two sets of arbitrary constants, one independent of s and one independent of r, can be chosen to achieve consistency with arbitrary values of ϕ and $\partial_t\phi$ on h. This suggests that causality holds, at least with respect to initial data on h.

The virtue of this argument is that it doesn't depend on the form of $V(\phi)$ at all, so it suggests that causality holds for arbitrary V in $1 + 1$-dimensional spacetime. This argument isn't quite a proof, because it doesn't address exactly how solutions of the finite-difference equation (14) relate to solutions of the continuous-spacetime equation (2), but it at least passes an important test: it is consistent with (a discretized version of) the result that was derived in section 3.

[35] This constant doesn't need to be small compared to the step-size ϵ_r, so this implies that a solution of equation (12) can even be singular ($\partial_r\varphi$ may be infinite) at $r = 0$ in the continuum limit.

12 An equation for characteristic hypersurfaces

Motivated by the claim quoted in section 7 that the boundaries of domains of influence are characteristic hypersurfaces, this section derives a general equation for the characteristic hypersurfaces of the equation of motion (2). Section 13 will use this result to relate characteristic hypersurfaces to the geometry of causal completions.

Write the original coordinates t and $\mathbf{x}$ collectively as x^a (the superscript is an index, not an exponent), with $t = x^0$ and $\mathbf{x} = (x^1, x^2, ..., x^D)$. Let $\overline{x}^a$ be another coordinate system, and let H be the hypersurface defined by $\overline{x}^0 = 0$. Let ∂_a denote the partial derivative with respect to x^a, with the other coordinates in that system held fixed, and let $\overline{\partial}_a$ denote the partial derivative with respect to $\overline{x}^a$, with the other coordinates in that system held fixed.

In the original coordinate system, the equation of motion (2) is

$$\eta^{ab}\partial_a\partial_b\phi + V'(\phi) = 0,$$

where η^{ab} are the components of the Minkowski metric in the mostly-minus convention. To write this in the new coordinate system, use the identities

$$\partial_a = (\partial_a\overline{x}^c)\overline{\partial}_c$$

to get

$$\eta^{ab}(\partial_a\overline{x}^c)\overline{\partial}_c(\partial_b\overline{x}^d)\overline{\partial}_d\phi + V'(\phi) = 0.$$

The derivative $\overline{\partial}_c$ acts on everything to its right, so this may also be written

$$\eta^{ab}(\partial_a\overline{x}^c)(\partial_b\overline{x}^d)\overline{\partial}_c\overline{\partial}_d\phi + \eta^{ab}(\partial_a\overline{x}^c)(\overline{\partial}_c\partial_b\overline{x}^d)\overline{\partial}_d\phi + V'(\phi) = 0. \tag{15}$$

The hypersurface H defined by $\overline{x}^0 = 0$ is a characteristic hypersurface if the coefficient of the $\overline{\partial}_0\overline{\partial}_0\phi$ term is zero (section 7). Equation (15) says that this condition may be written

$$\eta^{ab}(\partial_a\overline{x}^0)(\partial_b\overline{x}^0) = 0. \tag{16}$$

Equation (16) says that gradient of $\overline{x}^0$ is lightlike (article 48968), so the hypersurface $\overline{x}^0 = 0$ is a characteristic hypersurface if the gradient of $\overline{x}^0$ is lightlike.

13 Relationship to causal completion

The previous section showed that if the gradient of a function $\overline{x}^0$ is lightlike where $\overline{x}^0 = 0$, then the hypersurface H defined by $\overline{x}^0 = 0$ is a characteristic hypersurface for equation (2). This section uses that result to show that the future boundary of a causal completion is a characteristic hypersurface.

First, here's some notation. Let H_0 be the hypersurface $t = 0$, and let h_0 be a subset of H_0. Let $\hat{h}_0$ be the future causal completion of h_0, the region of spacetime with $t \geq 0$ that cannot be reached by any causal worldline unless the worldline also intersects h_0. The boundary of $\hat{h}_0$ consists of two parts: one part h_0 at $t = 0$, and another part h_c with $t > 0$.[36]

To relate that geometry to equation (16), consider any neighborhood of a point on h_c, and let $\overline{x}^a$ be a coordinate system in which that part of h_c is defined by the condition $\overline{x}^0 = 0$. Because of the way h_c was defined, it cannot have a timelike tangent vector at any point, and it must have one lightlike tangent vector at every point.[37] The condition for v^a to be the components of a tangent vector is $v^a\partial_a\overline{x}^0 = 0$, which may also be written $\eta_{ab}v^a u^b = 0$ with $u^b \equiv \eta^{bc}\partial_c\overline{x}^0$. The condition $\eta_{ab}v^a u^b = 0$ must be satisfied by one lightlike vector v and must be violated by every timelike vector v. This is impossible unless u is lightlike, which implies equation (16). This proves that h_c is a characteristic hypersurface for equation (2).

When combined with the intuition described in section 11, this suggests that h_c might also be the boundary of the domain of influence for the complement of h_0 in H_0. This would be consistent with the quote in section 7.

[36] In the example described in footnote 16, h_c is the curved surface of the cone. More generally, h_c can be called a **conoid** – an imprecise term for something resembling a cone (Hadamard (1923), chapter 3, section 1, page 72).

[37] Intuitively: h_c and h_0 share the same boundary, and h_0 is spacelike. The definition of h_c means that it can be constructed by deforming h_0 into the future (keeping its boundary fixed) until it is just barely no longer spacelike everywhere.

14 Lattice models and maximum speed: outline

Section 11 used a lattice model that came from discretizing the derivatives with respect to the "lightlike" coordinates r and s. Section 15 introduces a different lattice model, one that comes from discretizing the derivatives with respect to the nominal time (t) and space ($\mathbf{x}$) coordinates instead. In this model, changing the value of the field at one point in space cannot affect the value of the field in another point in space if the number of steps in space exceeds the number of steps in time. In this sense, the region in which two solutions differ from each other cannot grow any faster than a finite maximum speed. This is true without any restrictions on $V(\phi)$, which suggests that equation (2) always satisfies causality.

To turn this into a proof, we would need to answer these questions:

- **Does the lattice model give equation (2) in the continuum limit?** For the special case $V = 0$, Courant *et al* (1928) shows that the appropriate continuum limit does exist if the coefficients in the lattice model satisfy an inequality called the CFL stability condition, and section 16 shows that the CFL stability condition is necessary. That's a warning that the existence of an appropriate continuum limit is not as obvious as it might naïvely seem to be. A comparable analysis is not available for general V, as far as I know.

- **Does the maximum speed remain finite in the continuum limit?** Section 17 argues heuristically that the answer is yes – if the appropriate continuum limit exists. To test the reasoning, sections 18-19 analyze a different kind of equation, one that doesn't have a finite maximum speed in continuous spacetime even though its lattice version does. In that case, the same reasoning correctly predicts the absence of any finite maximum speed.

The key would be to prove that an appropriate continuum limit exists, which I don't know how to do when the equation of motion is nonlinear. As a substitute, section 20 presents computer results that would be surprising if an appropriate continuum limit didn't exist.

15 The lattice model

Let ϵ_t and ϵ_x denote the lattice spacings in the time and space directions, respectively. (Section 16 will explain why we should allow them to be different.) Let $\mathbf{e}_1, ..., \mathbf{e}_D$ be basis vectors for the spatial lattice, each with magnitude ϵ_x. The derivatives in the equation of motion (2) can be discretized like this:[38]

$$\begin{aligned}\ddot{\phi}(t,\mathbf{x}) &\to \frac{1}{\epsilon_t}\left(\frac{\phi(t+\epsilon_t,\mathbf{x})-\phi(t,\mathbf{x})}{\epsilon_t}-\frac{\phi(t,\mathbf{x})-\phi(t-\epsilon_t,\mathbf{x})}{\epsilon_t}\right)\\ &= \frac{\phi(t+\epsilon_t,\mathbf{x})+\phi(t-\epsilon_t,\mathbf{x})-2\phi(t,\mathbf{x})}{\epsilon_t^2}\\ \nabla^2\phi(t,\mathbf{x}) &\to \sum_j \frac{1}{\epsilon_x}\left(\frac{\phi(t,\mathbf{x}+\mathbf{e}_j)-\phi(t,\mathbf{x})}{\epsilon_x}-\frac{\phi(t,\mathbf{x})-\phi(t,\mathbf{x}-\mathbf{e}_j)}{\epsilon_x}\right)\\ &= \sum_j \frac{\phi(t,\mathbf{x}+\mathbf{e}_j)+\phi(t,\mathbf{x}-\mathbf{e}_j)-2\phi(t,\mathbf{x})}{\epsilon_x^2}.\end{aligned} \tag{17}$$

That gives this discrete version of the equation of motion (2):

$$\begin{aligned}\phi(t+\epsilon_t,\mathbf{x}) = 2\phi(t,\mathbf{x}) &- \phi(t-\epsilon_t,\mathbf{x})\\ &+\beta\sum_j\big(\phi(t,\mathbf{x}+\mathbf{e}_j)+\phi(t,\mathbf{x}-\mathbf{e}_j)-2\phi(t,\mathbf{x})\big)\\ &-\epsilon_t^2 V'\big(\phi(t,\mathbf{x})\big)\end{aligned} \tag{18}$$

with

$$\beta \equiv \left(\frac{\epsilon_t}{\epsilon_x}\right)^2. \tag{19}$$

Equation (18) determines the values of ϕ for all spatial points $\mathbf{x}$ at time $t+\epsilon_t$, if the values of ϕ at times t and $t-\epsilon_t$ are specified. In the continuous-spacetime limit, this corresponds to specifying the initial values of ϕ and $\dot{\phi}$.

[38] Article 71852

16 When does a continuum limit exist?

In the continuous-spacetime version (2), we can change the relative coefficients of the time- and space-derivative terms just by changing the units in which the time and space coordinates are expressed. Such a change of units has no real effect, because it can be compensated by rescaling the independent variables t and $\mathbf{x}$. This might lead us to expect that the value of β isn't really important, but that's not quite true.

This section considers the simplest case $V = 0$, for which equation (2) is called the **wave equation**. This section shows that when $V = 0$, the lattice version (18) is **unstable** when $\beta > 1/D$, meaning that a typical solution diverges exponentially as time passes, even for initial data that is relatively smooth compared to the spatial step size ϵ_x. We can avoid that instability by taking $\beta \leq 1/D$, in which case the lattice version of the wave equation becomes equivalent to its continuous-spacetime version in the appropriate limit, as long as we consider initial data ϕ and $\dot{\phi}$ that remain smooth as $\epsilon_x \to 0$ (with β held fixed).

To show that the lattice wave equation is unstable when $\beta > 1/D$, consider the ansatz[39]

$$\phi(n\epsilon_t, \mathbf{x}) = z^n \exp(i\mathbf{p} \cdot \mathbf{x}), \tag{20}$$

where the components of $\mathbf{p}$ are real constants and z is a complex number to be determined. Substitute this ansatz into the $V = 0$ version of equation (18) to get this equation for z:

$$z = 2 - \frac{1}{z} + \beta \sum_j \big(2\cos(\mathbf{p} \cdot \mathbf{e}_j) - 2\big),$$

which may be rearranged to get the quadratic equation

$$z^2 + 2\gamma z + 1 = 0 \tag{21}$$

[39] This approach to analyzing stability is called **von Neumann stability analysis**.

with

$$\gamma \equiv \beta \sum_j \big(1 - \cos(\mathbf{p}\cdot\mathbf{e}_j)\big) - 1. \tag{22}$$

The solutions of (21) are

$$z = -\gamma \pm \sqrt{\gamma^2 - 1}. \tag{23}$$

When $\gamma^2 \leq 1$, the two solutions for z are both complex numbers with magnitude $|z| = 1$. In this case, the magnitude of (20) remains constant in time. But when $\gamma^2 > 1$, the two solutions (23) are both real numbers, and one of them has magnitude greater than 1. In this case, the magnitude of (20) grows exponentially in time, growing by a factor of $z > 1$ with every discrete time-step. That's disastrous for a continuum limit in which any nonzero time-interval corresponds to infinitely many discrete time-steps. This shows that the condition $\gamma^2 \leq 1$ is a prerequisite for a sensible continuum limit. Use

$$0 \leq 1 - \cos\theta \leq 2$$

in equation (22) to see that the condition $\gamma^2 \leq 1$ holds for all $\mathbf{p}$ only if

$$\beta \leq \frac{1}{D}. \tag{24}$$

If β doesn't satisfy this condition, then $\gamma^2 > 1$ whenever the components of $\mathbf{p}$ are all small enough compared to $1/\epsilon_x$. Any initial data that becomes smooth in the continuum limit may be expressed (using a Fourier transform) as a superposition of solutions of the form (20) with such values of $\mathbf{p}$, so this shows that (24) is a necessary condition for a sensible continuum limit.

This result is called the **Courant-Friedrichs-Lewy (CFL) stability condition**. It's named after the authors of Courant *et al* (1928), who also showed that if the condition (24) is satisfied, then the wave equation does have a continuum limit that agrees with the continuous-spacetime version of the wave equation.

17 Maximum speed in the continuum limit

Equation (18) determines the values of ϕ for all spatial points $\mathbf{x}$ at time $t+\epsilon_t$, if the values of ϕ at times t and $t-\epsilon_t$ are specified. According to that equation, the effect of a disturbance at one point cannot propagate any faster than one space-step per time-step, so the maximum possible speed is

$$v_{\text{max}} \equiv \frac{\epsilon_x}{\epsilon_t}. \tag{25}$$

What happens to this restriction in the continuum limit?

First, we need to think about what "continuum limit" means. We could use the limits $\epsilon_t \to 0$ and $\epsilon_x \to 0$ as the definitions of the partial derivatives in the original equation of motion (2), but that would presume that the function $\phi(t,\mathbf{x})$ is defined for a continuum of values of t and $\mathbf{x}$, so that the arguments of $\phi(t,\mathbf{x})$ can be varied continuously. Here, a different perspective will be used: in the lattice model, the field $\phi(t,\mathbf{x})$ is defined only where its arguments are integer multiples of the fixed quantities ϵ_t and ϵ_x, and taking the "continuum limit" means considering field configurations that change only infinitesimally between consecutive values of those discrete arguments. For such configurations, a time T or distance X over which the field changes by a finite (not infinitesimal) amount must be infinite when expressed in units of ϵ_t or ϵ_x, but we can use "macroscopic" units in which T and X are finite. Units are arbitrary, but not completely arbitrary: we must choose the "macroscopic" units so that the time and space derivatives of the field are *finite* when expressed in those units.

Speed is a comparison between a spatial scale and a time scale ("How far does something move in a given amount of time?"). The question is whether the speed at which field disturbances propagate is finite or infinite when expressed in macroscopic units, given that the derivatives of the field are all finite when expressed in those units. The answer to this question depends on the form of the equation of motion. If we choose the configuration at two consecutive times $t-\epsilon_t$ and t, then equation (18) tells us what the configuration must be at all subsequent times. This means that the relationship between the macroscopic units of time and space is not arbitrary: it's constrained by equation (18) if we want the time and space derivatives of the field to be finite when expressed in macroscopic units, because that equation determines the solution's time-dependence.

In the special case $V = 0$, this constraint on the relationship between the macroscopic units of time and space must be expressible in terms of the quantity β defined in (19), because equation (18) depends on the parameters ϵ_t and ϵ_x only through β (when $V = 0$). This implies that β must be finite when expressed in macroscopic units. Since $\beta = 1/v_{\text{max}}^2$, this means that v_{max} must be finite when expressed in macroscopic units. In other words, when $V = 0$, the continuum limit of equation (18) satisfies causality: a disturbance in the field cannot propagate faster than a maximum speed that is *finite* when expressed in macroscopic units.

Section 18 tests this intuition by applying it to another model, namely the **heat equation**, which does not satisfy the causality principle. When applied to a lattice version of the heat equation, the same line of reasoning that was used above leads to the correct conclusion that the maximum speed should be *infinite* when expressed in macroscopic units.

The purpose of considering this intuition is that we might be able to use it even when $V \neq 0$ in equation (18). The word *might* is in that sentence because when $V \neq 0$, the quantity β by itself does not account for all of the ϵ_t and ϵ_x dependence in equation (18). However, the V term in equation (18) does not contribute directly to propagation: if the other terms on the right-hand side of equation (18) were absent, then the field could not propagate. Heuristically, this suggests[40] that the constraint on the relationship between the macroscopic units of time and space should depend only on β, so the qualitative conclusion reached above for the case $V = 0$ should still hold when $V \neq 0$.

[40] This is the weak link in the argument.

18 Testing the intuition, part 1

Section 17 used intuition to deduce that equation (2) has a finite maximum speed. To test that intuition, this section applies it to a different lattice model, one whose continuum version does not have a finite maximum speed even though the lattice model does. In this section, space is one-dimensional ($D = 1$), so $\mathbf{x}$ has only one component x.

The lattice model considered here is similar to the $V = 0$ version of (18), but with only one time-derivative instead of two:

$$\frac{\varphi(t+\epsilon_t,x)-\varphi(t,x)}{\epsilon_t} = \frac{\alpha}{\epsilon_x}\left(\frac{\varphi(t,x+\epsilon_x)-\varphi(t,x)}{\epsilon_x} - \frac{\varphi(t,x)-\varphi(t,x-\epsilon_x)}{\epsilon_x}\right), \tag{26}$$

which can be rearranged like this:

$$\varphi(t+\epsilon_t,x) = \varphi(t,x) + \frac{\alpha\epsilon_t}{\epsilon_x^2}\big(\varphi(t,x+\epsilon_x)+\varphi(t,x-\epsilon_x)-\varphi(t,x)\big). \tag{27}$$

This is a lattice version of the **heat equation**,[41] whose continuum version will be considered in section 19. Like the other lattice model (18), this one has a finite maximum speed: disturbances introduced at one point cannot propagate any faster than

$$v_{\text{max}} \equiv \frac{\epsilon_x}{\epsilon_t}, \tag{28}$$

just like equation (25). Just like in section 17, the question is whether this speed is finite or infinite when expressed in macroscopic units, given that the derivatives of the field are all finite when expressed in those units. We can answer this using the same intuition that was used in section 17, but here that intuition leads to a different conclusion.

If we choose the configuration of the field at one time, then equation (27) tells us what the configuration must be at later times. This means that equation (27)

[41] This is not the best discretization for numerical approximations. Discretizations that are better for that purpose are described in Bender and Tovbis (1997).

may constrain the relationship between the macroscopic units of time and space, if we want the time and space derivatives of the field to be finite when expressed in those units. This constraint on the relationship between the macroscopic units of time and space must be expressible in terms of the quantity

$$\frac{\alpha\epsilon_t}{\epsilon_x^2} = \frac{\alpha}{v_{\max}\epsilon_x}, \tag{29}$$

because equation (27) depends on the parameters ϵ_t and ϵ_x only through this quantity. This implies that the quantity (29) must be finite when expressed in macroscopic units. But the factor ϵ_x in (29) is *zero* when expressed in macroscopic units, so the quantity $v_{\max}$ must be *infinite* when expressed in macroscopic units. This is only an upper bound,[42] but leaves us with no reason to expect the continuum limit of equation (27) to satisfy causality, if the continuum limit exists at all. This conclusion is correct: the next section shows that the continuum version of (26) does not satisfy causality.

[42] Footnote **??** in section 17

19 Testing the intuition, part 2

In two-dimensional spacetime, the **heat equation** is

$$\frac{\partial\varphi}{\partial t}=\alpha\frac{\partial^2\varphi}{\partial x^2}, \tag{30}$$

with constant $\alpha > 0$. The lattice model studied in the previous section is a discretization of this. Equation (30) does not have any maximum speed, even though its lattice version does (equation 28). To see that the heat equation (30) does not have any finite maximum speed, consider the function

$$\varphi(t,x)=\frac{1}{\sqrt{t}}\int_{-\infty}^{\infty}dy\;f(y)\exp\left(-\frac{(x-y)^2}{4\alpha t}\right). \tag{31}$$

This function satisfies the heat equation (30) for all $t > 0$, because the left- and right-hand sides of (30) are both equal to

$$\frac{1}{\sqrt{t}}\int_{-\infty}^{\infty}dy\;f(y)\exp\left(-\frac{(x-y)^2}{4\alpha t}\right)\left(-\frac{1}{2t}+\frac{(x-y)^2}{4\alpha t^2}\right).$$

To deduce what happens at $t = 0$, use these properties:

- As $t \to 0$, the exponential factor in (31) approaches zero whenever $x-y\neq 0$, so the function $f(y)$ in the integrand might as well be replaced with $f(x)$ in that limit.

- If the function $f(y)$ is replaced with $f(x)$, then the integral is independent of t. (The proof is easy: just change the integration variable y to absorb both x and t.) Most importantly, this shows that the limit $t \to 0$ is finite.

Altogether, this shows that (31) satisfies the heat equation (30) with the initial condition $\varphi(0,x)\propto f(x)$. Now suppose that $\varphi(0,x)$ is nonzero only within a finite interval, say

$$\varphi(0,x)=f(x)=\begin{cases}1 & \text{if } -1<x<1,\\ 0 & \text{otherwise.}\end{cases}$$

Even with this initial condition, the solution (31) is nonzero for all $-\infty < x < \infty$ as soon as $t > 0$. This is clear from the fact that the integrand is positive for all x whenever $t > 0$. This shows that the heat equation (30) does not satisfy the causality principle: a disturbance in one location immediately affects the solution in all other locations. The effect in distant locations may be very small, but this article is concerned with *strict* causality, so any nonzero effect counts.

20 The emergence of isotropy

The reasoning in section 17 suggests that equation (2) has a finite maximum speed for any V, assuming that the lattice model (18) reproduces (2) in the continuum limit. However, the maximum speed deduced using that reasoning is only an upper bound. It can't be a tight bound. The reason is simple: when $D \geq 2$, equation (2) has rotation symmetry, but the lattice model (18) does not. In particular, the maximum speed in equation (2) is isotropic (the same in all directions), but the maximum speed in equation (18) is not. If equation (2) emerges from (18), then this discrepancy must resolve itself somehow. On the other hand, if equation (2) doesn't emerge from (18), then we would have no reason to expect this discrepancy to resolve itself. If computer calculations with the lattice model show a tendency for the maximum speed to become isotropic, then this is evidence that the lattice model does reproduce (2) in the continuum limit.

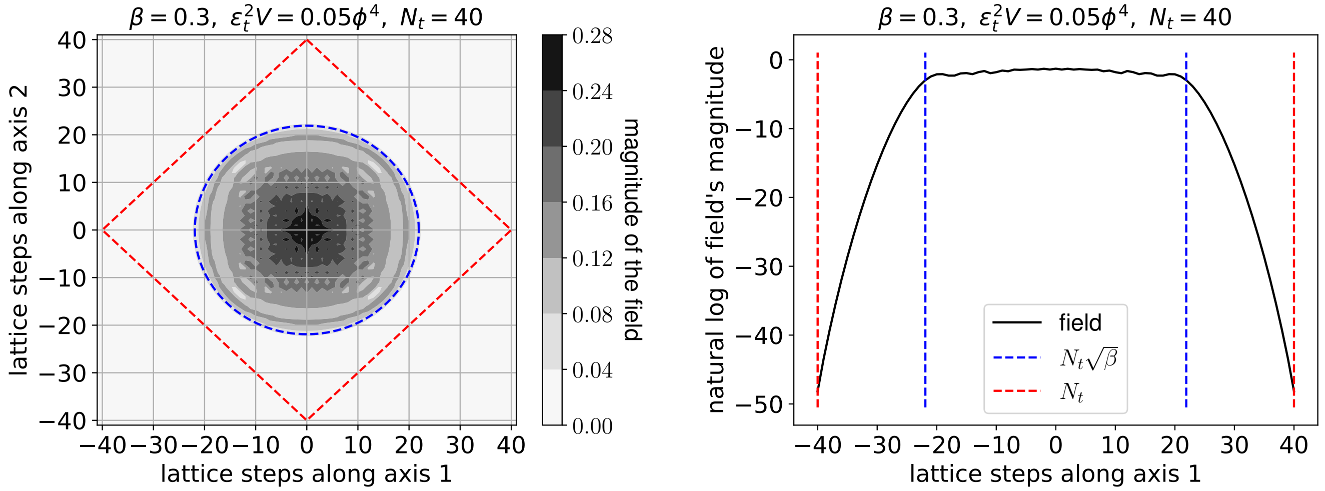
The results shown below were generated with a computer,[43] using the $D = 2$ version of the lattice equation (18) to propagate the field forward in time, starting with a configuration that is zero everywhere except at a single point. The pictures show a tendency for the field's magnitude to be concentrated within a (growing) *circle*, even though equation (18) only strictly constrains it to a (growing) diamond-shaped region that contains the circle. This tendency for an isotropic speed limit to emerge is what we would expect if the lattice model reduces to the original equation of motion (2) in the continuum limit. Otherwise, we would have no reason to expect such isotropy to emerge. In that sense, these computer results may be regarded as evidence (not proof) that the lattice model has a continuum limit consistent with the original equation of motion (2) even when $V \neq 0$.

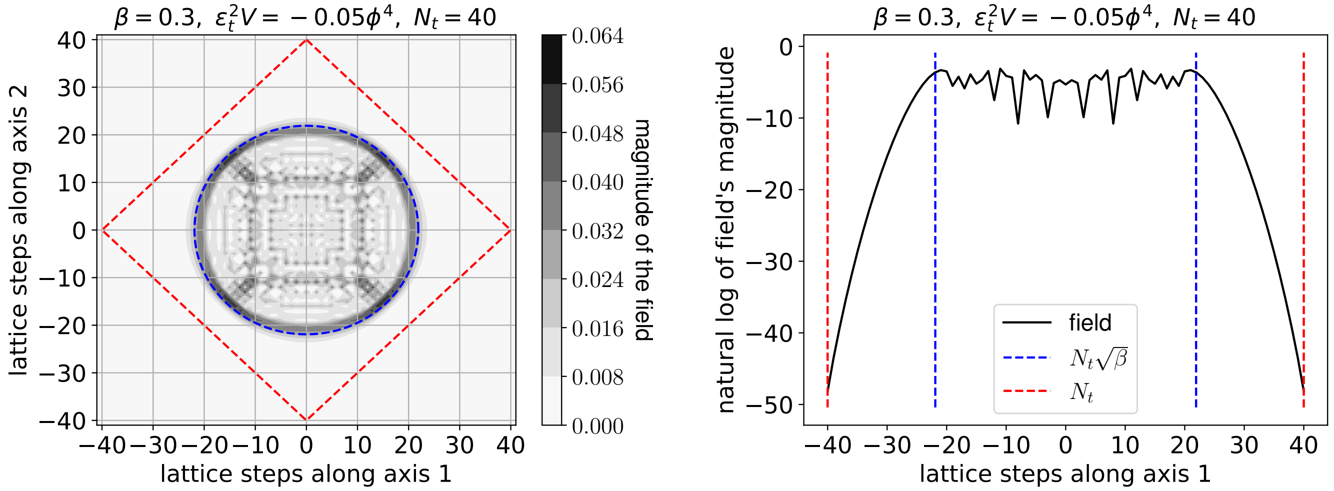
Each picture on the left shows a contour plot of the field's magnitude after N_t time steps, using the values of β, V, and N_t indicated in the figure's title. The dashed red diamond is the boundary of the set of points that can be reached from the center in N_t steps. The dashed blue circle has radius $N_t\sqrt{\beta}$, the significance of which is explained below. Each picture on the right shows a slice along a canonical

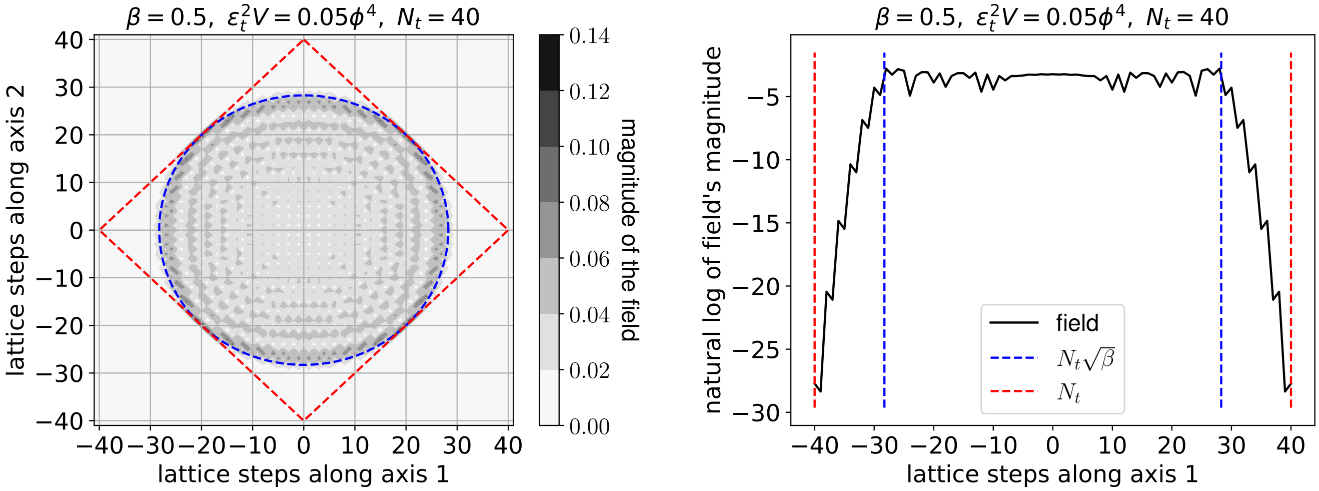
[43] The source code is posted here: https://cphysics.org/extras/98038a.html

axis, to show how quickly the (natural log of the) field's magnitude decreases beyond the radius $N_t\sqrt{\beta}$.

All three pairs of pictures use $V \propto \phi^4$, which makes the equation of motion (2) nonlinear. The first two pairs of pictures use $\beta = 0.3$, with $V > 0$ and $V < 0$, respectively. The third pair of pictures uses $\beta = 0.5$, which is the maximum value of β allowed by the CFL stability condition (24).







These pictures show that the field's magnitude tends to be concentrated inside a circle of radius $N_t\sqrt{\beta}$, even though points outside of this circle (but inside the diamond) can be reached from the center in N_t steps. This is significant because it is consistent with the emergence of an *isotropic* maximum speed as the continuum limit is approached, if such a limit exists at all. If such a limit didn't exist, then we would have no reason to expect isotropy to emerge, so these pictures can be regarded as evidence that a continuum limit does exist, even when the equation of motion is nonlinear and even when V doesn't have a lower bound.

The radius $N_t\sqrt{\beta}$ can be explained intuitively. The factor of β in equation (18) affects the number of time steps that are needed in order to accumulate a *significant* effect at the point $\mathbf{x}$, given a disturbance that occurred at the origin. Decreasing the value of β increases the number of time steps that are needed. (Example: if $\beta \lll 1$, then a large number of time-steps would be needed even just to accumulate a significant effect at the origin's nearest neighbors.) Assuming that an isotropic continuum limit exists, the region in which the field's magnitude is significant should approach an isotropic shape (a circle when $D = 2$) when N_t increases. For a given value of N_t, the intuition just described says that the radius of this circle should be an increasing function of β. To quantify this, we can use dimensional analysis: the only increasing function of β that is proportional to N_t and has the correct dimensions is $\propto N_t\sqrt{\beta}$. To fix the proportionality factor, consider the case

$\beta = 1$: in this case, equation (18) says that each time step carries the field's magnitude as-is from one point in space to the next, so the proportionality factor should be 1. Altogether, this intuition suggests the radius inside which the field's magnitude can be significant should be $N_t\sqrt{\beta}$. This is consistent with the pictures shown above.

To test this intuition, we can ask whether it's consistent with the CFL condition (section 16). Choose some point $\mathbf{x}$ in the spatial lattice, and let $N_\mathbf{x}$ denote the minimum number of spatial-lattice-steps needed to reach the point $N_\mathbf{x}$, starting from the origin. According to equation (18), a disturbance at the origin can affect the field at $\mathbf{x}$ if and only if the number N_t of time steps is $N_t \geq N_\mathbf{x}$. For the case $D = 2$, the boundary of this region is indicated in the pictures by the dashed red diamond. Along a canonical axis, the maximum distance that can be reached in N_t steps is N_t, expressed in units of ϵ_x. Along a diagonal (same number of steps in each of the D spatial dimensions), the pythagorean theorem says that the maximum distance that can be reached in N_t steps is $N_t/\sqrt{D}$, because the N_t steps must be divided equally among the D dimensions. This is a strict bound, independent of the value of β, but the intuition in the preceding paragraph suggests that the effect at $\mathbf{x}$ can be significant out to a radius $N_t\sqrt{\beta}$. That intuition assumed the existence of an isotropic continuum limit, so the existence of an isotropic continuum limit should be expected only if this β-dependent radius complies with the strict bound. This gives

$$N_t\sqrt{\beta} \leq N_t/\sqrt{D} \qquad \Rightarrow \qquad \beta \leq \frac{1}{D},$$

which is precisely the CFL condition that was derived rigorously in section 16, so the intuition that was used in the previous paragraph passes this test.

Altogether, the computer results shown in this section may be regarded as evidence (not proof) that equation (2) satisfies causality for any V, even if V doesn't have a lower bound.[44] This complements the evidence given in section 3, which was limited to the case $V \geq 0$, and it's consistent with the evidence given in section 12.

[44] Stability is a separate issue, as illustrated in section 4.

21 Huygen's principle in various dimensions

The previous sections were all focused on the principle of causality, which is a statement about what can(not) happen *outside* a region's light cone. For extra fun, this section mentions something interesting about what happens *inside* the light cone when $V = 0$ in equation (2).

When $V = 0$, equation (2) reduces to the **wave equation**. In three-dimensional space ($D = 3$), the wave equation satisfies **Huygen's principle**.[45] Precise statements of the principle are given in Balakrishnan (2004a) and in theorem 5.8 of Ben-Artzi (2015). Here's a rough translation: a wave equation satisfies Huygen's principle if a pulse remains a pulse – if a configuration of the field that is initially confined to a tiny region of space remains confined to a tiny neighborhood of that region's light cone, without leaving any "residue" deeper inside the light cone.

Huygen's principle depends on D, the number of dimensions of space, in an interesting way. The calculations reviewed in Balakrishnan (2004a) and Balakrishnan (2004b) show that Huygen's principle holds only for $D \in \{3, 5, 7, 9, ...\}$. In other words, it holds only when the number of dimensions of space is *odd* and *no less than three*. For other values of D, a pulse does not remain a pulse: as it propagates, it leaves a residue of nonzero magnitude inside the light cone.

[45] The name *Huygen's principle* seems to be used inconsistently, for at least two related-but-different things. The version I'm highlighting here is sometimes called the *Strong* Huygen's principle.

22 References

Balakrishnan, 2004a. “Wave propagation: odd is better, but three is best. 1. The formal solution of the wave equation” *Resonance* **9**: 30-38, https://www.ias.ac.in/article/fulltext/reso/009/06/0030-0038

Balakrishnan, 2004b. “Wave propagation: odd is better, but three is best. 2. Propagation in spaces of different dimensions” *Resonance* **9**: 8-17, http://repository.ias.ac.in/1079

Ben-Artzi, 2015. “Linear Wave Equations” https://jbenartzi.github.io/2015.Dispersive/files/5.pdf

Bender and Tovbis, 1997. “Continuum limit of lattice approximation schemes” *Journal of Mathematical Physics* **38**: 3700-3717, https://stars.library.ucf.edu/cgi/viewcontent.cgi?article=2847&context=facultybib1990

Bilgin and Kalantarov, 2018. “Non-existence of global solutions to nonlinear wave equations with positive initial energy” *Communications on Pure and Applied Analysis* **17**: 987-999

Berger, 2003. *A Panoramic View of Riemannian Geometry.* Springer

Cazenave *et al*, 2019. “Solutions with prescribed local blow-up surface for the nonlinear wave equation” https://arxiv.org/abs/1904.03893

Chatzikaleas and Donninger, 2017. “Stable blowup for the cubic wave equation in higher dimensions” https://arxiv.org/abs/1712.03833

Courant *et al*, 1928. “On the Partial Difference Equations of Mathematical Physics” republished in *IBM Journal of Research and Development* **11**: 215-234, http://www.stanford.edu/class/cme324/classics/courant-friedrichs-le
pdf

Garbarz and Palau, 2021. "A note on Haag duality" https://arxiv.org/abs/2108.01257

Hadamard, 1923. *Lectures on Cauchy's Problem in Linear Partial Differential Equations.* New Haven Yale University Press, https://archive.org/details/lecturesoncauchy00hadauoft

Keel and Tao, 1999. "Small data blowup for semilinear Klein-Gordon equations" *American Journal of Mathematics* **121**: 629-669, http://www.math.ucla.edu/~tao/preprints/blowup.ps.Z

Klainerman, 2006. "Partial Differential Equations" https://web.math.princeton.edu/~seri/homepage/papers/gowers-Aug4-2006.pdf

Robinett, 1978. "Do tachyons travel more slowly than light?" *Phys. Rev. D* **18**: 3610-3616, https://journals.aps.org/prd/abstract/10.1103/PhysRevD.18.3610

Vitagliano, 2014. "Characteristics, Bicharacteristics, and Geometric Singularities of Solutions of PDEs" *Int. J. Geom. Meth. Mod. Phys.* **11**: 1460039, https://arxiv.org/abs/1311.3477

Wang, 2015. "Lectures on Nonlinear Wave Equations" http://people.maths.ox.ac.uk/wangq1/Lecture_notes/nonlinear_wave_9.pdf

Witten, 2018. "Notes on Some Entanglement Properties of Quantum Field Theory" *Rev. Mod. Phys.* **90**: 45003, https://arxiv.org/abs/1803.04993

23 References in this series

Article **09894** (https://cphysics.org/article/09894):
"Tensor Fields on Smooth Manifolds" (version 2023-11-12)

Article **21916** (https://cphysics.org/article/21916):
"Local Observables in Quantum Field Theory" (version 2023-11-12)

Article **30983** (https://cphysics.org/article/30983):
"The Free Scalar Quantum Field: Particles" (version 2023-11-12)

Article **31738** (https://cphysics.org/article/31738):
"The Electromagnetic Field and Maxwell's Equations" (version 2022-02-18)

Article **48968** (https://cphysics.org/article/48968):
"The Geometry of Spacetime" (version 2024-02-25)

Article **49705** (https://cphysics.org/article/49705):
"Classical Scalar Fields and Local Conservation Laws" (version 2023-11-12)

Article **71852** (https://cphysics.org/article/71852):
"Treating Space as a Lattice" (version 2022-08-21)

Article **78463** (https://cphysics.org/article/78463):
"Energy, Momentum, and Angular Momentum in Classical Electrodynamics" (version 2023-12-15)