Imagine a perfectly elastic $d$-dimensional sheet, infinite in extent, immersed (possibly with self-intersection) in a Riemannian manifold $(M,g)$. One can think of this sheet as a static map $\phi : \Omega \to M$ from $d$-dimensional spatial domain into $M$ (in equilibrium with regards to surface tension) or as a freely evolving map $\phi : \mathbb{R} \times \Omega \to M$ from a spacetime $\mathbb{R} \times \Omega$ into $M$. (For simplicity, we will primarily focus on Euclidean spatial domains $\Omega = \mathbb{R}^d$ here.) What are the ideal equations of motion for this sheet?

For the evolution of a point particle $\phi : t \mapsto \phi(t)$ moving freely in $M$ (i.e. the evolutionary case when $d = 0$), the motion is described by the geodesic flow equation, which can be described in a number of different ways, all of which are of importance. In coordinates, the geodesic flow equation is given by the nonlinear ordinary differential equation

$$\partial_{tt} \phi = -\Gamma(\phi)(\partial_t \phi, \partial_t \phi),$$

where $\Gamma$ is the Christoffel symbol. In coordinate-free terms, this equation can be written as

$$D_t \partial_t \phi = 0,$$

where $D_t = (\phi^* \nabla)_t$ is the $\partial_t$ component of the pullback $\phi^* \nabla$ of the Levi–Civita connection $\nabla$ on the tangent bundle $TM$ to the pullback bundle $\phi^* TM$. In variational terms, the geodesic flow equation can be viewed as the formal Euler–Lagrange equation for the functional

$$\mathcal{L}(\phi) := \frac{1}{2} \int_{\mathbb{R}} |\partial_t \phi|^2_{g(\phi)} \, dt,$$

where $|\partial_t \phi|_{g(\phi)}$ denotes the magnitude of the tangent vector $\partial_t \phi$ at $\phi$ with respect to the metric $g$. Finally, in Hamiltonian terms, one can view the geodesic flow equation as the Hamiltonian flow associated to the energy

$$E(\phi[t]) := \frac{1}{2} |\partial_t \phi(t)|^2_{g(\phi(t))},$$

where we use $\phi[t] := (\phi(t), \partial_t \phi(t))$ to denote the full state of the particle $\phi$ at time $t$ (i.e. both its position and velocity).

For a static map $\phi : x \mapsto \phi(x)$ from a Euclidean space $\mathbb{R}^d$ into $M$, the equilibrium requirement is described mathematically by the harmonic map equation, which again can be described in a number of important ways. In coordinates, the harmonic map equation is written as

$$\Delta \phi = -\Gamma(\phi)(\partial_i \phi, \partial_j \phi),$$

where $\partial_i$ is differentiation in the $x_i$ spatial direction, $\Delta = \partial_i \partial_i$ is the usual Laplacian, and Roman indices such as $i$ are summed from 1 to $d$. In coordinate-free notation, this becomes

$$D_i \partial_i \phi = 0,$$
and in variational terms, harmonic maps are formal critical points for the Dirichlet energy functional
\[ \mathcal{L}(\phi) := \frac{1}{2} \int_{\mathbb{R}^d} \langle \partial_i \phi, \partial_i \phi \rangle_{g(\phi)} \, dx. \]
When the target manifold \( M \) is a Euclidean space, a harmonic map is the same concept as a harmonic function, and so the general harmonic map equation can be viewed as a nonlinear generalization of Laplace’s equation \( \Delta \phi = 0 \). A typical example of a harmonic map is the stereographic projection from \( \mathbb{R}^2 = \{(x, y, 0) : x, y \in \mathbb{R} \} \) to the unit sphere \( S^2 = \{(x, y, z) : x^2 + y^2 + (z - 1)^2 = 1 \} \) by projecting from the north pole \((0, 0, 2)\). Harmonic maps arise naturally in the theory of minimal surfaces, in conformal or complex geometry, and in the topology of manifolds.

The natural common generalization of the geodesic flow equation (in time) and the harmonic maps equation (in space) is the wave maps equation for an evolving map \( \phi : \mathbb{R}^{1+d} \to M \) from Minkowski spacetime \( \mathbb{R}^{1+d} = (\mathbb{R}^{1+d}, h) \) into \( M \), which in coordinates takes the form
\[ (0.1) \quad \Box \phi = -\Gamma(\phi)(\partial^\alpha \phi, \partial_\alpha \phi), \]
where \( \Box := \partial^\alpha \partial_\alpha \) is the d’Alambertian operator, and \( \alpha \) runs over the spacetime indices \( 0, \ldots, d \) and is raised and lowered according to the Minkowski metric \( dh^2 = -dt^2 + dx^2 \). In coordinate-free notation, this becomes
\[ (0.2) \quad D^\alpha \partial_\alpha \phi = 0. \]
In variational terms, wave maps are formal critical points of the functional
\[ (0.3) \quad \mathcal{L}(\phi) := \frac{1}{2} \int_{\mathbb{R}^{1+d}} \langle \partial^\alpha \phi, \partial_\alpha \phi \rangle_{g(\phi)} \, dx dt \]
and can also be viewed as the Hamiltonian flow associated to the energy functional
\[ (0.4) \quad E(\phi[t]) := \frac{1}{2} \int_{\mathbb{R}^d} |\partial_t \phi(t, x)|^2_{g(\phi(t, x))} + |\nabla_x \phi(t, x)|^2_{g(\phi(t, x))} \, dx. \]
Solutions to this equation are known as wave maps or sigma models. They serve as simplified models for other, more complicated, relativistic field equations, such as the Yang–Mills equations and Einstein’s equations for gravity; in particular, they are perhaps the simplest nonlinear wave equation that admits a nontrivial gauge symmetry. In the opposite direction, wave maps are a more complicated variant of the scalar semilinear wave equation \( \Box \phi = F(\phi) \), which has been intensively studied in recent decades. As such, wave maps are an important test case for extending the well-developed theory of scalar semilinear wave equations to more geometric settings.

As it turns out, there is a fifth description of wave maps, in addition to the above four \((0.1)–(0.4)\), which is of importance in the analysis of these equations (including the one in the book under review), in which one takes the derivative map \( d\phi \) (which one views as a section of the pullback bundle \( \phi^*TM \)), expresses that map in terms of an orthonormal frame (or “gauge”) for that bundle, and then recasts the wave map equation \((0.2)\) (together with the torsion-free property of the Levi–Civita connection) as a covariant div-curl system that enjoys an additional gauge freedom coming from the ability to arbitrarily change the orthonormal frame used to coordinatise the derivative map \( d\phi \); see [26] for details. However, for the purposes of this review we will not discuss this derivative formulation in much further detail here.
Being a nonlinear wave equation, the fundamental solvability problem for wave maps is the Cauchy problem, which we specify in the smooth category for simplicity: given a specified smooth initial datum $\phi[0] = (\phi(0), \partial_t \phi(0))$ obeying the obvious compatibility conditions (namely, that $\phi(0)$ lies on $M$ and $(\phi(0), \partial_t \phi(0))$ lies in the tangent bundle $TM$), does there exist a unique smooth solution $\phi$ to the wave maps equation (0.1) with this initial data? Furthermore, if the solution exists for all time, what are the asymptotics as the time coordinate $t$ goes to infinity, and if instead the solution only exists up to some finite time $T_*$, how does the solution “blow up” as $t$ approaches $T_*$? One can also ask related questions, such as quantitative estimates on the regularity, stability, and lifespan of solutions, as well as the question of what happens if one weakens the regularity hypotheses on the initial data or on the solution, but we will not focus on these additional interesting questions in this review.

From past experience with general nonlinear evolution equations, we know that there are two key features of the wave maps equation that will play a major role in answering the above questions. The first key feature is the conservation of the energy (0.4), as well as the closely related pointwise conservation $\partial_\alpha T^{\alpha\beta}$ of the stress-energy tensor

$$T^{\alpha\beta} = \frac{1}{2} \langle \partial^\alpha \phi, \partial^\beta \phi \rangle g(\phi) - \frac{1}{4} h^{\alpha\beta} \langle \partial^n \phi, \partial_\gamma \phi \rangle g(\phi).$$

Note that while a smooth map might not have finite energy due to slow decay at spatial infinity, it will always be locally of finite energy, which turns out to be sufficient for the question of global existence and regularity, due to the finite speed of propagation of the wave maps equation.

The second key feature is the scale invariance of the wave maps equation with respect to the scaling transformation $\phi \mapsto \phi^{(\lambda)}$, defined for any scaling parameter $\lambda > 0$ by

$$\phi^{(\lambda)}(t, x) := \phi \left( \frac{t}{\lambda}, \frac{x}{\lambda} \right),$$

which can be easily verified to preserve the class of wave maps. The case $d = 2$ of two spatial dimensions then plays a distinguished role, as it is the unique dimension for which the wave maps equation is energy-critical, in the sense that the energy (0.4) is invariant (i.e. dimensionless) with respect to the scaling (0.5). Informally, this means that for a given energy, the relative strength of the linear and nonlinear components of the wave maps equation (0.1) are equally matched, both at arbitrarily large and at arbitrarily small spatial scales. This is in contrast to the subcritical situation $d = 1$, in which the fine scale behavior is essentially linear, and one can easily establish global regularity of solutions from arbitrary smooth initial data (see [12], [3], [6], [16]), and also should be compared with the supercritical situation $d > 2$, in which one can construct self-similar solutions that blow up in finite time (at least when the target manifold is positively curved; see [2], [19], [22]).

Henceforth we restrict our attention to the energy-critical case $d = 2$. The study of critical equations (not just of wave equation type, but across all classes of PDE) has proven to be a fascinating and delicate story, requiring very sophisticated (and scale-invariant) estimates from harmonic analysis, and in particular in exploiting the phenomenon of concentration compactness, which is a major theme of the book under review. This study naturally splits into two subtopics: the perturbative
theory, in which one either makes a smallness hypothesis on the initial data (e.g. small energy), or works locally in time instead of globally; and the nonperturbative theory, in which one extends the perturbative theory into the regime of large data and long times.

For simpler equations, such as the semilinear wave equation, the perturbative theory can be handled by applying the contraction mapping theorem (or equivalently, the Picard iteration scheme) in a carefully chosen function space, using a number of harmonic analysis estimates on solutions to the linear wave equation, such as the Strichartz estimates. In the case of wave maps, the perturbative theory is surprisingly complicated; not only does one need to employ incredibly complicated spaces and estimates, but one must also perform an additional gauge transformation to prevent a logarithmic divergence coming from the interactions between scales. Nevertheless, this can all be accomplished, leading in particular to the conclusion that one has existence and uniqueness of global smooth wave maps into any reasonable target manifold as long as the initial data has sufficiently small energy; see \[10\], \[14\], \[33\], \[27\] for surveys of the long sequence of papers leading up to this result (which, for reasons of space, we are unable to survey here).

For the perturbative theory, the exact choice of target manifold does not play a decisive role (although some particularly simple manifolds, such as the sphere, allow for an easier treatment in the proofs). In contrast, in the nonperturbative theory there is an important distinction to be made between positively curved targets and negatively curved targets (which roughly speaking correspond to focusing and defocusing nonlinearities respectively for the semilinear wave equation). This can already be seen at the level of harmonic maps (which can be interpreted as the stationary solutions to the wave maps equation); positively curved targets can support nontrivial harmonic maps of finite energy (such as the stereographic projection to \(S^2\) mentioned earlier), whereas negatively curved targets cannot (as can be seen from the Bochner–Weitzenbock identity and an integration by parts; see e.g. \[5\] or \[27\]). In the positively curved case, it is now known that one can construct large energy solutions that blow up in finite time \[11\], \[15\]. In the negatively curved case (and in particular, in the case the target manifold is a hyperbolic space \(M = H^m\)), it was conjectured that no blowup occurs, and that solutions exist globally and remain smooth for all time. One of the main purposes of the text under review is to establish this conjecture for constant negative curvature surfaces, such as \(H^2\):

**Theorem 0.1** (Global regularity). Let \(M\) be a Riemann surface of constant negative curvature. Then for any smooth initial data \(\phi[0] : \mathbb{R}^2 \to TM\), there exists a unique smooth wave map \(\phi : \mathbb{R}^{1+2} \to M\) with this initial data.

This result was also simultaneously established by the author \[28\], \[32\] for constant negative curvature manifolds of any dimension, and by Sterbenz and Tataru \[20\], \[21\] for arbitrary smooth targets with bounded geometry; earlier results in these directions under additional symmetry conditions on the initial data may be found in \[4\], \[18\], \[17\], \[23\], \[24\], \[25\]. However, in contrast to these other works, the authors are able to go well beyond global regularity, and also obtain some (rather technical to state) estimates and qualitative conclusions on the asymptotic behavior of solutions (assuming, for sake of simplicity, that the initial data is constant outside of a compact set), and furthermore establish an important concentration compactness property of solutions, which is one of the strongest statements one can make about the well-posedness of a given equation.
To explain this further, we now step back briefly from wave maps to give an extremely abbreviated history of concentration compactness. A fundamental difficulty in functional analysis is the failure of the Bolzano–Weierstrass theorem in infinite dimensions: given a sequence of functions \( f_1, f_2, f_3, \ldots \) in some function space (e.g. the Sobolev space \( H^1(\mathbb{R}^d) \) for some \( d > 2 \)) of bounded norm, it is usually not the case that one can find a subsequence \( f_{n_j} \) which converges strongly to some limit \( v \), thus having a decomposition

\[
f_{n_j} = v + w_j.
\]

Standard counterexamples precluding such a decomposition include “moving bump” examples where each \( f_n \) is concentrated at a point \( x_n \) in space that goes to infinity as \( n \to \infty \), or “shrinking/growing bump” examples where each \( f_n \) is spread out over a spatial scale \( \lambda_n \) that is going to either zero or infinity as \( n \to \infty \). There are also countless “dispersed” examples in which the \( f_n \) do not behave like bump functions, but simply fluctuate randomly in space without converging strongly to any limit. This lack of compactness frustrates many strategies for understanding solutions to a PDE (for instance, by trying to locate an extremiser to a variational problem), and so a vast amount of effort has been devoted to finding substitutes for this lack of compactness. One general substitute in this regard is weak compactness: thanks to results such as the Banach–Alaoglu theorem, one can usually recover a version of the Bolzano–Weierstrass theorem if one is willing to allow the subsequence \( f_{n_j} \) to converge in the weak topology rather than the strong one. However, weak convergence is often insufficient for applications. A compromise is to work with notions of convergence intermediate between strong and weak convergence, such as convergence in a weaker norm than the original function space norm. For instance, if the initial function space was \( H^1(\mathbb{R}^d) \), one might study convergence in the Lebesgue space \( L^{2d/(d-2)}(\mathbb{R}^d) \), as the latter norm is controlled by the former norm thanks to the Sobolev embedding theorem. In this intermediate norm, the dispersed examples alluded to previously are no longer an obstruction (they converge to zero in intermediate norms), but the moving bump and shrinking/growing bump counterexamples are still present, as are superpositions of such examples. The concentration compactness phenomenon, which was first introduced and systematically developed by Lions [13], asserts, roughly speaking, that such superpositions are in fact the only obstruction to compactness; given any sequence \( f_1, f_2, \ldots \) bounded in (say) \( H^1(\mathbb{R}^d) \), one can extract a subsequence \( f_{n_j} \) which admits a profile decomposition

\[
f_{n_j} = \sum_k g_{j,k} v_k + w_j
\]

for some \( v_k \) independent of \( j \), for some \( w_j \) converging in the intermediate sense \( L^{2d/(d-2)}(\mathbb{R}^d) \) to zero, and where the \( g_{j,k} \) are elements of the relevant symmetry group acting on both \( H^1(\mathbb{R}^d) \) and \( L^{2d/(d-2)}(\mathbb{R}^d) \), which in this case is the affine group of translations and dilations (where the dilations are normalized to preserve the \( H^1(\mathbb{R}^d) \) and \( L^{2d/(d-2)}(\mathbb{R}^d) \) norms). Here we are glossing over several technical details about the nature of convergence of the sum \( \sum_k g_{j,k} v_k \). While this appears to be a significantly more complicated notion of convergence than (0.6), the fact that the error can be made small in a normed sense, rather than simply converging weakly to zero, is of importance in PDE applications (particularly because many stability results about PDE and their associated functionals are stated with respect
to norms rather than in weak topologies). Very roughly speaking, these concentration compactness results can be proven by iteratively locating places and scales where the functions \( f_{n_j} \), “concentrate”, renormalizing such components of \( f_{n_j} \) into a single profile \( v_k \), and then subtracting off the effect of that profile and continuing the iteration.

The original formulation of the concentration compactness phenomenon was phrased for sequences in linear spaces, such as Sobolev spaces, or the space of solutions to a certain linear differential equation with a finiteness condition imposed on some norm of the solution. However, in 1999, Bahouri and Gérard observed that a version of concentration compactness also held for certain nonlinear classes of functions, and specifically for finite energy solutions \( f_n \) to a certain energy-critical nonlinear wave equation. Roughly speaking, their main result was that given any such sequence of solutions with uniformly bounded energy, one could find a subsequence \( f_{n_j} \) with a profile decomposition (0.7), where \( v_k \) were also solutions to the same nonlinear equation, the error \( w_j \) converged to zero in suitable intermediate norms, and the \( g_{j,k} \) represent the symmetries of the space of solutions (in this case, the group generated by translations in both space and time, as well as scaling). The nonlinear proof, while being more complicated than the linear one, still followed the same basic strategy, namely to isolate locations and scales where the solutions \( f_n \) concentrated, and then removing the resulting profiles (using various approximate superposition principles to overcome the lack of perfect linearity).

In a seminal paper of Kenig and Merle, a general strategy was laid out in which the Bahouri–Gérard concentration compactness phenomenon (and variants thereof) could be used to demonstrate global regularity for critical nonlinear equations, together with some associated quantitative bounds. Suppressing many details, the strategy proceeds as follows. Suppose for sake of contradiction that one did not have global regularity or the associated bounds; then by standard perturbative theory, one could find a sequence of solutions \( f_n \) with finite energy, such that a certain spacetime norm of the \( f_n \) diverged to infinity as \( n \to \infty \). Furthermore, one could assume that the energy \( E_n \) of these solutions \( f_n \) converged to a critical energy \( E_{\text{crit}} \), below which all solutions were well behaved in suitable norms. Applying the profile decomposition (0.7), it turns out that one can then extract a single profile \( v \) whose energy is exactly the critical energy \( E_{\text{crit}} \). Furthermore, there is an important almost periodicity property to this profile, namely that at any given time it is localized in a single location in space and frequency, without dispersing across many locations or many scales. (Intuitively, the reason for this is that if such dispersion were to take place, then the profile would decouple into two or more component profiles of strictly smaller energy, which can be used to contradict the definition of the critical energy.) It is then often possible to combine this almost periodicity with other properties of solutions (such as monotonicity formulas of Morawetz or virial type) to obtain a contradiction, thus establishing global regularity. See surveys of this method and its variants, which remain among the most powerful tools currently known to attack critical nonlinear PDE.

Almost the entirety of the almost five hundred pages of the book under review is devoted to implementing this strategy for the energy-critical wave maps equation. There are several reasons why the argument is this lengthy. Firstly, as already mentioned briefly, in order to obtain a satisfactory perturbative theory, one has to pass to the derivative map formulation of wave maps, and then one has to
select a good gauge in order to keep the nonlinear terms at a manageable size. There are a number of gauges which are suitable for this purpose; in the setting of Riemann surfaces, it turns out that the Coulomb gauge (in which a certain divergence-free condition is imposed on the connection coefficients) is an acceptable choice. Unfortunately, the introduction of this gauge introduces a number of higher order terms into the nonlinearity, and a large part of the text is concerned with the establishment of estimates that can adequately control these terms.

Once this is done, the next stage is to run the Bahouri–Gérard concentration compactness method to decompose the wave map into profiles. Here, another difficulty emerges, which appears to be inherent to the wave maps equation and is not present in simpler models such as the semilinear wave equation. Namely, the principle of superposition partly fails when one attempts to superimpose a low frequency wave map with a high frequency wave map; the effect of the latter on the former remains negligible, but the effect of the former on the latter is nontrivial, with the low frequency component acting as a “magnetic” field that rotates the phases of the high frequency component. This partial interaction between frequencies requires an extremely careful profile decomposition, in which the lowest frequency components of a sequence of wave maps are extracted first, and the next lowest frequency components extracted with the assistance of the linearized equation around the previous components, and so forth until all components have been extracted.

Finally, a minimal energy profile is extracted, and by applying local energy estimates of Morawetz (or Pohozaev) type, it is shown that these profiles behave self-similarly or statically in a certain limit, so that they are associated to a non-trivial harmonic map (either from the plane, or from hyperbolic space). Known results about harmonic maps, combined with the negative curvature of the target, can then be used to obtain a contradiction and prove the main theorem.

In all, the book provides a detailed and careful presentation of one of the deepest results available for the wave maps equation, and is a remarkable technical achievement that exemplifies the vast amount of progress that the field of nonlinear wave and dispersive equations has experienced in recent decades.

References


28. T. Tao, *Global regularity of wave maps III. Large energy from $R^{1+2}$ to hyperbolic spaces*, preprint.


30. T. Tao, *Global regularity of wave maps V. Large data local wellposedness in the energy class*, preprint.


32. T. Tao, *Global regularity of wave maps VII. Control of delocalised or dispersed solutions*, preprint.


TERENCE TAO
DEPARTMENT OF MATHEMATICS, UCLA, LOS ANGELES, CALIFORNIA 90095-1555

E-mail address: tao@math.ucla.edu