SIAM J. Appl. Math.
Vol. 80, No. 6, pp. 2560–2585

CONVOLUTIONAL NEURAL NETWORKS IN PHASE SPACE AND INVERSE PROBLEMS[*]

GUNTHER UHLMANN[†] AND YIRAN WANG[‡]

Abstract. We study inverse problems consisting of determining medium properties using the responses to probing waves from the machine learning point of view. Based on the analysis of propagation of waves and their nonlinear interactions, we construct a deep convolutional neural network to reconstruct the coefficients of nonlinear wave equations that model the medium properties. Furthermore, for given approximation accuracy, we obtain the depth and number of units of the network and their quantitative dependence on the complexity of the medium.

Key words. convolutional neural networks, inverse problems

AMS subject classification. 35R30

DOI. 10.1137/19M1294484

1. Introduction.

Consider nonlinear acoustic wave equations on $\mathbb{R}^3$ of the form

$$\begin{aligned}(\partial_t^2 - c^2(x)\Delta)u(t,x) + F(t,x,u(t,x)) = f(t,x), \quad t > 0, \quad x \in \mathbb{R}^3,\\ u(t,x) = 0, \quad t \leq 0, \quad x \in \mathbb{R}^3,\end{aligned} \tag{1.1}$$

where $c(x) > 0$ is the wave speed, $f(t,x)$ is the source term, $\Delta = \sum_{i=1}^3 \partial_{x^i}^2$ is the Laplacian on $\mathbb{R}^3$, and $F(t,x,u)$ is a smooth function in t, x, and u. We are mainly interested in the case when F is nonlinear in u. As we only consider local problems, we assume that $c(x)$ is nontrapping without loss of generality. One can think of (1.1) as modeling acoustic waves generated by the source $f(t,x)$ traveling in a medium with certain nonlinear mechanism. The coefficients $c(x), F(t,x,u)$ characterize linear and nonlinear properties of the medium. The inverse problem that we formulate precisely in section 2 is the determination of the wave speed $c(x)$ and the nonlinear term $F(t,x,u)$ by measuring the response of waves traveling through the medium. From the machine learning point of view, the inverse problem is to learn material properties (characterized by c and F) from the data (the source and wave responses). In this work, our goal is to develop a neural network structure for solving both the forward and the inverse problems. We remark that our method applies to a large class of hyperbolic equations on manifolds describing wave phenomena; however, for simplicity, we will use (1.1) as the example.

The linearized problem, that is, when $F(t,x,u)$ is linear in u and typically with the hyperbolic Dirichlet-to-Neumann data, has been studied extensively in the literature. There are well-developed methods such as the boundary control (BC) method; see [15] for an overview. However, the nonlinear problems that we study in this work are not always solvable by linearization. Some recent progress has been made towards

[*] Received by the editors October 21, 2019; accepted for publication (in revised form) September 21, 2020; published electronically December 15, 2020.
https://doi.org/10.1137/19M1294484

Funding: The work of the first author was partially supported by NSF, a Walker Family Professorship at UW, and a Si-Yuan Professorship at HKUST.

[†] Department of Mathematics, University of Washington, Seattle, WA 98195 USA, and Institute for Advanced Study, the Hong Kong University of Science and Technology, Clear Water Bay, New Territories, Hong Kong, China (gunther@math.washington.edu).

[‡] Department of Mathematics, Emory University, Atlanta, GA 30322 USA (yrwang.math@gmail.com).

solving these problems by exploiting the nonlinear interactions of waves, beginning with the work of Kurylev, Lassas, and Uhlmann [17]. The phenomenon that nonlinear interactions of waves could generate new waves has been known for a while and has been observed in many physical experiments. Mathematically this phenomenon has been studied from the point of view of interactions of singularities by the notable work of Bony [2], Melrose and Ritter [29], and Rauch and Reed [34] among others. See also Beals [1] for an overview of the subject in the 1980s and 1990s. The idea introduced in [17] is that by using distorted plane waves concentrated near fixed directions, one can keep track of their interactions and the newly generated waves. One can characterize the "features" of these waves in the data, which eventually leads to the determination of the parameters of the equation.

The numerical method we develop for solving this problem is based on these ideas in the theoretical work. From the machine learning/deep learning point of view, one natural attempt is to construct a neural network for the map from the data to c, F directly. However, this data driven approach leaves everything in a black box (and hence less interpretable) and requires a large amount of data. As we will see later, for the inverse problem that we consider, acquisition of full data is unlikely. Thus our approach is to combine the physical model (1.1) and the data. The idea bears some similarity to the PDE learning approach, which has gained considerable popularity recently. Among many works in this direction, we mention the physics-informed network in [33], DeepONet in [26], and PDE-Net 2.0 in [25]; see also [35] for more PDE motivated networks. Our study belongs to this category, but we address issues related to wave phenomena for (1.1). At this point, we mainly focus on theoretical questions about the network design and properties. We hope to address the numerical implementation and practical issues elsewhere.

The informal version of our main theorem is the following main result.

Main Result. *We construct a deep convolutional neural network (in section* 9*, see Figure* 3*) with M levels, K units on each level, and parameter set Θ such that for all data (f, u) where u is the solution of* (1.1) *(in some open set $\mathscr{V}$) with source f compactly supported and $\|f\| < \epsilon$, the network generates an approximation function $h(f; \Theta)$ satisfying*

$$\|u - h(f; \Theta)\| < C_M \epsilon^M.$$

The norms are specified in Theorem 10.1*. The numbers K and M depend on the complexity of $c(t, x), F(t, x, u)$. The parameters Θ can be used to reconstruct $c(x), F(t, x, u)$ (see section* 10*).*

In constructing the network, we make use of the similarity between the iteration scheme for solving nonlinear equations and the deep forward network. This leads to an interpretation of the nonlinear effects of the activation function and a better choice of the activation function for the wave equation. Roughly speaking, the units in each level of the neural network represent small wave units. Units in deeper levels capture the effects of wave propagation and nonlinear interactions. In fact, for general source term f (not necessarily distorted plane waves), we think of it as consisting of sufficiently many small wave units, and the network captures the interaction among them. This point of view resembles that in applied harmonic analysis; see, for instance, [6, 16]. From the theoretical work mentioned above, it is natural to work in the phase space and consider the high frequency information in the wave units, which we take as the "features" in this machine learning problem. A main part of the construction is to show how these "features" interact with each other and propagate through the network. This is novel even for the conventional convolutional network; see [27].

As a result of the analysis, the depth and number of units of the network can be expressed in terms of the complexity of the parameter functions $c(x), F(t,x,u)$. For highly nonlinear functions, a deeper network should be used to reveal such effects and produce better approximations.

The organization of the paper is as follows. In section 2, we formulate the inverse problem to be considered in this article. Then we compare the iteration schemes of the deep forward networks (section 3) and for solving wave equations (section 4). We propose the basic network structure in section 5 and discuss some issues in the architecture related to the wave phenomena. The resolution of these issues leads us to the convolutional neural network in section 9, but before that, we need to discuss the nonlinear interactions of conormal waves (section 6), the estimates of linear propagation (section 7), and nonlinear effects (section 8). Finally, we prove the approximation properties of the network in section 10. The paper finishes with some concluding remarks in section 11.

2. The inverse problem.

We consider two types of inverse problems for (1.1) with different types of data. In this work, we shall work with the first problem exclusively.

2.1. The source perturbation problem.

Let $f(t,x)$ be compactly supported. For $T > 0$ fixed, consider

$$\begin{aligned}(\partial_t^2 - c^2(x)\Delta)u(t,x) + F(t,x,u) = f(t,x), \quad t \in (-\infty, T], \quad x \in \mathbb{R}^3,\\ u(t,x) = 0, \quad t \leq 0, \quad x \in \mathbb{R}^3.\end{aligned} \tag{2.1}$$

It is known (also see section 4 for a precise statement) that for $f \in H^s([0,T]\times\mathbb{R}^3)$ sufficiently small and compactly supported, there is a unique solution $u \in H^{s+1}([0,T]\times\mathbb{R}^3)$. We denote this solution map by $u = L(f)$. We remark that we do not pursue the optimal regularity result in this work.

We want to determine c and F in the region where the wave can travel to. It is convenient to formulate the problem using the space-time nature of wave propagation. Let

$$g = -dt^2 + c^{-2}(x)dx^2$$

be the Lorentzian metric so that the corresponding d'Alembert operator is $\Box_g = \partial_t^2 - c^2(x)\Delta$. We denote $\mathscr{M} = \mathbb{R}^4$ and consider the Lorentzian manifold $(\mathscr{M}, g)$. Let $\widehat{\mu}(s) \subset \mathscr{M}$ be a time-like geodesic where $s \in [-1,1]$. In general relativity, this represents the world line of a freely falling observer. Let $\mathscr{V} \subset \mathscr{M}$ be an open relatively compact neighborhood of $\widehat{\mu}([s_-, s_+])$ where $-1 < s_- < s_+ < 1$. We denote $\mathscr{M}(T) = (-\infty, T] \times \mathbb{R}^3$ and choose $T > 0$ such that $\mathscr{V} \subset \mathscr{M}(T)$. Let $p_\pm = \widehat{\mu}(s_\pm)$. See the left of Figure 1 .

We recall some notions of causalities. For $p, q \in \mathscr{M}$, we denote $p \ll q$ $(p \leq q)$ if $p \neq q$ and p can be joined to q by a future pointing time-like (causal) curve. We denote $p \leq q$ if $p = q$ or $p < q$. The chronological future of $p \in M$ is the set $I^+(p) = \{q \in \mathscr{M} : p \ll q\}$. The causal future of $p \in M$ is $J^+(p) = \{q \in \mathscr{M} : q \leq p\}$. The chronological past and causal past are denoted by $I^-(p)$ and $J^-(p)$, respectively. For any set $A \subset \mathscr{M}$, we denote the causal future by $J^\pm(A) = \cup_{p\in A} J^\pm(p)$. Also, we denote $J(p,q) = J^+(p) \cap J^-(q)$ and $I(p,q) = I^+(p) \cap I^-(q)$.

Let f be supported in $\mathscr{V}$, and we measure the wave u in $\mathscr{V}$. We consider the data set

$$\mathcal{D}_{sour,\epsilon} \doteq \{(f, u|_{\mathscr{V}}) : u = L(f), f \in H^s_{comp}(\mathscr{V}), s > 1, \|f\|_{H^s} < \epsilon\},$$

where $\epsilon > 0$ is small such that the well-posedness of (2.1) is guaranteed. The inverse problem is to determine $c(x)$ and $F(t, x, u)$ on $I(p_-, p_+)$ from $\mathcal{D}_{sour,\epsilon}$; see Figure 1. Notice that $I(p_-, p_+)$ is the largest set that the wave u can travel to from $\mathscr{V}$ and return to $\mathscr{V}$. The goal of this work is to construct a neural network to approximate the map $f \to L(f)|_{\mathscr{V}}$ from which we find approximations of c, F. We observe that here the data consists of all possible pairs $(f, u|_{\mathscr{V}})$.

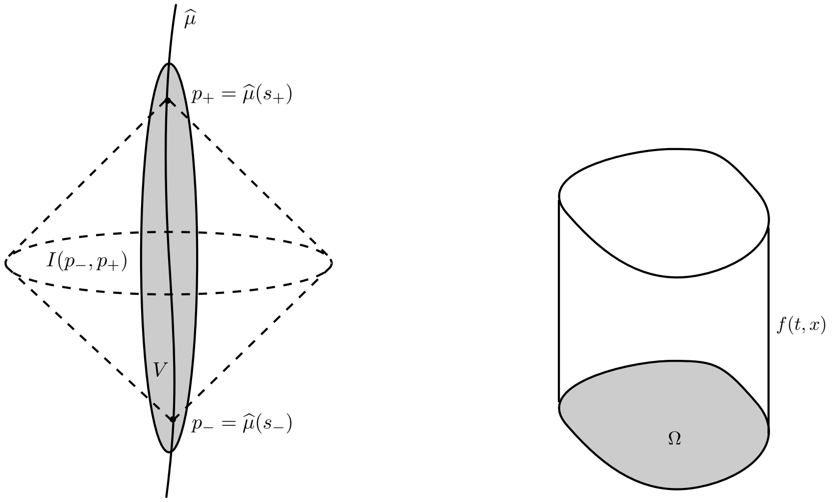

FIG. 1. *Two types of inverse problem. Left: The source perturbation problem. Right: The boundary value problem.*

We remark that this formulation of the inverse problem was introduced for the Einstein equations in [18], which has a concrete physical interpretation, that is, to determine space-time structures (e.g., topological, differentiable structure and the metric) from actively generated gravitational perturbations measured near a freely falling observer. In fact, the Einstein equation in wave gauge is a second order quasilinear hyperbolic system. The problem has been further studied in [22] for Einstein–Maxwell equations and [38] for more general source fields. One of our motivations is to develop an algorithm to understand the gravitational wave interactions in these works. For semilinear wave equations on globally hyperbolic Lorentzian manifolds, the problem was studied in [17] and [21].

2.2. The hyperbolic Dirichlet-to-Neumann problem. For the second type of inverse problem the information is given on the boundary. We consider the wave equation (1.1) on a bounded domain $\Omega \subset \mathbb{R}^3$ with smooth boundary $\partial\Omega$. See the right of Figure 1. For fixed $T > 0$, consider

$$\begin{gathered}(\partial_t^2 - c^2(x)\Delta)u(t,x) + F(t,x,u(t,x)) = 0, \quad (t,x) \in [0,T] \times \Omega, \\ u(t,x) = f(t,x), \quad t \le T, x \in \partial\Omega, \\ u(t,x) = 0, \quad t \le 0, x \in \Omega.\end{gathered} \tag{2.2}$$

For $f \in H^s([0,T] \times \partial\Omega)$ sufficiently small and regular, and compactly supported, the problem is well-posed. See, for example, [5] for the treatment of Cauchy data and [8, 31]. We can define the Dirichlet-to-Neumann map as

$$\Lambda(f) = \nu \cdot \partial u|_{[0,T]\times\partial\Omega},$$

where ν is the outward normal vector to $\partial\Omega$. We consider the data set

$$\mathcal{D}_{DtN,\epsilon} \doteq \{(f, \Lambda(f)) : f \in H^s_{comp}([0,T] \times \partial\Omega), s > 1, \|f\|_{H^s} < \epsilon\},$$

where $\epsilon > 0$ is small such that (2.2) is well-posed. The inverse problem is to determine $c(x)$ and $F(t, x, u)$ from this data set. We remark that on unbounded domain, one can formulate the problem as a scattering problem. Again, our goal here is to construct a neural network to approximate the map $f \to \Lambda(f)$, from which we find approximations of c, F.

For this setup, Nakamura and Watanabe [31] considered the one dimensional quasilinear wave equation, which is further generalized in Nakamura and Vashisth [32] for systems in one dimension. The nonlinear elastic system is of particular interest because of its applications in geophysics and rock sciences. For example, one is interested in determining the underground formation of the Earth using nonlinear responses of seismic waves because the contrast in nonlinear parameters is stronger than that in linear ones; see [19]. In de Hoop, Uhlmann, and Wang [8], the authors analyzed the nonlinear interaction of two elastic waves, and the inverse problems of determining elastic parameters is addressed as well.

3. Review of deep feedforward networks.

We briefly review the iteration scheme of deep feedforward networks and refer the reader to [10] for more details. In general, the goal of deep feedforward networks is to approximate some function $y = \mathcal{F}^*(x) : \mathbb{R}^n \to \mathbb{R}^m$. For example, in classification problems, the function returns the number of classes the data x belongs to. The feedforward network defines a mapping $y = \mathcal{F}(x; \theta)$ in which θ is the parameter. The value of the parameter θ can be learned by solving an optimization problem. The result $\mathcal{F}(x, \theta)$ is an approximation of the function $\mathcal{F}^*(x)$.

There are many variants of deep feedforward networks. We illustrate using the multi-layer perceptrons (MLPs). The construction of an MLP consists of a sequence of compositions of linear mappings (the perceptron) followed by nonlinear maps called the activation function. Usually, the linear mapping is taken to be the affine transformation

$$\alpha(x; \theta) = Ax + b,$$

where $A \in \mathbb{R}^n$ and $b \in \mathbb{R}$ consist of the parameter $\theta = \{A, b\}$. There are many choices of the activation function in practice. A commonly used one is the rectified linear unit (ReLU)

$$\rho(t) = \max\{0, t\}, \quad t \in \mathbb{R}.$$

To build the MLP, we start from the input data x and call it $h^{(0)} = x$. This forms the first layer of the network. We introduce K units,

$$\widetilde{h}^{(1)}_k = \alpha(h^{(0)}, \theta^{(1)}_k) = A^{(1)}_k h^{(0)} + b^{(1)}_k, \quad k = 1, 2, \ldots, K,$$

where $\theta^{(1)}_k = \{A^{(1)}_k, b^{(1)}_k\}$. Then let $h^{(1)}_k = \rho(\alpha(h^{(0)}, \theta^{(1)}_k))$, where ρ applies componentwisely. We obtain $h^{(1)} = (h^{(1)}_k)^K_{k=1}$ as the second layer, also called the hidden layer of the network. One must realize that without the activation function, $h^{(1)}$ would be just a linear function of $h^{(0)}$. We continue this process and obtain schematically

$$h^{(n)} = \rho(\alpha(h^{(n-1)}; \theta^{(n)})), \quad n = 1, 2, \ldots, N,$$

which defines the nth layer of the network. Here, N is called the depth of the network and $\theta^{(n)} = \{A^{(n)}_k, b^{(n)}_k, k = 1, 2, \ldots, K\}$ denotes the collection of parameters at level n.

Eventually, we obtain the approximation function from $h^{(N)}$. For example, $\mathcal{F}(x;\theta) = \sum_{k=1}^{K} \beta_k h_k^{(N)}$, where θ is the collection of parameters $\theta^{(n)}, n = 1, 2, \dots, N$, and $\beta_k, k = 1, \dots, K$.

To find the parameters θ, we need to solve an optimization problem on some training data set $\mathscr{X}$. The cost function can be formulated as

$$J(\theta) = \sum_{x \in \mathscr{X}} ||\mathcal{F}^*(x) - \mathcal{F}(x;\theta)||^2$$

in the l^2 norm for vector spaces. The cost function is usually nonconvex and the problem is solved by a gradient descent based method such as stochastic gradient descent and back-propagation. We refer the readers to [10, 23] for more information.

On the theoretical level, it is important to understand the approximation ability of the network. It is shown in [11, 12, 24] that MLPs can approximate any Borel measurable function, which is known as the universal approximation property. We recall and formulate two such theorems below with special attention to the regularity of activation functions. The first theorem is for continuous activation functions, and the second one is for C^l activation functions.

Let $\mathcal{G} : \mathbb{R} \to \mathbb{R}$ be a Borel measurable function. Let $\mathscr{A}$ be the set of affine transformations from $\mathbb{R}^n$ to $\mathbb{R}$; namely $\alpha \in \mathscr{A}$ means $\alpha(x) = Ax + b, x \in \mathbb{R}^n$ with some $A \in \mathbb{R}^n, b \in \mathbb{R}$. We define

$$\Sigma(\mathcal{G}) = \left\{ f : \mathbb{R}^n \to \mathbb{R} | f(x) = \sum_{j=1}^{q} \beta_j \mathcal{G}(\alpha_j(x)), x \in \mathbb{R}^n, \beta_j \in \mathbb{R}, \alpha_j \in \mathscr{A}, q = 1, 2, \dots \right\}.$$

THEOREM 3.1 (Theorem 2.1 of [11]). *Let $\mathcal{G} : \mathbb{R} \to \mathbb{R}$ be a continuous nonconstant function. Then for any compact set $\mathscr{K} \subset \mathbb{R}^n$, $\Sigma(\mathcal{G})$ is dense in $C^0(\mathscr{K})$ with respect to*

$$\delta_{\mathscr{K}}(f, g) = \sup_{x \in \mathscr{K}} |f(x) - g(x)|, \quad f, g \in C^0(\mathbb{R}^n).$$

THEOREM 3.2 (Corollary 3.4 of [12]). *Let $\mathcal{G} : \mathbb{R} \to \mathbb{R}$ be a function in $C^l(\mathbb{R})$ with nonnegative integer l satisfying*

$$\int_{\mathbb{R}^n} \left| \frac{d^l}{dx^l} \mathcal{G}(x) \right| dx < \infty. \tag{3.1}$$

Then for any compact set $\mathscr{K} \subset \mathbb{R}^n$, $\Sigma(\mathcal{G})$ is dense in $H^m(\mathbb{R}^n)$ for $m \le l$ with respect to

$$\delta_{\mathscr{K}}^m(f, g) = \sum_{|\alpha| \le m} \sup_{x \in \mathscr{K}} |D^\alpha f(x) - D^\alpha g(x)|, \quad f, g \in H^m(\mathbb{R}^n).$$

Another important variant of feedforward networks is the convolutional neural network where the affine transformation is replaced by convolutions. See [23, 28]. We refer readers to [10] for the motivation and advantages of this type of network.

4. The iteration scheme. We review the iteration scheme for solving nonlinear wave equations because it shares some similarities to the architecture in deep feedforward networks. The material is rather classical; however, we want to show in this section to what extent each iteration step reveals nonlinear effects. To illustrate the idea, we take the polynomial nonlinear function

$$F(t, x, u) = a(t, x)u^2 + b(t, x)u^3 + c(t, x)u^4$$

as an example. The coefficients a, b, c reflect the nonlinearity in increasing orders. Also, we shall consider the small source perturbation problem for the wave equation

$$Pu(t,x) + F(t,x,u) = \epsilon f(t,x), \quad (t,x) \in \mathscr{M}(T), \tag{4.1}$$

where f is compactly supported, ϵ is a small parameter, and we denote the linear wave operator by $P = \partial_t^2 - c^2(x)\Delta$. These two simplifications will be removed eventually.

Let v be the solution of the linearized equation on $\mathscr{M}(T)$,

$$Pv = f.$$

It is well known that there is a fundamental solution $Q = P^{-1}$. We write $v = Q(f)$. Let u be the solution of (4.1). Then formally we have

$$P(u - \epsilon v) + F(u) = 0 \Longrightarrow u = \epsilon v - Q(F(u)).$$

Here, we omitted the dependence of F on t, x in the notation. Now we let $u^{(1)} = \epsilon v$ be the linearized solution and set

$$\begin{aligned} u^{(2)} &= \epsilon v - Q(F(u^{(1)})) \\ &= \epsilon v - \epsilon^2 Q(av^2) + O(\epsilon^3). \end{aligned}$$

We observe that modulo $O(\epsilon^3)$ terms, the coefficients a appear in $u^{(2)}$ and this is associated with the quadratic nonlinearity. We continue this procedure to get

$$\begin{aligned} u^{(3)} &= \epsilon v - Q(F(u^{(2)})) \\ &= \epsilon v - \epsilon^2 Q(av^2) + 2\epsilon^3 Q(avQ(av^2)) - \epsilon^3 Q(bv^3) + O(\epsilon^4), \end{aligned}$$

and another iteration gives

$$\begin{aligned} u^{(4)} &= \epsilon v - Q(F(u^{(3)})) \\ &= \epsilon v - \epsilon^2 Q(av^2) + 2\epsilon^3 Q(avQ(av^2)) - \epsilon^3 Q(bv^3) \\ &+ \epsilon^4[-Q(cv^4) + 2Q(avQ(bv^3)) + 3Q(bv^2 Q(av^2)) - 4Q(avQ(avQ(av^2)))] + O(\epsilon^5). \end{aligned}$$

The point is that for each $i = 1, 2, 3$, modulo $O(\epsilon^i)$ terms, we should expect to see the nonlinear coefficients in $u^{(i)}$. One continues the procedure to obtain the sequence $u^{(n)}$. The fact is that $u^{(n)}$ converges to the solution u in a proper sense.

Proposition 4.1. *Consider the nonlinear wave equation*

$$\begin{gathered} Pu(t,x) + F(t,x,u(t,x)) = f(t,x), \quad (t,x) \in \mathscr{M}(T), \\ u(t,x) = 0, \quad (t,x) \in \mathscr{M}(0). \end{gathered}$$

We assume that F is a smooth function with $F(t,x,0) = F_u(t,x,0) = 0$. For fixed $T > 0$, there exists ϵ_0 such that for f compactly supported in $\mathscr{M}(T)\backslash\mathscr{M}(0)$ with $\|f\|_{H^s(\mathscr{M})} \le \epsilon, s > 1, 0 < \epsilon < \epsilon_0$, the sequence $u^{(n)}$ defined iteratively by

$$u^{(1)} = Q(f), \quad u^{(n)} = u^{(0)} - Q(F(t,x,u^{(n-1)})),\ n \ge 2,$$

converges to a unique solution $u \in H^{s+1}(\mathscr{M}(T))$. Moreover, we have the estimates

$$\|u^{(n)} - u\|_{H^{s+1}(\mathscr{M}(T))} < C_n \epsilon^n, \quad \|u\|_{H^{s+1}(\mathscr{M}(T))} < C\epsilon,$$

where $C_n, C > 0$, depends on c, F and C_n depends on n as well.

Proof. First, we recall that $Q : H^s_{comp}(\mathscr{M}(T)) \to H^{s+1}_{loc}(\mathscr{M}(T))$ is bounded; see, for example, [7, Proposition 5.6]. So there is $C_Q > 0$ depending on c such that

$$\|Qf\|_{H^{s+1}} \le C_Q \|f\|_{H^s}.$$

For $s > 1$, the space $H^{s+1}(\mathscr{M}(T))$ is closed under multiplication. Moreover, $F(t,x,u) \in H^{s+1}(\mathscr{M}(T))$ for any smooth function F and $u \in H^{s+1}(\mathscr{M}(T))$; see [36]. By Sobolev embedding, $H^{s+1}(\mathscr{M}(T)) \subset C^r(\mathscr{M}(T))$ with $r < s-1$. In particular, $u^{(n)} \in H^{s+1}(\mathscr{M}(T)) \subset C^0(\mathscr{M}(T))$ are continuous for $s > 1$.

We want to show that $u^{(n)}$ form a Cauchy sequence. For convenience, we take $u^{(0)} = 0$. Suppose f is supported in a compact set $K \subset \mathscr{M}(T)$. By finite speed of propagation for linear wave equations, we know that each $u^{(n)}, n \ge 1$, is supported in $J_+(K)$. We shall assume $u^{(n)}$ is supported in $J_+(K) \cap \mathscr{M}(T)$.

Now we consider $u^{(m)} - u^{(n)}, m, n \ge 1$, satisfying

$$\begin{gathered} P(u^{(m)} - u^{(n)}) = -[F(t,x,u^{(m-1)}) - F(t,x,u^{(n-1)})], \quad (t,x) \in \mathscr{M}(T), \\ u^{(m)} - u^{(n)} = 0, \quad (t,x) \in \mathscr{M}(0). \end{gathered}$$

We obtain

$$\|u^{(m)} - u^{(n)}\|_{H^{s+1}} \le C_Q \|F(t,x,u^{(m-1)}) - F(t,x,u^{(n-1)})\|_{H^s}. \tag{4.2}$$

First we take $n = 1$ to get

$$\|u^{(m)} - u^{(1)}\|_{H^{s+1}} \le C_Q \|F(t,x,u^{(m-1)})\|_{H^s}.$$

Then we write

$$F(t,x,u^{(m-1)}) = \left(\frac{1}{2}\int_0^1 \partial_u^2 F(t,x,\tau u^{(m-1)}) dt\right)(u^{(m-1)})^2.$$

Because F is smooth, and $u^{(n)} \in H^{s+1}$, we can use Moser-type estimates (see [37, Proposition 3.9], which also works for $F(t,x,u)$ by minor modifications of the proof) to obtain that for $(t,x) \in J_+(K)$ and $\tau \in [0,1]$,

$$\|\partial_u^2 F(t,x,u^{(m-1)}) - \partial_u^2 F(t,x,0)\|_\infty \le C\|u^{(m-1)}\|_\infty (1 + \|u^{(m-1)}\|_{H^{s+1}}),$$

where C depends on $|\partial_u^k F(t,x,u)|$ for $k \le s+1$. Thus,

$$\|\partial_u^2 F(t,x,u^{(m-1)})\|_\infty \le C_F + C\|u^{(m-1)}\|_\infty (1 + \|u^{(m-1)}\|_{H^{s+1}}),$$

and we have

$$\|u^{(m)} - u^{(1)}\|_{H^{s+1}} \le C_Q[C_F + C\|u^{(m-1)}\|_\infty (1 + \|u^{(m-1)}\|_{H^{s+1}})]\|u^{(m-1)}\|^2_{H^{s+1}}.$$

Now we use induction and assume that $\|u^{(m-1)} - u^{(1)}\|_{H^{s+1}} < \epsilon$ for ϵ sufficiently small. This implies that $\|u^{(m-1)}\|_{H^{s+1}} \le C_0\epsilon$ for some constant C_0. We see that

$$\|u^{(m)} - u^{(1)}\|_{H^{s+1}} \le \epsilon\left(\epsilon C_0^2 C_Q[C_F + C\epsilon(1+\epsilon)]\right). \tag{4.3}$$

So we just need to take $\epsilon < \epsilon_0$ with $\epsilon_0 C_0^2 C_Q[C_F + C\epsilon_0(1+\epsilon_0)] < 1$, and we obtain $\|u^{(m)} - u^{(1)}\|_{H^{s+1}} < \epsilon$. This finishes the induction and shows that

$$\|u^{(m)}\|_{H^{s+1}} \le C_0\epsilon \tag{4.4}$$

are bounded for all m.

Next, we return to (4.2) and write

$$F(t,x,u^{(m-1)}) - F(t,x,u^{(n-1)}) = \Big(\int_0^1 \partial_u F(t,x,u^{(m-1)} + \tau u^{(n-1)})dt\Big)(u^{(m-1)} - u^{(n-1)}).$$

As $\partial_u F(t,x,0) = 0$, we use Moser-type estimates again to get

$$\|\partial_u F(t,x,u^{(m-1)}+\tau u^{(n-1)})\|_\infty \leq C\|u^{(m-1)}+\tau u^{(n-1)}\|_\infty(1+\|u^{(m-1)}+\tau u^{(n-1)}\|_{H^{s+1}}) \leq C_1\epsilon$$

for all $\tau \in [0,1]$. Now we use the fact that $u^{(n)}$ are continuous and bounded on $J_+(K)$ to get

$$\|u^{(m)} - u^{(n)}\|_{H^{s+1}} \leq C_Q C_F C_1 \epsilon \|u^{(m-1)} - u^{(n-1)}\|_{H^{s+1}},$$

which implies that for $m > n$

$$\|u^{(m)} - u^{(n)}\|_{H^{s+1}} \leq (C_Q C_F C_1 \epsilon)^{n-1}\|u^{(m-n+1)} - u^{(1)}\|_{H^{s+1}} \leq (C_Q C_F C_1 \epsilon)^{n-1}\epsilon,$$

where we have used (4.3). By possibly shrinking ϵ_0 further so that $C_Q C_F C_1 \epsilon < 1$ for $\epsilon < \epsilon_0$, we see that $u^{(m)}$ is a Cauchy sequence, and it converges to some $u \in H^{s+1}$. Then we have the estimates

$$\|u^{(n)} - u\|_{H^{s+1}} \leq C_n \epsilon^n$$

for some constant C_n depending on n. The estimates of $\|u\|_{H^{s+1}}$ follow from the triangle inequality and (4.4). □

We make a few remarks. (1) The same argument works for quasilinear wave equations. (2) The proof works for short time (i.e., for small T) instead of for small data. (3) Here we mainly consider F to be nonlinear. If F has linear terms, that is, $F_u(t,x,0) \neq 0$, then the same arguments work throughout. We just have to replace P by $\widetilde{P} = P + F_u(t,x,0)u$ and change Q to $\widetilde{Q} = \widetilde{P}^{-1}$. In case $F(t,x,0) \neq 0$, one can solve $Pw = -F(t,x,0)$ first and repeat the argument to get

$$\|u^{(n)} - (u-w)\|_{H^{s+1}(\mathscr{M}(T))} \leq C_n \epsilon^n.$$

So, in principle, one can remove the small data assumption. For simplicity, we stick with this assumption in the rest of the paper.

5. The basic network structure.

At least on the theoretical level, we can construct a simple neural network by comparing the iteration scheme for solving the wave equation and the MLP: we replace the linear mapping (affine transformation) in the MLP by the solution operator of the linear wave equation, and we replace the activation function by $F(t,x,u)$. See Figure 2.

This could be used for solving (1.1). Let us consider the continuous model for the moment. Let $h^{(-1)} = f$ be the source term of the wave equation. This is the input data for the network. We set

$$h^{(0)} = Q(f)$$

as the first layer, which is just the linearized solution. Let $F(t,x,z)$ be a smooth function which is the activation function now. We get

$$g^{(1)} = F(t,x,Qh^{(0)}).$$

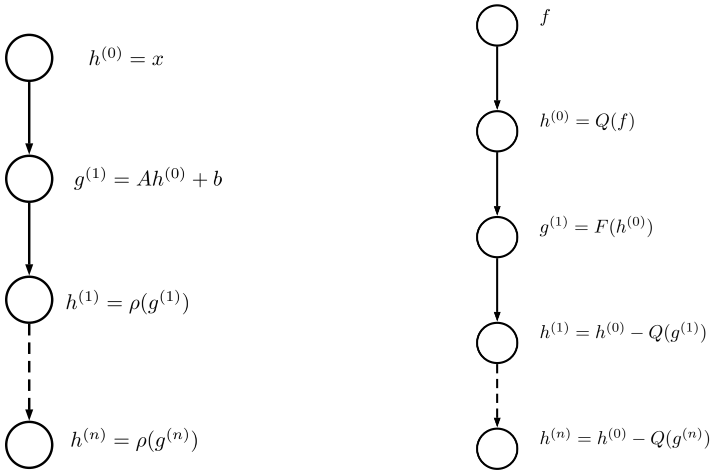
$h^{(0)} = x$
$g^{(1)} = Ah^{(0)} + b$
$h^{(1)} = \rho(g^{(1)})$
$h^{(n)} = \rho(g^{(n)})$
f
$h^{(0)} = Q(f)$
$g^{(1)} = F(h^{(0)})$
$h^{(1)} = h^{(0)} - Q(g^{(1)})$
$h^{(n)} = h^{(0)} - Q(g^{(n)})$

FIG. 2. *A schematic comparison of the feedforward network and the iteration scheme for solving wave equations.*

It is better that we think of this as the hidden layer. We apply the linear operation to get

$$h^{(1)} = h^{(0)} - QF(t, x, Qh^{(0)}). \tag{5.1}$$

This defines the iteration scheme and generates the second layer of the network. We then continue to obtain

$$h^{(n)} = h^{(0)} - QF(t, x, Qh^{(n-1)}), \quad n = 1, 2, \dots, n,$$

and the output is $h^{(N)}$. In this setting, we see that all the layers and parameters in the MLP have concrete meaning: the layers represent the propagation and nonlinear interactions in the solution, and the parameters represent the significance of the nonlinearities. The result in the previous section suggests that a deeper network produces better approximations.

Since we are interested in the inverse problem, we approach it as follows. First, the fundamental solution Q of P is related to the wave speed $c(x)$ that we aim to reconstruct. Thus we treat Q as the physical model and later use an oscillatory integral representation with explicit dependency on $c(x)$ to approximate Q. We will learn $c(x)$ with other network parameters, similarly to [26]. We remark that Q is a linear operator, and, roughly speaking, part of Q is a Fourier integral operator (FIO) and another part is a pseudodifferential operator. There are recent works on how to learn FIOs and pseudodifferential operators using neural networks; see [3] and [16]. It seems possible that this operation can be replaced by a network, although we will not use it here. Next, for the unknown nonlinear function $F(t, x, u)$, we will use a network which approximates F without an a priori model (such as polynomials). Finally, we use the iteration scheme as shown in Figure 2 to get the whole network.

Let us first use the MLP approximation for F to build the network. Let $h^{(-1)} = f$ be the source term of the wave equation, and set

$$h^{(0)} = Q(f)$$

as the first layer. For $h^{(0)}$, we introduce K units and apply affine transformations to get

$$q^{(1)} = \sum_{k=1}^{K} \gamma_k \rho(A_k(t, x, h^{(0)}) + B_k), \tag{5.2}$$

where $\rho(t, x, u)$ is an activation function to be specified below and A_k, B_k define an affine transformation on $(t, x, u) \in \mathbb{R}^{n+2}$. This step is supposed to approximate $F(t, x, u)$. Apply the linear operation to get

$$h^{(1)} = h^{(0)} - Q(q^{(1)}). \tag{5.3}$$

This completes the first step, and the iteration scheme is

$$\begin{aligned} h^{(n)} &= h^{(0)} - Q(q^{(n)}), \\ q^{(n)} &= \sum_{k=1}^{K} \gamma_k \rho(A_k(t, x, h^{(n-1)}) + B_k), \quad n = 1, 2, \dots, N. \end{aligned} \tag{5.4}$$

We thus obtain the output $h^{(N)}$. The parameter Θ consists of $\theta = \{\gamma_k, A_k, B_k, k = 1, 2, \dots, K\}$ and $c(x)$, but we treat $c(x)$ as the model parameter. We will not discuss further discretization of the parameters and networks as they do not matter for the approximation properties we address in this work. Using Theorems 3.1 and 3.2 and Proposition 4.1, we immediately obtain the following theorem.

THEOREM 5.1. *Consider the inverse problem for wave equations with sources. Assume*

1. $(f, u) \in \mathcal{D}_{sour,\epsilon}$ *for* $s > 1$ *and* ϵ *sufficiently small.*
2. $F(t, x, u)$ *is a smooth function with* $F(t, x, 0) = F_u(t, x, 0) = 0$.

Consider the network defined by iteration (5.3) *with activation function* $\rho : \mathbb{R}^3 \times \mathbb{R} \to \mathbb{R}$ *in* $C^l(\mathbb{R}^4)$ *with nonnegative integer* $l \leq s$. *Then there exist* $K > 0$ *and parameters* θ *such that*

$$\|u - h^{(N)}\|_{H^{l+1}(\mathcal{V})} \leq C_N \epsilon^{N+1}, \tag{5.5}$$

where C_N *is a constant independent of* f *and* u.

There are some issues with the network described by (5.4). First, it is important to realize that there are losses in the regularity of (5.5), and this is essential for understanding our construction. For example, the ReLU activation function $\rho(x) = x_+$ is C^0. So the network only approximates the solution u in the H^1 norm. Obviously, $\rho(x)$ introduces new singularities to the network. Although Q is a linear operator, it is nonlocal. Thus the new singularities might be propagated to other units. This issue does not show up in usual deep neural networks. In fact, the added singularity should help in solving image classification problems from the singularity point of view, but not for our problem.

Second, Theorems 3.1 and 3.2 do not provide any estimate on the number of units. In fact, to keep up with the ϵ^{N+1} error, the approximation error of $F(t, x, u)$ from the MLPs should be within ϵ^{N+1} instead of ϵ. Thus one is not making good use of the nonlinearity. Roughly speaking, we think of the "features" in this inverse problem as the H^s or C^r singularities of the solution (or, more precisely, the wave fronts in phase space). This is similar to the "edges" in images. An important phenomenon in nonlinear wave propagation is that nonlinear interactions of waves could produce new waves. This has been observed in physical applications and studied mathematically and is known as the nonlinear interaction of singularities and propagation of singularities for wave operators.

Our goal in sections 6–10 is to develop a neural network that addresses these issues. In section 6, we briefly discuss the nonlinear effects using conormal waves. In fact, data generated from such waves can be used as the training data. This consideration is used in section 7 to find an approximation of the linear operator Q and in section 8 for constructing approximations of the nonlinear term . Finally, we propose the network in section 9 and prove the main theorem in section 10.

6. Nonlinear interactions of conormal waves.

Conormal distributions have simple wave front sets and have been proven to be useful for analyzing nonlinear wave interactions. In general, the singularities generated from the nonlinear interactions could be rather complicated as shown by Beals's example; see [1]. In this section, we briefly discuss the results in [17, 21] to explain why it is helpful to look at the information in the phase space.

We use Lagrangian distributions from Hörmander [13]. Let $\mathscr{X}$ be an n dimensional smooth manifold and Λ be a smooth conic Lagrangian submanifold of $T^*\mathscr{X}\backslash 0$. We denote by $I^\mu(\mathscr{X}, \Lambda)$ the Lagrangian distribution of order μ associated with Λ. In particular, for $\mathscr{U}$ open in $\mathscr{X}$, let $\phi(x,\xi) : \mathscr{U} \times \mathbb{R}^N \to \mathbb{R}$ be a smooth nondegenerate phase function that locally parametrizes Λ, i.e.,

$$\{(x, d_x\phi) : x \in \mathscr{U}, d_\xi\phi = 0\} \subset \Lambda.$$

Then $u \in I^\mu(\mathscr{X}, \Lambda)$ can be locally written as a finite sum of oscillatory integrals

$$\int e^{i\phi(x,\xi)} a(x,\xi) d\xi, \quad a \in S^{\mu+\frac{n}{4}-\frac{N}{2}}(\mathscr{U} \times \mathbb{R}^N),$$

where $S^\bullet(\bullet)$ denotes the standard symbol class; see [13]. For $u \in I^\mu(\mathscr{X}, \Lambda)$, we know that the wave front set $\mathrm{WF}(u) \subset \Lambda$ and $u \in H^s(\mathscr{X})$ for any $s < -\mu - \frac{n}{4}$. The principal symbol of u is well defined in $S^{\mu+\frac{n}{4}}(\Lambda; \Omega^{\frac{1}{2}})/S^{\mu+\frac{n}{4}-1}(\Lambda; \Omega^{\frac{1}{2}})$, where $\Omega^{\frac{1}{2}}$ denotes the half-density bundle on Λ. See section 25.1 of [13]. For our problem, we can trivialize the bundle in local coordinates.

For a submanifold $\mathscr{Y} \subset \mathscr{X}$, the conormal bundle $N^*\mathscr{Y}$ is a Lagrangian submanifold of $T^*\mathscr{X}$. Distributions in $I^\mu(\mathscr{X}, N^*\mathscr{Y})$ are called conormal distributions to $\mathscr{Y}$. In local coordinates $x = (x', x''), x' \in \mathbb{R}^k, x'' \in \mathbb{R}^{n-k}$ such that $\mathscr{Y} = \{x' = 0\}$. Let $\xi = (\xi', \xi'')$ be the dual variable; then $N^*\mathscr{Y} = \{x' = 0, \xi'' = 0\}$. We can write $u \in I^\mu(\mathscr{X}, N^*\mathscr{Y})$ as

$$u = \int e^{ix'\xi'} a(x'', \xi') d\xi', \quad a \in S^{\mu+\frac{n}{4}-\frac{k}{2}}(\mathbb{R}^{n-k}_{x''}; \mathbb{R}^k_{\xi'}).$$

In this case, the principal symbol is

$$\sigma(u) = (2\pi)^{\frac{n}{4}-\frac{k}{2}} a_0(x'', \xi')|dx''|^{\frac{1}{2}}|d\xi'|^{\frac{1}{2}},$$

where $a_0 \in S^{\mu+\frac{n}{4}-\frac{k}{2}}(\mathbb{R}^{n-k}_{x''}; \mathbb{R}^k_{\xi'})$ is such that $a - a_0 \in S^{\mu+\frac{n}{4}-\frac{k}{2}-1}(\mathbb{R}^{n-k}_{x''}; \mathbb{R}^k_{\xi'})$. See [13].

Using four conormal waves and asymptotic analysis with multiple parameters, we can identify the leading terms in the solution that contains the new wave. This idea is introduced in Kurylev, Lassas, and Uhlmann [17] and further developed in Lassas, Uhlmann, and Wang [21]. We again consider the polynomial nonlinear function

$F(t,x,u) = a(t,x)u^2 + b(t,x)u^3 + c(t,x)u^4$. We refine the iteration method in section 4 by introducing four small parameters to locate the nonlinear interactions.

Let $f_i, i = 1,2,3,4$, be compactly supported, and set $f = \sum_{i=1}^4 \epsilon_i f_i$. We let $v_i = Q(f_i), i = 1,2,3,4$, be the linearized solution. Here, we shall assume that $v_i \in I^\mu(\mathscr{M}, N^*\mathcal{Y}_i)$, where $\mathcal{Y}_i$ are codimension one submanifolds of $\mathscr{M}$. These are called distorted plane waves; see [17, 21] for the details of construction in different contexts. With the source f, the linearized solution of u is $v = \sum_{i=1}^4 \epsilon_i v_i$. Let u be the solution of the nonlinear equation. We use the iteration scheme in section 4 to get

$$u = v + \sum \epsilon_i \epsilon_j Q(a v_i v_j) + \mathcal{R},$$

where the remainder term $\mathcal{R} = \sum_{i=1}^4 O(\epsilon_i^2)$ and the summation is over $i, j = 1,2,3,4$. In this approach, self-interactions of linearized waves are not considered. We iterate another two times to obtain

$$\begin{aligned} u &= v - Q(F(v - Q(F(v - Q(F((v - Q(F(u)))))))) \\ &= v + \sum_{i,j} Q(a v_i v_j) + \sum_{i,j,k} \epsilon_i\epsilon_j\epsilon_k [Q(b v_i v_j v_k) + 2Q(a v_i Q(a v_j v_k))] \\ &+ \epsilon_1\epsilon_2\epsilon_3\epsilon_4 \sum_{i,j,k,l} [Q(c v_i v_j v_k v_l) + Q(a v_i Q(b v_j v_k v_l)) + Q(b v_i v_j Q(a v_k v_l)) + Q(a v_i Q(a v_j Q(a v_k v_l)))] + \mathcal{R}. \end{aligned}$$

We observe that the $\epsilon_i\epsilon_j$ terms reflect the interaction of two waves v_i, v_j, and the $\epsilon_i\epsilon_j\epsilon_k$ terms reflect the interaction of three waves. The $\epsilon_1\epsilon_2\epsilon_3\epsilon_4$ terms are particularly useful. It is worth noting that these terms can be obtained from

$$\partial_{\epsilon_1}\partial_{\epsilon_2}\partial_{\epsilon_3}\partial_{\epsilon_4} u|_{\epsilon_1=\epsilon_2=\epsilon_3=\epsilon_4=0}. \tag{6.1}$$

Suppose $\mathcal{Y}_i, i = 1,2,3,4$, intersect at a point $q \in I(p_-, p_+)$ transversally. The work [17, 21] shows that (6.1) at q contains new singularities which are conormal to $T_q^*\mathscr{M}\backslash 0$. In other words, the term contains a point source. The singularity can be propagated back to the region $\mathscr{V}$ and hence is observable in the data. Moreover, the leading order terms of the symbol of the conormal distributions are determined, and they can be expressed in terms of the linear and nonlinear coefficients of the wave equation. The conclusion is that given all data $(f,u) \in \mathcal{D}_{sour,\epsilon}$, one can determine these coefficients in many cases up to diffeomorphisms; see [18, 21] for details. For illustration purposes, we formulate a simple version of the uniqueness result below.

THEOREM 6.1. *Let $c_1(x), c_2(x)$ be two smooth functions on $\mathbb{R}^3$, and let g_1, g_2 be an associated Lorentzian metric. Let $\mathscr{V}$ be a neighborhood of time like geodesics $\widehat{\mu}_i \subset \mathscr{M}$. Let $-1 < s_- < s_+ < 1$ and $p_i^\pm = \widehat{\mu}_i(s_\pm)$. Consider the nonlinear wave equation, $i = 1,2$,*

$$\begin{gathered} (\partial_t^2 - c_i^2(x)\Delta)u(t,x) + F_i(t,x,u(t,x)) = f(t,x), \quad (t,x) \in \mathscr{M}(T), \\ u(t,x) = 0, \quad (t,x) \in \mathscr{M}(0), \end{gathered}$$

where $F_i(t,x,u)$ are smooth such that $\partial_u^k F(t,x,0) \neq 0, x \in \mathscr{M}$, for some $k \geq 2$. Assume that for δ sufficiently small, the data set

$$\mathcal{D}^i_{sour,\epsilon} = \{(f, u|_{\mathscr{V}}) : f \in C_0^4(\mathscr{V}), \|f\|_{C^4} < \epsilon, u \textit{ is the solution of nonlinear wave equation}\},$$

$i = 1,2$, is the same. Then we have $c_1(x) = c_2(x)$ on $I(p_1^-, p_1^+) = I(p_2^-, p_2^+)$ and

$$\partial_u^k F_1(t,x,0) = \partial_u^k F_2(t,x,0), \quad k \geq 4.$$

Sketch of proof. Because $c_i(x)$ does not depend on t, we consider the linearized problem and apply Tataru's unique continuation theorem to conclude that $c_1 = c_2$ on $I(p_1^-, p_1^+) = I(p_2^-, p_2^+)$. The determination of F_i follows from [21, Theorem 1.3]. □

We remark that if c depends on t or more generally one works with a globally hyperbolic Lorentzian metric g, Tataru's unique continuation result does not apply. One needs the full analysis in [17, 21], and the determination is unique up to a conformal diffeomorphism in general. We also remark that the results are further applied to the Einstein equations coupled with scalar field equations or Maxwell equations; see [18, 22, 38].

7. The linear wave propagation.

We find the approximation of Q in this section. Consider the variable-coefficients linear wave equation

$$\begin{gathered} Pu = (\partial_t^2 - c^2(x)\Delta)u = f, \quad (t,x) \in (0,\infty)\times\mathbb{R}^3, \\ u = 0, \quad t < 0. \end{gathered}$$

The fundamental solution $Q = P^{-1}$ is well understood. For our purpose, we need the microlocal structure of Q. In Melrose and Uhlmann [30], a full symbolic construction was carried out, and the Schwartz kernel K_Q of Q was found to be a paired Lagrangian distribution. We recall that for two Lagrangians $\Lambda_0, \Lambda_1 \subset T^*\mathscr{X}$ which intersect cleanly at a codimension k submanifold, i.e.,

$$T_p\Lambda_0 \cap T_p\Lambda_1 = T_p(\Lambda_0 \cap \Lambda_1) \quad \forall p \in \Lambda_0 \cap \Lambda_1,$$

the paired Lagrangian distribution associated with (Λ_0, Λ_1) is denoted by $I^{p,l}(\Lambda_0, \Lambda_1)$. For $u \in I^{p,l}(\Lambda_0, \Lambda_1)$, we know that $\mathrm{WF}(u) \subset \Lambda_0 \cup \Lambda_1$. Microlocally away from the intersection $\Lambda_0 \cap \Lambda_1$, $u \in I^{p+l}(\Lambda_0 \backslash \Lambda_1)$ and $u \in I^{p}(\Lambda_1 \backslash \Lambda_0)$ are Lagrangian distributions on the corresponding Lagrangians.

Let (t, x, τ, ξ) be local coordinates of $T^*\mathscr{M}$, and let $\mathcal{P}(t,x,\tau,\xi) = |\tau|^2 - c^2(x)|\xi|^2$ be the principal symbol of P. Let Σ be the characteristic set

$$\Sigma = \{(t,x,\tau,\xi) \in T^*\mathscr{M} : \mathcal{P}(t,x,\tau,\xi) = 0\}.$$

The Hamilton vector field of $\mathcal{P}$ is denoted by $H_{\mathcal{P}}$, and in local coordinates

$$H_{\mathcal{P}} = \sum_{i=1}^{4}\left(\frac{\partial\mathcal{P}}{\partial\zeta_i}\frac{\partial}{\partial z_i} - \frac{\partial\mathcal{P}}{\partial z_i}\frac{\partial}{\partial\zeta_i}\right), \quad z = (t, x_1, x_2, x_3),\, \zeta = (\tau, \xi_1, \xi_2, \xi_3).$$

The integral curves of $H_{\mathcal{P}}$ in Σ are called null-bicharacteristics. Let $\mathrm{Diag} = \{(z,z') \in \mathscr{M}\times\mathscr{M} : z = z'\}$ be the diagonal, and denote by

$$N^*\mathrm{Diag} = \{(z,\zeta,z',\zeta') \in T^*(\mathscr{M}\times\mathscr{M})\backslash 0 : z = z', \zeta' = -\zeta\}$$

the conormal bundle of Diag minus the zero section. We let Λ_c be the Lagrangian submanifold in $T^*(\mathscr{M}\times\mathscr{M})$ obtained by flowing out $N^*\mathrm{Diag}\cap\Sigma$ under $H_{\mathcal{P}}$. Here, we regard $\Sigma, H_{\mathcal{P}}$ as objects on product manifold $T^*\mathscr{M}\times T^*\mathscr{M}$ by lifting from the left factor. More explicitly,

$$\Lambda_c = \{(z,\zeta,z',\zeta') \in T^*(\mathscr{M}\times\mathscr{M})\backslash 0 :\ (z,\zeta) \text{ lies on a bicharacteristic from } (z', -\zeta')\}.$$

The canonical relation is denoted by

$$\Lambda_c' = \{(z,\zeta,z',\zeta') \in T^*(\mathscr{M})\backslash 0 \times T^*(\mathscr{M})\backslash 0 : (z,\zeta,z',-\zeta') \in \Lambda_c\}.$$

We also call the map $S(z', \zeta') = (z, \zeta)$ if $(z, \zeta, z', \zeta') \in \Lambda'_c$ the canonical relation. This map can be found explicitly by solving the Hamilton field equations. Let $\gamma(s) = (\alpha(s), \beta(s)) : [0, \infty) \to \mathscr{M} \times \mathbb{R}^4$ be the null-bicharacteristics from (z', ζ'). Then we have

$$\begin{gathered} \frac{d\alpha(s)}{ds} = \frac{\partial \mathcal{P}}{\partial \zeta}, \quad \frac{d\beta(s)}{ds} = -\frac{\partial \mathcal{P}}{\partial z}, \\ \alpha(0) = z', \quad \beta(0) = \zeta'. \end{gathered} \tag{7.1}$$

Then $S(z', \zeta') = \gamma(s_0)$ where $\alpha(s_0) = z$. It is shown in [30] that for linear differential operator P, the causal inverse $Q \in I^{-\frac{3}{2}, -\frac{1}{2}}(N^*\text{Diag}, \Lambda_c)$ is such that $PQ = \text{Id}$ on $\mathscr{E}'(\mathscr{M})$. Also, from [7, Proposition 5.6], we know that $Q : H^s_{\text{comp}}(\mathscr{M}) \to H^{s+1}_{\text{loc}}(\mathscr{M})$ is continuous for $s \in \mathbb{R}$.

We will find and use the leading order term in Q as its approximation. We follow the parametrix construction in [30, Proposition 6.6]; see also [9, section 5.1]. The conditions (6.1)–(6.6) of [30] are satisfied; thus the flow out of $\partial\Lambda_c$ under $H_{\mathcal{P}}$ is an embedded Lagrangian submanifold with boundary. We look for $Q_0 \in I^{-\frac{3}{2}, -\frac{1}{2}}(N^*\text{Diag}, \Lambda_c)$ to solve $PQ_0 - \text{Id} = 0$ with errors of lower orders. First we have

$$\sigma(\text{Id}) = \mathcal{P}(z, \zeta)\sigma(Q_0)|_{N^*\text{Diag}}.$$

So on $N^*\text{Diag}$, we have $\sigma(Q_0) = \mathcal{P}(z, \zeta)^{-1}$. Then from [30, Theorem 4.13], we obtain the (nonzero) initial condition of $\sigma(Q_0)$ on $\Lambda_c \cap N^*\text{Diag}$. We solve on Λ_c

$$(i\mathscr{L}_{H_{\mathcal{P}}} + \mathcal{P}_{sub})\sigma(Q_0) = 0, \tag{7.2}$$

where $\mathscr{L}$ denotes the Lie derivative acting on half-density factors and $\mathcal{P}_{sub} = -\frac{1}{2i}\sum_{j=0}^{3} \frac{\partial^2 \mathcal{P}}{\partial z_j \partial \zeta_j}$ is the subprincipal symbol. Along the null-bicharacteristics from (z', ζ') to (z, ζ), the equation is a transport equation, and we get the solution $\sigma(Q_0)(z, \zeta, z', \zeta')$, which is nonvanishing. Using the canonical relation, we can write it as

$$\sigma(Q_0)(S(z', \zeta'); z', \zeta').$$

So we find $Q_0 \in I^{-\frac{3}{2}, -\frac{1}{2}}(N^*\text{Diag}, \Lambda_c)$ such that

$$Q - Q_0 \in I^{-\frac{5}{2}, -\frac{1}{2}}(N^*\text{Diag}, \Lambda_c).$$

Using the L^2 estimates of FIOs with paired Lagrangian kernel (see [14, Theorem 3.3]), we obtain $Q - Q_0 : H^s_{comp}(\mathscr{M}) \to H^{s+2}_{loc}(\mathscr{M})$.

8. Estimates of nonlinear effects.

Suppose $u \in H^{s+1}(\mathscr{M})$ and $F(t, x, u)$ is smooth in t, x, u. Because in the linear wave propagation we are only concerned with the leading order singularities in u, we actually have $F(u) = F(w + \mathcal{R})$, where $u = w + \mathcal{R}$ and $\mathcal{R} \in H^{s+2}$. In this section, we show that $F(u)$ can be approximated by a function $\widetilde{F}(w)$ with difference $F(u) - \widetilde{F}(w) \in H^{s+2}$ and such that $\widetilde{F}(w)$ captures the major nonlinear effects. Actually, we shall work in the phase space and make use of Bony's paraproducts. The approximation function $\widetilde{F}$ is eventually constructed by a convolutional neural network.

Instead of conormal distributions, for which we looked at the singularities in conical decomposition of the phase space, we use Sobolev and Hölder functions and start with the dyadic decomposition of Coifman and Meyer [4]. For $K > 1$ fixed, we set

$$\mathscr{C}_p = \{\xi \in \mathbb{R}^n : K^{-1}2^p \le |\xi| \le K2^{p+1}\}$$

and $\mathscr{C}_0 = \{\xi \in \mathbb{R}^n : |\xi| \le 2K\}$. Then $\{\mathscr{C}_p\}_0^\infty$ form an open covering of $\mathbb{R}^n$. Let ψ_j be a partition of unity:

$$1 = \sum_{j=0}^\infty \psi_j(\xi), \quad \psi_j \in C_0^\infty, \quad \text{supp } \psi_j \subset \mathscr{C}_j.$$

Actually, one can begin with $\psi_0(\xi)$ which is equal to 1 for $|\xi| \le K$ and 0 for $|\xi| > 2K$. Then set $\Psi_j(\xi) = \psi_0(2^{-j}\xi)$, and set $\psi_j(\xi) = \Psi_j(\xi) - \Psi_{j-1}(\xi)$. For any $u \in \mathscr{S}'(\mathbb{R}^n)$, the Paley–Littlewood decomposition of u is

$$\{u_p\}_0^\infty, \text{ where } u_p = \mathscr{F}^{-1}(\psi_p(\xi)\widehat{u}(\xi)), p = 0, 1, 2, \ldots.$$

Hereafter, $\mathscr{F}, \mathscr{F}^{-1}$ are used to denote Fourier and inverse Fourier transforms. The hat notation is also used for the Fourier transform when convenient. We have $u = \sum_{p=0}^\infty u_p$ in the topology of $\mathscr{S}'(\mathbb{R}^n)$; see, e.g., [36]. We recall that Sobolev and Hölder functions can be characterized using Paley–Littlewood decompositions. The Sobolev space $H^s(\mathbb{R}^n)$ is defined as

$$H^s(\mathbb{R}^n) = \{u \in \mathscr{S}'(\mathbb{R}^n) : (1 + |\xi|^2)^{s/2}\widehat{u}(\xi) \in L^2(\mathbb{R}^n)\}$$

with norm

$$\|u\|_{H^s} = \|(1 + |\xi|^2)^{s/2}\widehat{u}(\xi)\|_{L^2}.$$

Then $u \in H^s(\mathbb{R}^n)$ if and only if $u = \sum_{p=0}^\infty u_p$, where $\widehat{u}_p$ are supported in $\mathscr{C}_p$ and

$$\|u_p\|_{L^2} \le c_p 2^{-ps}, \quad \{c_p\} \in l^2.$$

Consider the Hölder space $C^\alpha(\mathbb{R}^n), \alpha > 0$ noninteger, equipped with the norm

$$\|u\|_{C^\alpha} = \sum_{|\lambda| \le [\alpha]} \|\partial^\lambda u\|_{C^\beta}.$$

When α are integers, it is necessary to use the Zygmund space $C_*^\alpha(\mathbb{R}^n)$. In particular, $C_*^\alpha = C^\alpha$ if α is not integer. Otherwise, $C^\alpha \subset C_*^\alpha$. The characterization is that $u \in C_*^\alpha(\mathbb{R}^n)$ if and only if $u = \sum_{p=0}^\infty u_p$, where $\widehat{u}_p$ are supported in $\mathscr{C}_p$ and

$$\|u_p\|_{L^2} \le c 2^{-p\alpha}.$$

Let $a \in C^r(\mathbb{R}^n)$ and $f \in H^s(\mathbb{R}^n)$. The paraproduct of a and f, introduced by Bony [2], is

$$T_a f = \sum_{k \ge 1} (\Psi_{k-1}(D)a)(\psi_{k+1}(D)f) = \sum_{p=2}^\infty \sum_{q=0}^{p-2} a_q f_p, \tag{8.1}$$

where $\Psi_k(\xi) = \sum_{j=0}^k \psi_j(\xi)$. If we denote $\mathscr{B}_p = \{\xi \in \mathbb{R}^n : |\xi| \le K2^{p+1}\}$, then Ψ_k is supported in $\mathscr{B}_k$. Using the characterization of H^s, C^r functions, we see that $T_a f \in H^s(\mathbb{R}^n)$ and

$$af = T_a f + \mathcal{R}, \quad \mathcal{R} \in H^{s+r}(\mathbb{R}^n).$$

So the difference is a more regular term for $r > 0$. Furthermore, we have that if $u \in C^r(\mathbb{R}^n) \cap H^s(\mathbb{R}^n), r, s > 0$, and $F(u)$ is smooth in u, then

$$F(u) = T_{F'(u)}u + \mathcal{R}, \quad \mathcal{R} \in H^{s+r}(\mathbb{R}^n).$$

See [36, Proposition 3.2.C]. We remark that the paraproduct does not throw away all nonlinear effects, which is evident from the definition (8.1).

There are several equivalent variants of paraproducts; see [36]. The one convenient for our purpose is to introduce a convolution kernel in the phase space, which is also done in Bony [2]. Choose $\chi \in C^\infty(\mathbb{R}^n \times \mathbb{R}^n)$ homogeneous of degree 0 outside a compact set such that $\chi(\xi,\eta) = 0$ for $|\xi| > \frac{1}{2}|\eta|$ and $\chi(\xi,\eta) = 1$ for $|\xi| < |\eta|/16$ and $|\eta| > 2$. Then the paraproduct can be written as

$$\begin{aligned} T_a^\chi f(x) &= (2\pi)^{-n} \int e^{ix\xi}\chi(\xi-\eta,\eta)\widehat{a}(\xi-\eta)\widehat{f}(\eta)d\eta d\xi \\ &= (2\pi)^{-n} \int e^{ix(\xi+\eta)}\chi(\xi,\eta)\widehat{a}(\xi)\widehat{f}(\eta)d\xi d\eta. \end{aligned}$$

We see that $\widehat{T_a^\chi f}$ is a convolution of $\widehat{a}, \widehat{f}$ with kernel χ. Here, we emphasize the dependence on χ. We also use the notation $T_a^\chi f = T^\chi(a; f)$.

We use paraproducts to construct a network for approximating composite functions $F(u), u \in H^s(\mathbb{R}^n)$. (Here, F is only a function of u and not of x.) Let $h^{(0)} = u$ be the first level of the network. We perform affine transformations and use paraproducts as the activation function to get

$$h^{(1)} = T^\chi(h^{(0)}; a_1 h^{(0)} + b_1), \tag{8.2}$$

where a_1, b_1 are constants. Then we continue to get

$$h^{(n)} = T^\chi(h^{(0)}; a_n h^{(n-1)} + b_n), \quad n = 1, 2, \ldots, N. \tag{8.3}$$

We remark that by taking the Fourier transform, the network is a convolutional network with kernel χ and $\widehat{h}^{(n)}$ and the network provides an approximation of $\widehat{F}(u)$ in the phase space. These two points of view will be used interchangeably below.

Our goal is to analyze the difference $F(u) - h^{(n)}(u)$ and show the approximation property of the network. We prove that the error terms consist of a spatial error which is controlled by the nonlinearity of F and a phase space error term which is controlled by the regularity of u. For $R > 0$, let $\Psi_R(\xi), \xi \in \mathbb{R}^n$, be a smooth cut-off function such that $\Psi_R(\xi) = 0$ if $|\xi| < R$ and $\Psi_R(\xi) = 1$ if $|\xi| > 2R$. We denote by $\Psi_R(D)$ the pseudodifferential operator with symbol Ψ_R.

PROPOSITION 8.1. *We assume that*

1. $u \in H^s(\mathbb{R}^n) \cap C^r(\mathbb{R}^n), r, s > 0$ *with* $\|u\|_\infty < \epsilon$;
2. $F(u)$ *is a smooth function of* u.

Then there exist constant parameters $a_i, b_i, i = 1, \ldots, N$, *such that for the* $h^{(N)}(u)$ *obtained in* (8.3), *we have* $F(u) - h^{(N)}(u) = \mathcal{R}_{sp} + \mathcal{R}_{ph}$, *where* $\mathcal{R}_{sp} \in H^s(\mathbb{R}^n), \mathcal{R}_{ph} \in H^{s+r}(\mathbb{R}^n)$, *and*

$$\|\mathcal{R}_{sp}\|_{H^s} < C_F \epsilon^{N+1}, \quad \|\Psi_R(D)\mathcal{R}_{ph}\|_{H^s} = O(R^{-r}).$$

Here C_F *is a constant such that* $\sup_{\|u\|_\infty < \epsilon} |\partial_u^{N+1} F(u)| \leq C_F$.

Proof. The proof is straightforward. First we use Taylor expansion of F based at 0:

$$F(u) = \sum_{n=0}^{N} \frac{\partial_u^n F(0)}{n!} u^n + \mathcal{R}_{sp}, \quad |\mathcal{R}_{sp}| \leq C_F |u|^{N+1}.$$

Next, we let $p(u) = \sum_{n=0}^{N} \frac{\partial_u^n F(0)}{n!} u^n$ and rewrite it as

$$p(u) = a_n u(\cdots a_3 u(a_2 u(a_1 u + b_1) + b_2) + b_3 \cdots) + b_n,$$

where a_i, b_i are constants related to $\frac{\partial^n F(0)}{n!}$. We now replace the products by paraproducts. First,

$$a_2 u(a_1 u + b_1) + b_2 = a_2 T^\chi(u; a_1 u + b_1) + b_2 + \mathcal{R}_2,$$

where $\mathcal{R}_2 \in H^{s+r}(\mathbb{R}^n)$. Next, we get

$$a_3(a_2 u(a_1 u + b_1) + b_2) + b_3 = a_3 T^\chi(u; a_2 T^\chi(u; a_1 u + b_1) + b_2) + a_3 T^\chi(u; \mathcal{R}_2) + \mathcal{R}_3,$$

where $\mathcal{R}_3 \in H^{s+r}(\mathbb{R}^n)$ and we also have $a_3 T^\chi(u; \mathcal{R}_2) \in H^{s+2r}(\mathbb{R}^n)$. Continuing this procedure, we get the function $h^{(N)}$ such that $p(u) - h^{(N)}(u) = \mathcal{R}_{ph} \in H^{s+r}(\mathbb{R}^n)$. This finishes the proof. □

Next, we show the stability of the network with respect to regular perturbations.

COROLLARY 8.2. *We assume that*

1. $u, w \in H^s(\mathbb{R}^n) \cap C^r(\mathbb{R}^n), r, s > 0$, *with* $\|u\|_\infty, \|w\|_\infty < \epsilon$;
2. $u = w + \mathcal{R}$ *with* $\mathcal{R} \in H^{s+m}(\mathbb{R}^n), m > 0$;
3. $F(u)$ *is a smooth function.*

Then there exist constant parameters $a_i, b_i, i = 1, \ldots, N$, *such that for the* $h^{(N)}(w)$ *obtained in* (8.3), *we have* $F(u) - h^{(N)}(w) = \mathcal{R}_{sp} + \mathcal{R}_{ph}$, *where* $\mathcal{R}_{sp} \in H^s(\mathbb{R}^n), \mathcal{R}_{ph} \in H^{s+r}(\mathbb{R}^n)$, *and*

$$\|\mathcal{R}_{sp}\|_{H^s} < C_F \epsilon^{N+1}, \quad \|\Psi_R(D)\mathcal{R}_{ph}\|_{H^s} = O(R^{-t}),$$

where $t = \min(m, r)$ *and* C_F *is the same as in Proposition* 8.1.

Proof. The spatial error is the same as in the previous proposition. So we consider

$$p(u) = \sum_{n=0}^{N} \frac{\partial_u^n F(0)}{n!}(w + \mathcal{R})^n = \sum_{n=0}^{N} \frac{\partial_u^n F(0)}{n!} w^n + \mathcal{R}_0,$$

where $\mathcal{R}_0 \in H^{s+m}(\mathbb{R}^n)$ because it is a finite sum of products of $w \in C^r(\mathbb{R}^n)$ and $\mathcal{R} \in H^m(\mathbb{R}^n)$. This finishes the proof. □

We remark that if $F(u)$ is such that $F(0) = F'(0) = 0$, we see from the proof that the estimate of $\mathcal{R}_{ph}$ is actually $\|\Psi_R(D)\mathcal{R}_{ph}\|_{H^s} = O(\epsilon^2 R^{-t})$. Finally, we make a remark in relation to the usual convolutional networks; see, e.g., [23, 28]. Let $u, v \in H^s(\mathbb{R}^n) \cap C^r(\mathbb{R}^n)$, and we consider the paraproduct of u, v:

$$\widehat{T_v^\chi u}(x) = \int \chi(\xi - \eta, \eta)\widehat{u}(\xi - \eta)\widehat{u}(\eta) d\eta.$$

Now we use the Paley–Littlewood decomposition and consider for M large

$$u \simeq \sum_{p=0}^{M} u_p = \sum_{p=0}^{M} \mathcal{F}^{-1}(\psi_p(\xi)\widehat{u}(\xi)).$$

Then formally we obtain

$$\widehat{T_v^\chi u}(x) \simeq \int \sum_{p=0}^{M} \chi(\xi - \eta, \eta)\psi_p(\xi - \eta)\widehat{v}(\xi - \eta)\widehat{u}(\eta) d\eta.$$

One can think of $\chi^p(\xi,\eta) = \chi(\xi,\eta)\psi_p(\xi)$ as the convolutional kernel at different scales. Here, $\widehat{v}(\xi-\eta)$ plays an important role because it captures the nonlinear effects. One could approximate this $\widehat{v}(\xi-\eta)$ on the support of χ^p using a set of parameters, and one would obtain the usual convolutional network.

9. Convolutional neural network in phase space.

We construct the network for approximating the map $f \to L(f)|_{\mathscr{V}}$ described in section 2. Here, $\mathscr{V}$ is an open relatively compact set of $\mathscr{M}$, and f is the source function supported in $\mathscr{V}$. This is the input data for the network.

We first choose a finite open covering $\mathscr{U}_i, i = 1,\dots,K$, of a compact region $\mathscr{K} \subset \mathscr{M}$ that contains $I(p_-,p_+)$, where $p_\pm \in \mathscr{V}$. Then for $c(x)$ close to some fixed $c_0(x)$, the construction does not depend on $I(p_-,p_+)$ which depends on c. Let $\operatorname{diam}(\mathscr{U})$ be the diameter of set $\mathscr{U} \subset \mathscr{M}$, and we assume that $\operatorname{diam}(\mathscr{U}_i) < \delta, i = 1,2,\dots,K$. We see that K is at least $O(\delta^{-2})$. Let ϕ_i be a partition of unity subordinated to $\mathscr{U}_i$:

$$\phi_i \in C_0^\infty(\mathscr{U}_i), \quad \sum_{i=1}^K \phi_i = 1.$$

For any $f \in H^s(\mathscr{M})$, we write $f_i = \phi_i f \in H^s(\mathscr{M})$ and get $f = \sum_{i=1}^K f_i$. Then we define the 0th layer (input) of the network to be

$$h^{(0)} = \{\widehat{f_i}\}_{i=1}^K.$$

In particular, the number of units for this level is K. When $\mathscr{U}_i$ is taken sufficiently small, $\widehat{f_i}$ is a good approximation of the wave front set of f at $\mathscr{U}_i$.

Next, we solve the wave equation approximately from each $\mathscr{U}_i$ to $\mathscr{U}_j$. This means that we solve

$$Pu = f_i \text{ in } \mathscr{M}$$

and get $u|_{\mathscr{U}_j}$. This makes sense if $\mathscr{U}_i \cap J_+(\mathscr{U}_j) \neq \emptyset$. We remark that one can regard u as the wave-packet generated by a point source if $\mathscr{U}_i$ is sufficiently small. We shall use the leading term of Q on Λ_c with principal symbol $\sigma(Q_0)$. For each pair $\mathscr{U}_i, \mathscr{U}_j$, we further decompose the operation as follows. Let S_{ij} be an invertible matrix which is an approximation of the canonical relation S. Then we set

$$h_{ij}^{(1)}(\zeta) = c_{ij}|\zeta|^{-1} h_i^{(0)}(S_{ij}\zeta), \quad \zeta \in \mathbb{R}^4, \quad i \neq j, \tag{9.1}$$

which we think of as an approximation of $\widehat{\phi_i u}$ (to be justified later). If $\mathscr{U}_i \cap J_+(\mathscr{U}_j) = \emptyset$, we should take $c_{ij} = 0$. This step solves wave propagation. On $\mathscr{U}_j$ itself, we solve using Q_0 on $N^*\text{Diag}$, so

$$h_{jj}^{(1)}(\zeta) = \frac{\gamma(\zeta) h_j^{(0)}(\zeta)}{|\tau|^2 - c_{jj}^2|\xi|^2}, \quad \zeta = (\tau,\xi) \in \mathbb{R}^4. \tag{9.2}$$

One can think of c_{jj} as the constant wave speed on $\mathscr{U}_j$, and $\gamma(\zeta)$ is a cut-off function away from the light-like directions. In particular, let $\mathcal{P}_j(\zeta) = |\tau|^2 - c_{jj}^2|\xi|^2$. Then we take $\gamma(\zeta) = 0$ when $\mathcal{P}_j(\zeta) < \delta$ and $\gamma(\zeta) = 1$ if $\mathcal{P}_j(\zeta) > 2\delta$. We collect the effects on each $\mathscr{U}_i$ and let

$$h^{(1)} = \{h_i^{(1)}\}_{i=1}^K, \quad h_i^{(1)} = \sum_{j=1}^K h_{ij}^{(1)}.$$

We shall take this as the first layer of the network. See Figure 3. We note that this step is based on the wave equation model. We also remark that if Q were an FIO, then one would only have (9.1) and not (9.2).

To obtain the next layer, we need to take into account the nonlinear effects and find approximations of $\widehat{F}$. Now we use the convolutional network constructed in section 8 on each $\mathscr{U}_i$. Here, on each $\mathscr{U}_i$, we take $F(t_i, x_i, u)$ for some $(t_i, x_i) \in \mathscr{U}_i$ as the approximation of $F(t, x, u)$ and apply the network (8.2), (8.3). The parameters are $\theta_{ik} \doteq \{a_{ik}, b_{ik}\}, i = 1, 2, \dots, K; k = 1, 2, \dots, M$. We denote the obtained approximation function on $\mathscr{U}_i$ by $v_i^{(1)}, i = 1, 2, \dots, K$. Next, we take $v_j^{(1)}$ as the source to solve the wave equation in phase space as before to get $h_{ij}^{(2)}$ on $\mathscr{U}_i$. Again, the terms are no longer supported on $\mathscr{U}_i$ so we collect the terms on each $\mathscr{U}_i$ to obtain the second layer

$$h_i^{(2)} = h_i^{(1)} - \sum_{j=1}^{M} h_{ij}^{(2)}.$$

This layer collects the linear and quadratic effects in the solution. See Figure 3 for the illustration of the structure.

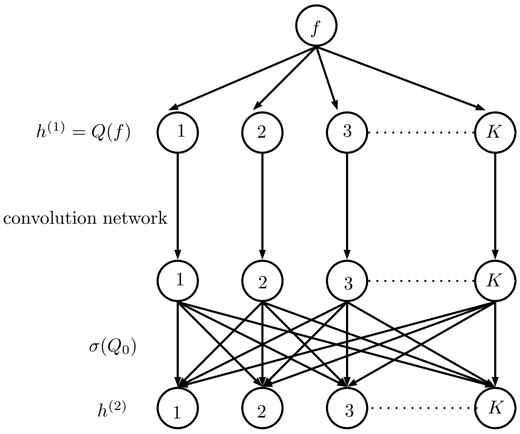

FIG. 3. *Illustration of the network structure. $h^{(1)}$ is the first layer obtained from the input f. $h^{(2)}$ is the second layer of the network obtained from the convolution network and application of Q. This defines the iteration scheme.*

We continue the procedure. We construct $h_{ij}^{(n)}$ from $h_i^{(n-1)}$ using (9.1), (9.2). Then we use the network to get $h_i^{(n)}$ from $h_{ij}^{(n)}$. The parameter of the network is

$$\Theta = \{c_{ij}, S_{ij} : i, j = 1, \dots, K\} \cup \{\theta_{ik} : i = 1, \dots, K, k = 1, \dots, M\},$$

where we take $S_{ii} = \mathrm{Id}$ as the identity. The first set on the right-hand side is the model parameters, and the second set is the parameters of the neural network. We denote the final output by $h^{(M)}(f; \Theta)$.

To determine the parameters, we need to solve the optimization problem on training data with the cost function

$$J_i = \|\widehat{u}_i - h_i^{(M)}(f; \Theta)\|^2, \quad i \in \mathscr{I},$$

where the index set $\mathscr{I}$ is such that $\mathscr{U}_i \subset \mathscr{V}, i \in \mathscr{I}$, and $u_i = \phi_i u$ is supported on $\mathscr{U}_i$. We shall specify the proper norm after the analysis in next section. As we pointed out earlier, the distorted plane waves discussed in section 6 can be used as training data.

10. The approximation theorem.

Theorem 10.1. *Consider the inverse problem for nonlinear wave equations with sources formulated in section* 2. *Assume that*

1. $c(x), F(t,x,u)$ *are smooth functions;* $F(t,x,0) = F_u(t,x,0) = 0$*;*
2. $(f,u) \in \mathcal{D}_{sour,\epsilon}$ *with* $\epsilon < \epsilon_0, s > 1$*, where* ϵ_0 *is as in Proposition* 4.1.

Consider the convolutional network constructed in section 9 *with depth* $M \geq 0$ *and* $K \geq 1$ *units for each level. Let* $\mathscr{I}$ *be the index set so that* $\mathscr{U}_i \subset \mathscr{V}$*, and let* $u_i = \phi_i u$*. Assume that* diam$(\mathscr{U}_i) < \delta$ *for some* $\delta > 0$*. Then there exist parameter sets* Θ *and* M, K *such that the function* $h^{(M)}(f;\Theta)$ *generated by the neural network satisfies for* $i \in \mathscr{I}$

$$\|\Psi_R(\zeta)\langle\zeta\rangle^{(s+1)/2}(\widehat{u}_i(\zeta) - h_i^{(M)}(f;\Theta))\|_{L^2(\mathbb{R}^4)} \leq C(1+\delta)\epsilon^M,$$

where $R > R_0$ *and the constants* C*,* R_0*, and* K *depend on* $M, \delta, \epsilon_0, c(x)$ *and* $F(t,x,u)$*.*

We make a few remarks before giving the proof. This theorem indicates that it is better to solve the optimization problem in the phase space and consider high frequency information. We shall see in the proof that the error comes from two sources. One is $\|u - u^{(n)}\|_{H^{s+1}} \leq C_n\epsilon^n$ from Proposition 4.1, where C_n is found in the proof. The other one is

$$\|\Psi_R(\zeta)\langle\zeta\rangle^{(s+1)/2}(\widehat{u}_i^{(n)}(\zeta) - h_i^{(M)}(f;\Theta))\|_{L^2(\mathbb{R}^4)} \leq C\delta\epsilon^n,$$

and we shall see that C depends on ϵ_0, M, and

$$\sup_{x\in I(p_-,p_+)} |c(x)|, \quad \sup_{x\in I(p_-,p_+)} |\partial_x c(x)|, \quad \sup_{(t,x)\in I(p_-,p_+),|s|<1} |\partial_s^k \partial_{(t,x)} F(t,x,s)|, \quad k \leq M.$$

Here, $x \in I(p_-,p_+)$ means x in the projection of $I(p_-,p_+)$ to $\mathbb{R}^3$. By taking δ small (necessarily increasing the number of units K), we obtain better approximation results. Indeed, when $\delta \to 0$, the set $\mathscr{U}_i$ approaches a point. Essentially what matters in the units of the network is just the wave front sets of u so the estimates become more accurate. Finally, we remark that in view of the uniqueness result Theorem 6.1 and its proof, one can take the training data consisting of sufficiently many conormal waves that are supported on each $\mathscr{U}_i, i \in \mathscr{I}$.

Proof of Theorem 10.1. Because $s > 1$, we know from Proposition 4.1 and Sobolev embedding that the solution $u \in H^{s+1} \subset C^r, r < s-1$. We start with the first level, which is $h^{(1)}$. This involves solving the wave equation using Q_0. Recall that the open coverings $\mathscr{U}_i, i = 1, 2, \ldots, K$, for $I(p_-,p_+)$ and $f_i = \phi_i f \in H^s$ are compactly supported on $\mathscr{U}_i$.

We first solve $Pv = f_i$ away from $\mathscr{U}_i$. From section 7, we know that $v - Q_0(f_i) \in H^{s+2}$. Away from $\mathscr{U}_i$, it suffices to consider $Q_0 \in I^{-\frac{3}{2}}(\Lambda_c \backslash N^*\text{Diag})$. So we can write

$$Q_0(f_i)(z) = \int e^{i\phi(z,z',\theta)}a(z,z',\theta)f_i(z')dz'd\theta,$$

where ϕ is a homogeneous nondegenerate phase function that parametrizes the Lagrangian Λ_c locally near (z,z'); namely,

$$\Lambda_c = \{(z,\zeta,z',\zeta') \in T^*(\mathscr{M}\times\mathscr{M})\backslash 0 : \zeta = \phi_z, \zeta' = -\phi_{z'}, \phi_\theta = 0\},$$

and a is a smooth function homogeneous of degree -1 in θ for $|\theta| > 1$. Consider $z \in \mathscr{U}_j, z' \in \mathscr{U}_i, i \neq j$, and choose constant $\widetilde{c}_{ij}$ such that

$$|\widetilde{c}_{ij}\langle\theta\rangle^{-1} - a(z, z', \theta)| \leq C\delta\langle\theta\rangle^{-1}.$$

Here, C depends on the symbol a. Because the symbol is obtained by solving the transport equation (7.2) involving $c(x)$ and its first derivatives along null-bicharacteristics, from the stability of ODEs, we see that C depends on $||c||_{C^1(I(p_-,p_+))}$. Note that here $c(x)$ is smooth, and only the constant depends on the C^1 seminorm. Then we have

$$\phi_i(z)Q_0(f_i)(z) - \phi_i(z)\int e^{i\phi(z,z',\theta)}\widetilde{c}_{ij}\langle\theta\rangle^{-1}f_i(z')dz'd\theta = \mathcal{R}_1,$$
$$\text{where } \|\mathcal{R}_1\|_{H^{s+1}(\mathscr{U}_i)} \leq C\delta\|f\|_{H^s(\mathscr{V})} \leq C\delta\epsilon.$$

Now we denote

$$I(z) = \phi_i(z)\int e^{i\phi(z,z',\theta)}\widetilde{c}_{ij}\langle\theta\rangle^{-1}f_i(z')dz'd\theta$$

and take the Fourier transform to get

$$\begin{aligned}\widehat{I}(\zeta) &\doteq \int e^{-iz\zeta}e^{i\phi(z,z',\theta)}\phi_i(z)\widetilde{c}_{ij}\langle\theta\rangle^{-1}f_j(z')dz'd\theta dz\\ &= (2\pi)^{-4}\int e^{-iz\zeta}e^{i\phi(z,z',\theta)}e^{iz'\eta}\phi_i(z)\widetilde{c}_{ij}\langle\theta\rangle^{-1}\widehat{f}_j(\eta)dz'd\theta dz d\eta.\end{aligned}$$

For this oscillatory integral, the phase function is

$$\Phi(z, z', \zeta, \theta, \eta) = -z\zeta + \phi(z, z', \theta) + z'\eta,$$

which is nondegenerate and homogeneous of degree one in ζ, η, θ. The critical points are

$$\Phi_\theta = \phi_\theta = 0, \quad \Phi_\eta = z' = 0, \quad \Phi_z = -\zeta + \phi_z = 0, \quad \Phi_{z'} = \eta + \phi_{z'} = 0.$$

Suppose $(z, \zeta; z', \zeta') \in \Lambda_c'$ and we choose local coordinates so that $z' = 0$. Using stationary phase arguments (e.g., [9, Proposition 1.2.4]), we obtain that

$$\widehat{I}(\zeta) = c_{ij}\langle\zeta\rangle^{-1}\widehat{f}_j(\zeta') + \mathcal{R}_1'$$

with new parameters c_{ij} and where $\langle\zeta\rangle^{(s+2)/2}\mathcal{R}_1' \in L^2$ and $\|\mathcal{F}^{-1}\mathcal{R}_1'\|_{H^{s+2}} = O(\epsilon)$.

Recall that $S(z, \zeta) = (z', \zeta')$. Let S_{ij} be a 4×4 matrix such that $|S_{ij}\zeta - S(z, \zeta)| \leq C\delta|\zeta|$ for $z \in \mathscr{U}_i, z' \in \mathscr{U}_j, |\zeta| > 1$. Here, C depends on S. But we know from section 7 that S is the solution of ODEs (7.1) with coefficients depending on $c(x)$ and its first derivatives. By the stability of ODEs, we see that C depends on $||c||_{C^1(I(p_-,p_+))}$. Then we get

$$\widehat{I}(\zeta) = c_{ij}\langle\zeta\rangle^{-1}\widehat{f}_j(S_{ij}\zeta) + \mathcal{R}_2 = h_{ij}^{(1)}(f; \Theta) + \mathcal{R}_2, \quad i \neq j.$$

Here, Θ is the collection of parameters of the network including c_{ij}, S_{ij}. To estimate $\mathcal{R}_2$, we recall that $f_j \in H^s$ and we have for $k \leq s$ that

$$\begin{aligned}\langle\zeta\rangle^k|\widehat{f}_j(\zeta + \delta\zeta) - \widehat{f}_j(\zeta)| &= \langle\zeta\rangle^k\left|\int(e^{iz(\zeta+\delta\zeta)} - e^{iz\zeta})f_j(z)dz\right|\\ &\leq C\delta\sum_{|\alpha|=k}\left|\int\partial_z^\alpha(e^{iz(\zeta+\delta\zeta)} - e^{iz\zeta})f_j(z)dz\right| \leq C\delta\|f\|_{H^k}\end{aligned}$$

because f_j is supported in $\mathscr{U}_i \subset I(p_-, p_+)$. So the constant C depends on the size of $I(p_-, p_+)$. Let $v_i = \phi_i v$. Therefore, we proved that

$$\|\Psi_R(\zeta)\langle\zeta\rangle^{(s+1)/2}(\widehat{v}_i - h^{(1)}_{ij}(f;\Theta))\|_{L^2} \leq C\delta\|f\|_{H^s(\mathscr{V})} + O(\epsilon R^{-1}) \leq C\epsilon(\delta + R^{-1}) \leq C\epsilon\delta \tag{10.1}$$

if $R > 1/\delta$ is large enough.

Next, we consider solving $Pv = f_j$ on $\mathscr{U}_j$. We want to use Q_0 on $N^*\text{Diag}$, which is a pseudodifferential operator, so we ignore the part on Λ_c. So we introduce a microlocal cut-off Φ supported sufficiently close to $\Lambda_c \cap N^*\text{Diag}$. In particular, we let $\chi(z,\zeta)$ be smooth in $T^*\mathscr{M}$ and $\chi(z,\zeta) = 1$ in $\mathcal{P}(z,\zeta) < \delta$ and $\chi(z,\zeta) = 0$ in $\mathcal{P}(z,\zeta) > 2\delta$. Then let $\widetilde{\chi}(t)$ be a smooth cut-off function so that $\widetilde{\chi}(t) = 1$ for $|t| < \delta$ and $\widetilde{\chi}(t) = 0$ for $|t| > 2\delta$. Then we set $\Phi(z,\zeta,z',\zeta') = \chi(z,\zeta)\widetilde{\chi}(|z-z'| + |\zeta-\zeta'|)$. Because $\text{diam}(\mathscr{U}_i) < \delta$, we still have a δ order error. More precisely,

$$\phi_i(z)Q_0(f_j) - \phi_i(z)\int e^{i(z-z')\zeta}(1-\chi(z,\zeta))\frac{f_j(z')}{\mathcal{P}(z,\zeta)}dz'd\zeta = \mathcal{R}_3,$$

where $\|\mathcal{R}_3\|_{H^{s+1}} \leq C\delta\|f\|_{H^s(\mathscr{V})} \leq C\delta\epsilon$ and C depends on the symbol only. Then we estimate

$$\phi_i(z)\int e^{i(z-z')\zeta}(1-\chi(z,\zeta))\frac{f_j(z')}{\mathcal{P}(z,\zeta)}dz'd\zeta - \phi_i(z)\int e^{i(z-z')\zeta}(1-\chi(z,\zeta))\frac{f_j(z')}{|\tau|^2 - c_{jj}^2|\xi|^2}dz'd\zeta = \mathcal{R}_4.$$

Using the same argument as we used for (10.1), we see that $\|\mathcal{R}_4\|_{H^{s+2}} \leq C\delta\|f\|_{H^s} \leq C\delta\epsilon$ if $|c_{jj}^2 - c^2(x)| \leq C\delta$ on $\mathscr{U}_i$. So we have proved (10.1) for $v_j = \phi_j v$.

Now we consider the second layer $h^{(2)}$, and we need the nonlinear function $F(t,x,u)$. On each $\mathscr{U}_i$, we write $F(t,x,u)$ in Taylor expansions:

$$F(t,x,u) = \sum_{j=2}^{M} a_j^{(i)}(t,x)u^j + O(|u|^{M+1}).$$

Let a_{ij} be constants so that $|a_{ij} - a_j^{(i)}(t,x)| < C\delta$, where C is some constant depending on the sup norm $|\partial_{(t,x)}a_j^{(i)}(t,x)|_\infty, i = 1,2,\ldots,K$. Thus for $p(u) = \sum_{j=2}^M a_{ij}u^j$, we obtain that $|F(t,x,u) - p(u)| \leq C\delta|u|^2$. Let $v_j^{(1)}$ be obtained from $h_j^{(1)}$ in the network and v_j be the solution of linearized wave equation. Using (10.1), we apply Proposition 8.1, Corollary 8.2, and the remark after them to get that $\widehat{p(v_j)} - v_j^{(1)} = \widehat{\mathcal{R}}_{sp} + \widehat{\mathcal{R}}_{ph}$, and they satisfy

$$\|\mathcal{R}_{sp}\|_{H^{s+1}} < C\delta\epsilon^2, \quad \|\Psi_R(D)\mathcal{R}_{ph}\|_{H^{s+1}} \leq C\epsilon^2\delta R^{-r},$$

where $r < s-1$. Next, we can apply the argument above to solve the linear wave equation $Pw = v_j^{(1)}$ on each $\mathscr{U}_i$ to get w_i. Then we obtain the term $h_{ij}^{(2)}$ in the network and

$$\widehat{w}_i - \sum_{j=1}^{K} h_{ij}^{(2)} = \mathcal{R}_5,$$

where

$$\|\Psi_R(\zeta)(1+|\zeta|)^{(s+1)/2}\mathcal{R}_5\|_{L^2} \leq C\delta\epsilon^2(1+R^{-r})$$

for R large enough. Together with Proposition 4.1, we complete the analysis for the iteration step in the network and obtain

$$\|\Psi_R(\zeta)\langle\zeta\rangle^{(s+1)/2}(\widehat{u}_i(\zeta) - h_i^{(2)}(f;\Theta))\|_{L^2(\mathbb{R}^4)} \leq C_2\epsilon^2 + C\delta\epsilon^2(1 + R^{-r}),$$

where C_2 is the constant in Proposition 4.1 which depends on c and F. The proof is finished by induction. ∎

11. Concluding remarks.

To summarize, we constructed a deep neural network to solve the inverse problems related to nonlinear wave equations. The network combines the physical model describing the wave phenomena into a data driven network structure. Moreover, based on theoretical works for the inverse problem, we proposeed constructing a cost function which captures the information in the phase space, and we obtained quantitative approximation results.

It is useful to remark on some issues that we have not explored in detail. First, it is not hard to see that our network generalizes to a large class of nonlinear hyperbolic equations on manifolds, although we have focused on (1.1). In fact, for strictly hyperbolic operators P, one can construct an approximation of $Q = P^{-1}$ as Fourier integral operators (FIOs), thanks to the work of Hörmander and Melrose and Uhlmann [30]. Thus, the treatment in section 7 can be generalized to this class of equations. The treatments for the nonlinear term and the iteration scheme remain unchanged. Actually, one can include the nonlinear terms F which are less regular such as Sobolev functions. Therefore, the network proposed in section 9 can be generalized to a large class of equations. Also, generalization to higher dimensions is straightforward, although theoretical results for the inverse problems are less known.

Second, let us discuss the reconstruction of $c(x)$ and $F(t, x, u)$ on each $\mathscr{U}_i$ from the parameters. For fixed $i = 1, 2, \dots, K$, consider the collection of $\theta_{ik}, k = 1, 2, \dots M$, which are the parameter sets on $\mathscr{U}_i$. From the construction of the network and the proof, it is easy to see that $p_i(u)$ constructed from the network is the approximation of $F(t, x, u)$ on $\mathscr{U}_i$ in the sense that

$$|F(t, x, u) - p_i(u)| < C\delta\epsilon, \quad (t, x) \in \mathscr{U}_i, \quad |u| < \epsilon.$$

The reconstruction of $c(x)$ on $\mathscr{U}_i$ is c_{ii}, and by the proof of Theorem 10.1 we have

$$|c(x) - c_{ii}| \leq C\delta \text{ on } \mathscr{U}_i.$$

Not much is known about the stability of the inverse problem in section 2, but we add that quite recently, Hölder stability for determining the nonlinear term was proved in [20] for some semilinear wave equations. We also remark that although we do not use it, our network also finds the approximation of the canonical relations which are the S_{ij} in the parameters. These can also be used to reconstruct $c(x)$; see, for example, [15].

There are practical issues that require further consideration. For example, we have not discussed the discretization of the network, the training of network, which involves solving a nontrivial nonconvex optimization problem, or the well-known stability problem related to many local minima. We plan to investigate these issues in future works. There are now plenty of numerical studies in PDE learning. In [33], the authors studied the data driven PDE discovery using a physics-informed network in which some of the network parameters are coefficients of the equation. Despite the lack of theoretical support, numerical results there do look promising for the reconstruction of parameters in the differential equations.

Acknowledgment. The authors thank the anonymous referee for a careful reading and many thoughtful suggestions.

REFERENCES

[1] M. Beals, *Propagation and Interaction of Singularities in Nonlinear Hyperbolic Problems*, Progr. Math. 3, Birkhäuser Boston, Boston, MA, 1989.

[2] J.-M. Bony, *Calcul symbolique et propagation des singularités pour les équations aux dérivées partielles non linéaires*, Ann. Sci. École Norm. Sup. (4), 14 (1981), pp. 209–246.

[3] T. A. Bubba, M. Galinier, M. Lassas, M. Prato, L. Ratti, and S. Siltanen, *Deep Neural Networks for Inverse Problems with Pseudodifferential Operators: An Application to Limited-Angle Tomography*, preprint, https://arxiv.org/abs/2006.01620, 2020.

[4] R. R. Coifman and Y. Meyer, *Au delà des opérateurs pseudo-différentiels*, Astérisque 57, Société Mathématique de France, Paris, France, 1978.

[5] C. Dafermos and W. Hrusa, *Energy methods for quasilinear hyperbolic initial-boundary value problems. Applications to elastodynamics*, Arch. Ration. Mech. Anal., 87 (1985), pp. 267–292.

[6] M. de Hoop, H. Smith, G. Uhlmann, and R. van der Hilst, *Seismic imaging with the generalized Radon transform: A curvelet transform perspective*, Inverse Problems, 25 (2009), 025005.

[7] M. de Hoop, G. Uhlmann, and A. Vasy, *Diffraction from conormal singularities*, Ann. Sci. Éc. Norm. Supér. (4), 48 (2015), pp. 351–408.

[8] M. de Hoop, G. Uhlmann, and Y. Wang, *Nonlinear interaction of waves in elastodynamics and an inverse problem*, Math. Ann., 376 (2020), pp. 765–795.

[9] J. J. Duistermaat, *Fourier Integral Operators*, Progr. Math. 130, Birkhäuser Boston, Boston, MA, 1996.

[10] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, Cambridge, MA, 2016.

[11] K. Hornik, M. Stinchcombe, and H. White, *Multilayer feedforward networks are universal approximators*, Neural Networks, 2 (1989), pp. 359–366.

[12] K. Hornik, M. Stinchcombe, and H. White, *Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks*, Neural Networks, 3 (1990), pp. 551–560.

[13] L. Hörmander, *The Analysis of Linear Partial Differential Operators* IV: *Fourier Integral Operators*, Classics Math., Springer-Verlag, Berlin, 2009.

[14] A. Greenleaf and G. Uhlmann, *Estimates for singular Radon transforms and pseudodifferential operators with singular symbols*, J. Funct. Anal., 89 (1990), pp. 202–232.

[15] A. Kachalov, M. Lassas, and Y. Kurylev, *Inverse Boundary Spectral Problems*, Chapman and Hall/CRC, Boca Raton, FL, 2001.

[16] K. Kothari, I. Dokmanić, and M. de Hoop, *Learning the Geometry of Wave-Based Imaging*, preprint, https://arxiv.org/abs/2006.05854v1, 2020.

[17] Y. Kurylev, M. Lassas, and G. Uhlmann, *Inverse problems for Lorentzian manifolds and nonlinear hyperbolic equations*, Invent. Math., 212 (2018), pp. 781–857.

[18] Y. Kurylev, M. Lassas, and G. Uhlmann, *Inverse Problems in Spacetime* I: *Inverse Problems for Einstein Equations*, extended preprint, https://arxiv.org/abs/1405.4503, 2014.

[19] B. N. Kuvshinov, T. J. H. Smit, and X. H. Campman, *Non-linear interaction of elastic waves in rocks*, Geophys. J. Internat., 194 (2013), pp. 1920–1940.

[20] M. Lassas, T. Liimatainen, L. Potenciano-Machado, and T. Tyni, *Uniqueness and Stability of an Inverse Problem for a Semi-Linear Wave Equation*, preprint, https://arxiv.org/abs/2006.13193, 2020.

[21] M. Lassas, G. Uhlmann, and Y. Wang, *Inverse problems for semilinear wave equations on Lorentzian manifolds*, Comm. Math. Phys., 360 (2018), pp. 555–609.

[22] M. Lassas, G. Uhlmann, and Y. Wang, *Determination of Vacuum Space-Times from the Einstein-Maxwell Equations*, preprint, https://arxiv.org/abs/1703.10704, 2017.

[23] Y. LeCun, Y. Bengio, and G. Hinton, *Deep learning*, Nature, 521 (2015), 436.

[24] M. Leshno, V. Y. Lin, A. Pinkus, and S. Schocken, *Multilayer feedforward networks with a non-polynomial activation function can approximate any function*, Neural Networks, 6 (1993), pp. 861–867.

[25] Z. Long, Y. Lu, and B. Dong, *PDE-Net* 2.0*: Learning PDEs from data with a numeric-symbolic hybrid deep network*, J. Comput. Phys., 399 (2019), 108925.

[26] L. Lu, P. Jin, and G. E. Karniadakis, *DeepONet: Learning Nonlinear Operators for Identifying Differential Equations Based on the Universal Approximation Theorem of Operators*, preprint, https://arxiv.org/abs/1910.03193, 2019.

[27] S. Mallat, *Group invariant scattering*, Comm. Pure Appl. Math., 65 (2012), pp. 1331–1398.

[28] S. Mallat, *Understanding deep convolutional networks*, Phil. Trans. R. Soc. A, 374 (2016), 20150203.

[29] R. Melrose and N. Ritter, *Interaction of nonlinear progressing waves for semilinear wave equations*, Ann. of Math. (2), 121 (1985), pp. 187–213.

[30] R. Melrose and G. Uhlmann, *Lagrangian intersection and the Cauchy problem*, Comm. Pure Appl. Math., 32 (1979), pp. 483–519.

[31] G. Nakamura and M. Watanabe, *An inverse boundary value problem for a nonlinear wave equation*, Inverse Probl. Imaging, 2 (2008), pp. 121–131.

[32] G. Nakamura and M. Vashisth, *Inverse Boundary Value Problem for Non-linear Hyperbolic Partial Differential Equations*, preprint, https://arxiv.org/abs/1712.09945, 2017.

[33] M. Raissi, P. Perdikaris, and G. E. Karniadakis, *Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations*, J. Comput. Phys., 378 (2019), pp. 686–707.

[34] J. Rauch and M. Reed, *Singularities produced by the nonlinear interaction of three progressing waves; examples*, Comm. Partial Differential Equations, 7 (1982), pp. 1117–1133.

[35] L. Ruthotto and E. Haber, *Deep neural networks motivated by partial differential equations*, J. Math. Imaging Vision, 62 (2020), pp. 352–364.

[36] M. E. Taylor, *Pseudodifferential Operators and Nonlinear Partial Differential Equations*, Birkhäuser Boston, Boston, MA, 1991.

[37] M. E. Taylor, *Partial Differential Equations* II: *Qualitative Studies of Linear Equations*, 2nd ed., Appl. Math. Sci. 116, Springer, New York, 2011.

[38] G. Uhlmann and Y. Wang, *Determination of space-time structures from gravitational perturbations*, Comm. Pure Appl. Math., 73 (2020), pp. 1315–1367.