Stability, consistency, and convergence of numerical discretizations

Douglas N. Arnold, School of Mathematics, University of Minnesota

Overview

A problem in differential equations can rarely be solved analytically, and so often is discretized, resulting in a discrete problem which can be solved in a finite sequence of algebraic operations, efficiently implementable on a computer. The *error* in a discretization is the difference between the solution of the original problem and the solution of the discrete problem, which must be defined so that the difference makes sense and can be quantified. *Consistency* of a discretization refers to a quantitative measure of the extent to which the exact solution satisfies the discrete problem. *Stability* of a discretization refers to a quantitative measure of the well-posedness of the discrete problem. A fundamental result in numerical analysis is that *the error of a discretization may be bounded in terms of its consistency and stability.*

A framework for assessing discretizations

Many different approaches are used to discretize differential equations: finite differences, finite elements, spectral methods, integral equation approaches, etc. Despite the diversity of methods, fundamental concepts such as error, consistency, and stability are relevant to all of them. Here we describe a framework general enough to encompass all these methods, although we do restrict to linear problems to avoid many complications. To understand the definitions, it is good to keep some concrete examples in mind, and so we start with two of these.

A finite difference method

As a first example, consider the solution of the Poisson equation, $\Delta u = f$, on a domain $\Omega \subset \mathbb{R}^2$, subject to the Dirichlet boundary condition $u = 0$ on $\partial\Omega$. One possible discretization is a finite difference method, which we describe in the case $\Omega = (0,1) \times (0,1)$ is the unit square. Making reference to Figure 1, let $h = 1/n$, $n > 1$ integer, be the grid size, and define the grid domain, $\Omega_h = \{\, (lh, mh) \,|\, 0 < l, m < n \,\}$, as the set of grid points in Ω. The nearest neighbors of a grid point $p = (p_1, p_2)$ are the four grid points $p_W = (p_1 - h, p_2)$, $p_E = (p_1 + h, p_2)$, $p_S = (p_1, p_2 - h)$, and $p_N = (p_1, p_2 + h)$. The grid points which do not themselves belong to Ω, but which have a nearest neighbor in Ω constitute the grid boundary, $\partial\Omega_h$, and we set $\bar\Omega_h = \Omega_h \cup \partial\Omega_h$. Now let $v : \bar\Omega_h \to \mathbb{R}$ be a grid function. Its five-point Laplacian $\Delta_h v$ is defined by

$$\Delta_h v(p) = \frac{v(p_E) + v(p_W) + v(p_S) + v(p_N) - 4v(p)}{h^2}, \quad p \in \Omega_h.$$

The finite difference discretization then seeks $u_h : \bar\Omega_h \to \mathbb{R}$ satisfying

$$\Delta_h u_h(p) = f(p),\ p \in \Omega_h, \quad u_h(p) = 0,\ p \in \partial\Omega_h.$$

If we regard as unknowns the $N = (n-1)^2$ values $u_h(p)$ for $p \in \Omega_h$, this gives us a systems of N linear equations in N unknowns which may be solved very efficiently.

Article from *Encyclopedia of Applied and Computational Mathematics,* Springer-Verlag 2015.

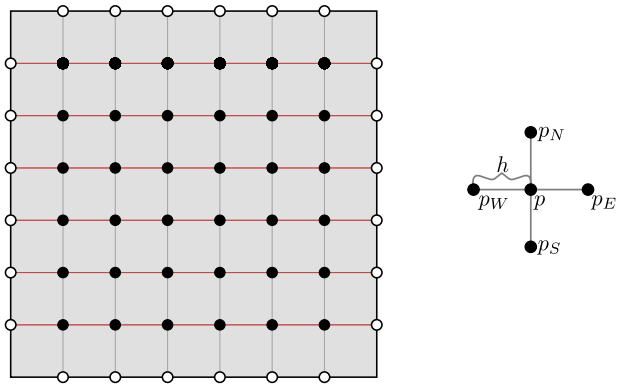

Fig. 1. The grid domain $\bar{\Omega}_h$ consists of the points in Ω_h, marked with solid dots, and in $\partial\Omega_h$, marked with hollow dots. On the right is the stencil of the five-point Laplacian, which consists of a grid point p and its four nearest neighbors.

A finite element method

A second example of a discretization is provided by a finite element solution of the same problem. In this case we assume that Ω is a polygon furnished with a triangulation $\mathcal{T}_h$, such as pictured in Figure 2. The finite element method seeks a function $u_h : \Omega \to \mathbb{R}$ which is continuous and piecewise linear with respect to the mesh and vanishing on $\partial\Omega$, and which satisfies

$$-\int_\Omega \nabla u_h \cdot \nabla v \, dx = \int_\Omega f v \, dx,$$

for all test functions v which are themselves continuous and piecewise linear with respect to the mesh and vanish on $\partial\Omega$. If we choose a basis for this set of space of test functions, then the computation of u_h may be reduced to an efficiently solvable system of N linear equations in N unknowns, where, in this case, N is the number of interior vertices in the triangulation.

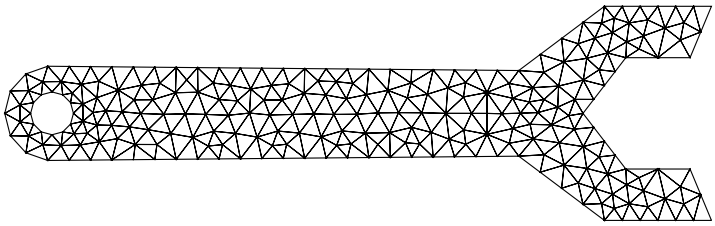

Fig. 2. A finite element mesh of the domain Ω. The solution is sought as a piecewise linear function with respect to the mesh.

Discretization

We may treat both these examples, and many other discretizations, in a common framework. We regard the discrete operator as a linear map L_h from a vector space V_h, called the discrete solution space, to a second vector space W_h, called the discrete data space. In the case of the finite difference operator, the discrete solution space is the space of mesh functions on $\bar{\Omega}_h$ which vanish on $\partial\Omega_h$, the discrete data space is the space of mesh functions on Ω_h, and the discrete operator $L_h = \Delta_h$, the five-point Laplacian. In the case of the finite element method, V_h is the space of continuous piecewise linear functions with respect to the given triangulation that vanish on $\partial\Omega$, and $W_h = V_h^*$, the dual space of V_h. The operator L_h is given by

$$(L_h w)(v) = -\int_\Omega \nabla w \cdot \nabla v \, dx, \quad w, v \in V_h.$$

For the finite difference method we define the discrete data $f_h \in W_h$ by $f_h = f|_{\Omega_h}$, while for the finite element method $f_h \in W_h$ is given by $f_h(v) = \int f v \, dx$. In both cases, the discrete solution $u_h \in V_h$ is found by solving the discrete equation

$$L_h u_h = f_h. \tag{1}$$

Of course, a minimal requirement on the discretization is that the finite dimensional linear system (1) has a unique solution, i.e., that the associated matrix is invertible (so V_h and W_h must have the same dimension). Then the discrete solution u_h is well-defined. The primary goal of numerical analysis to ensure that the discrete solution is a good approximation of the true solution u in an appropriate sense.

Representative and error

Since we are interested in the difference between u and u_h, we must bring these into a common vector space, where the difference makes sense. To this end, we suppose that a *representative* $U_h \in V_h$ of u is given. The representative is taken to be an element of V_h which, though not practically computable, is a good approximation of u. For the finite difference method a natural choice of representative is the grid function $U_h = u|_{\Omega_h}$. If we show that the difference $U_h - u_h$ is small, we know that the grid values $u_h(p)$ which determine the discrete solution are close to the exact values $u(p)$. For the finite element method, a good possibility for U_h is the piecewise linear interpolant of u, that is, U_h is the piecewise linear function that coincides with u at each vertex of the triangulation. Another popular possibility is to take U_h to be the best approximation of u in V_h in an appropriate norm. In any case, the quantity $U_h - u_h$, which is the difference between the representative of the true solution and the discrete solution, defines the *error* of the discretization.

At this point we have made our goal more concrete: we wish to ensure that the error, $U_h - u_h \in V_h$, is small. To render this quantitative we need to select a norm on the finite dimensional vector space V_h with which to measure the error. The choice of norm is an important aspect of the problem presentation, and an appropriate choice must reflect the goal of the computation. For example, in some applications, a large error at a single point of the domain could be catastrophic, while in others only the average error over the domain is significant. In yet other cases, derivatives of u are the true quantities of interest. These cases would lead to different choices of norms. We shall denote the chosen norm of $v \in V_h$ by $\|v\|_h$. Thus we now have a quantitative goal for our computation: that the error $\|U_h - u_h\|_h$ be sufficiently small.

Consistency and stability

Consistency error

Having used the representative U_h of the solution to define the error, we also use it to define a second sort of error, the *consistency error*, also sometimes called the truncation error. The consistency error is defined to be $L_h U_h - f_h$, which is an element of W_h. Now U_h represents the true solution u, so the consistency error should be understood as a quantity measuring the extent to which the true solution satisfies the discrete equation (1). Since $Lu = f$, the consistency error should be small if L_h is a good representative of L and f_h a good representative of f. In order to relate the norm of the error to the consistency error, we need a norm on the discrete data space W_h as well. We denote this norm by $\|w\|'_h$ for $w \in W_h$ and so our measure of the consistency error is $\|L_h U_h - f_h\|'_h$.

Stability

If a problem in differential equations is well-posed, then, by definition, the solution u depends continuously on the data f. On the discrete level, this continuous dependence is called *stability*. Thus stability refers to the continuity of the mapping $L_h^{-1} : W_h \to V_h$, which takes the discrete data f_h to the discrete solution u_h. Stability is a matter of degree, and an unstable discretization is one for which the modulus of continuity of L_h^{-1} is very large.

To illustrate the notion of instability, and to motivate the quantitative measure of stability we shall introduce below, we consider a simpler numerical problem than the discretization of a differential equation. Suppose we wish to compute the definite integral

$$\gamma_{n+1} = \int_0^1 x^n e^{x-1}\, dx, \tag{2}$$

for $n = 15$. Using integration by parts, we obtain a simple recipe to compute the integral in short sequence of arithmetic operations:

$$\gamma_{n+1} = 1 - n\gamma_n,\ n = 1, \ldots, 15, \qquad \gamma_1 = 1 - e^{-1} = 0.632121\ldots. \tag{3}$$

Now suppose we carry out this computation, beginning with $\gamma_1 = 0.632121$ (so truncated after six decimal places). We then find that $\gamma_{16} = -576,909$, which is truly a massive error, since the correct value is $\gamma_{16} = 0.0590175\ldots$. If we think of (3) as a discrete solution operator (analogous to L_h^{-1} above) taking the data γ_1 to the solution γ_{16}, then it is a highly unstable scheme: a perturbation of the data of less than 10^{-6} leads to a change in the solution of nearly 6×10^5. In fact, it is easy to see that for (3), a perturbation ϵ in the data leads to an error of $15! \times \epsilon$ in solution—a huge instability. It is important to note that the numerical computation of the integral (2) is not a difficult numerical problem. It could be easily computed with Simpson's rule, for example. The crime here is solving the problem with the unstable algorithm (3).

Returning to the case of the discretization (1), imagine that we perturb the discrete data f_h to some $\tilde{f}_h = f_h + \epsilon_h$, resulting in a perturbation of the discrete solution to $\tilde{u}_h = L_h^{-1}\tilde{f}_h$. Using the norms in W_h and V_h to measure the perturbations and then computing the ratio, we obtain

$$\frac{\text{solution perturbation}}{\text{data perturbation}} = \frac{\|\tilde{u}_h - u_h\|_h}{\|\tilde{f}_h - f_h\|'_h} = \frac{\|L_h^{-1}\epsilon_h\|_h}{\|\epsilon_h\|'_h}.$$

We define the *stability constant* C_h^{stab}, which is our quantitative measure of stability, as the maximum value this ratio achieves for any perturbation ϵ_h of the data. In other words, the stability constant is the norm of the operator L_h^{-1}:

$$C_h^{\text{stab}} = \sup_{0 \neq \epsilon_h \in W_h} \frac{\|L_h^{-1}\epsilon_h\|_h}{\|\epsilon_h\|'_h} = \|L_h^{-1}\|_{\mathcal{L}(W_h, V_h)}.$$

Relating consistency, stability, and error

The fundamental error bound

Let us summarize the ingredients we have introduced in our framework to assess a discretization:

- the discrete solution space, V_h, a finite dimensional vector space, normed by $\|\cdot\|_h$

- the discrete data space, W_h, a finite dimensional vector space, normed by $\|\cdot\|'_h$
- the discrete operator, $L_h : V_h \to W_h$, an invertible linear operator
- the discrete data $f_h \in W_h$
- the discrete solution u_h determined by the equation $L_h u_h = f_h$
- the solution representative $U_h \in V_h$
- the error $U_h - u_h \in V_h$
- the consistency error $L_h U_h - f_h \in W_h$
- the stability constant $C_h^{\text{stab}} = \|L_h^{-1}\|_{\mathcal{L}(W_h,V_h)}$

With this framework in place, we may prove a rigorous error bound, stating that the error is bounded by the product of the stability constant and the consistency error:

$$\|U_h - u_h\|_h \le C_h^{\text{stab}} \|L_h U_h - f_h\|'_h. \tag{4}$$

The proof is straightforward. Since L_h is invertible,

$$U_h - u_h = L_h^{-1}[L_h(U_h - u_h)] = L_h^{-1}(L_h U_h - L_h u_h) = L_h^{-1}(L_h U_h - f_h).$$

Taking norms, gives

$$\|U_h - u_h\|_h \le \|L_h^{-1}\|_{\mathcal{L}(W_h,V_h)} \|L_h U_h - f_h\|'_h,$$

as claimed.

The fundamental theorem

A discretization of a differential equation always entails a certain amount of error. If the error is not small enough for the needs of the application, one generally refines the discretization, for example using a finer grid size in a finite difference method or a triangulation with smaller elements in a finite element method. Thus we may consider a whole sequence or family of discretizations, corresponding to finer and finer grids or triangulations or whatever. It is conventional to parametrize these by a positive real number h called the discretization parameter. For example, in the finite difference method, we may use the same h as before, the grid size, and in the finite element method we can take h to be the maximal triangle diameter, or something related to it. We shall call such a family of discretizations a discretization scheme. The scheme is called *convergent* if the error norm $\|U_h - u_h\|_h$ tends to 0 as h tends to 0. Clearly convergence is a highly desirable property: it means that we can achieve whatever level of accuracy we need, as long as we do a fine enough computation. Two more definitions apply to a discretization scheme. The scheme is *consistent* if the consistency error norm $\|L_h U_h - f_h\|'_h$ tends to 0 with h. The scheme is *stable* if the stability constant C_h^{stab} is bounded uniformly in h: $C_h^{\text{stab}} \le C^{\text{stab}}$ for some number C^{stab} and all h. From the fundamental error bound, we immediately obtain what may be called the fundamental theorem of numerical analysis: *a discretization scheme which is consistent and stable is convergent.*

Historical perspective

Consistency essentially requires that the discrete equations defining the approximate solution are at least approximately satisfied by the true solution. This is an evident requirement, and has implicitly guided the construction of virtually all discretization methods, from the earliest examples. Bounds on the consistency error are often not difficult to obtain. For finite difference methods, for example, they may be derived from Taylor's theorem, and, for finite element methods, from simple approximation

theory. Stability is another matter. Its central role was not understood until the mid-twentieth century, and there are still many differential equations for which it is difficult to devise or to assess stable methods.

That consistency alone is insufficient for the convergence of a finite difference method was pointed out in a seminal paper of Courant, Friedrichs, and Lewy [2] in 1928. They considered the one-dimensional wave equation and used a finite difference method, analogous to the five-point Laplacian, with a space-time grid of points (jh, lk) with $0 \le j \le n$, $0 \le l \le m$ integers and $h, k > 0$ giving the spatial and temporal grid size, respectively. It is easy to bound the consistency error by $O(h^2 + k^2)$, so setting $k = \lambda h$ for some constant $\lambda > 0$ and letting h tend to 0, one obtains a consistent scheme. However, by comparing the domains of dependence of the true solution and of the discrete solution on the initial data, one sees that this method, though consistent, cannot be convergent if $\lambda > 1$.

Twenty years later, the property of stability of discretizations began to emerge in the work of von Neumann and his collaborators. First, in von Neumann's work with Goldstine on solving systems of linear equations [4], they studied the magnification of round-off error by the repeated algebraic operations involved, somewhat like the simple example (3) of an unstable recursion considered above. A few years later, in a 1950 article with Charney and Fjørtoft [1] on numerical solution of a convection diffusion equation arising in atmospheric modeling, the authors clearly highlighted the importance of what they called computational stability of the finite difference equations, and they used Fourier analysis techniques to assess the stability of their method. This approach developed into von Neumann stability analysis, still one of the most widely used techniques for determining stability of finite difference methods for evolution equations.

During the 1950s, there was a great deal of study of the nature of stability of finite difference equations for initial value problems, achieving its capstone in the 1956 survey paper [3] of Lax and Richtmeyer. In that context, they formulated the definition of stability given above and proved that, for a consistent difference approximation, stability ensured convergence.

Techniques for ensuring stability

Finite difference methods

We first consider an initial value problem, for example the heat equation or wave equation, discretized by a finite difference method using grid size h and time step k. The finite difference method advances the solution from some initial time t_0 to a terminal time T by a sequence of steps, with the lth step advancing the discrete solution from time $(l-1)k$ to time lk. At each time level lk the discrete solution is a spatial grid function u_h^l, and so the finite difference method defines an operator $G(h, k)$ mapping u_h^{l-1} to u_h^l, called the *amplification matrix*. Since the amplification matrix is applied many times in the course of the calculation ($m = (T - t_0)/k$ times to be precise, a number which tends to infinity as k tends to 0), the solution at the final step u_h^m involves a high power of the amplification matrix, namely $G(h, k)^m$, applied to the data u_h^0. Therefore the stability constant will depend on a bound for $\|G(h, k)^m\|$. Usually this can only be obtained by showing that $\|G(h, k)\| \le 1$, or, at most, $\|G(h, k)\| \le 1 + O(k)$. As a simple example, we may consider an initial value problem for the heat equation with homogeneous boundary conditions on the unit square:

$$\frac{\partial u}{\partial t} = \Delta u, \quad x \in \Omega,\ 0 < t \le T,$$
$$u(x,t) = 0, \quad x \in \partial\Omega,\ 0 < t \le T, \qquad u(x,0) = u_0(x), \quad x \in \Omega,$$

which we discretize with the five-point Laplacian and forward differences in time:

$$\begin{aligned} \frac{u^l(p) - u^{l-1}(p)}{k} &= \Delta_h u^{l-1}(p), \quad p \in \Omega_h,\ 0 < l \le m, \\ u^l(p) = 0, \quad p \in \partial\Omega_h,\ 0 < l \le m, &\qquad u^0(p) = u_0(p), \quad p \in \Omega_h. \end{aligned} \tag{5}$$

In this case the norm condition on the amplification matrix $\|G(h,k)\| \le 1$ holds if $4k \le h^2$, but not otherwise, and, indeed, it can be shown that this discretization scheme is stable, if and only if that condition is satisfied. Figure 3 illustrates the tremendous difference between a stable and unstable choice of time step.

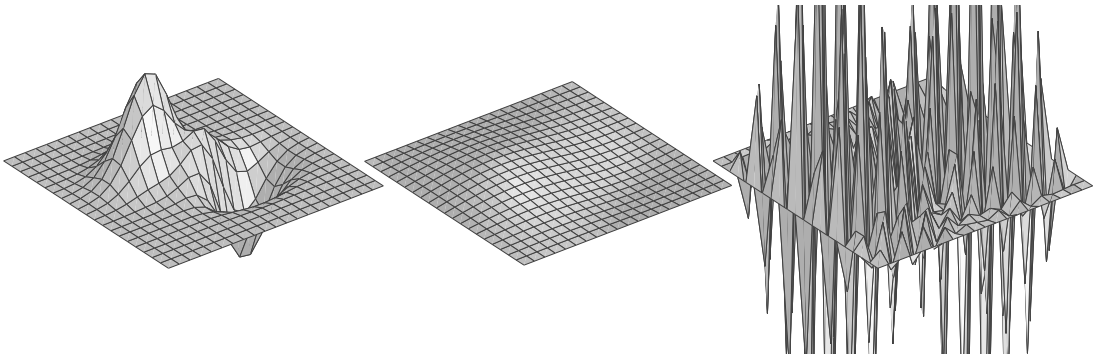

Fig. 3. Finite difference solution of the heat equation using (5). Left: initial data. Middle: discrete solution at $t = 0.03$ computed with $h = 1/20$, $k = 1/2,000$ (stable). Right: same computation with $k = 1/1,000$ (unstable).

Several methods are used to bound the norm of the amplification matrix. If an L^∞ norm is chosen, one can often use a discrete maximum principle based on the structure of the matrix. If an L^2 norm is chosen, then Fourier analysis may be used if the problem has constant coefficients and simple enough boundary conditions. In other circumstances, more sophisticated matrix or eigenvalue analysis is used.

For time-independent PDEs, such as the Poisson equation, the requirement is to show that the inverse of the discretization operator is bounded uniformly in the grid size h. Similar techniques as for the time-dependent problems are applied.

Galerkin methods

Galerkin methods, of which finite element methods are an important case, treat a problem which can be put into the form: find $u \in V$ such that $B(u,v) = F(v)$ for all $v \in V$. Here V is a Hilbert space, $B : V \times V \to \mathbb{R}$ is a bounded bilinear form, and $F \in V^*$, the dual space of V. (Many generalizations are possible, e.g., to the case where B acts on two different Hilbert spaces, or the case of Banach spaces.) This problem is equivalent to a problem in operator form, find u such $Lu = F$, where the operator $L : V \to V^*$ is defined by $Lu(v) = B(u,v)$. An example is the Dirichlet problem for the Poisson equation considered earlier. Then $V = \mathring{H}^1(\Omega)$, $B(u,v) = \int_\Omega \nabla u \cdot \nabla v \, dx$, and $F(v) = \int_\Omega fv \, dx$. The operator is $L = -\Delta : \mathring{H}^1(\Omega) \to \mathring{H}^1(\Omega)^*$.

A Galerkin method is a discretization which seeks u_h in a subspace V_h of V satisfying $B(u_h, v) = F(v)$ for all $v \in V_h$. The finite element method discussed above

took V_h to be the subspace of continuous piecewise linears. If the bilinear form B is coercive in the sense that there exists a constant $\gamma > 0$ for which

$$B(v, v) \geq \gamma \|v\|_V^2, \quad v \in V,$$

then stability of the Galerkin method with respect to the V norm is automatic. No matter how the subspace V_h is chosen, the stability constant is bounded by $1/\gamma$. If the bilinear form is not coercive (or if we consider a norm other than the norm in which the bilinear form is coercive), then finding stable subspaces for Galerkin's method may be quite difficult. As a very simple example, consider a problem on the unit interval $I = (0, 1)$, to find $(\sigma, u) \in H^1(I) \times L^2(I)$ such that

$$\int_0^1 \sigma\tau \, dx + \int_0^1 \tau' u \, dx + \int_0^1 \sigma' v \, dx = \int_0^1 f v \, dx, \quad (\tau, v) \in H^1(I) \times L^2(I). \tag{6}$$

This is a weak formulation of system $\sigma = u'$, $\sigma' = f$, with Dirichlet boundary conditions (which arise from this weak formulation as natural boundary conditions), so this is another form of the Dirichlet problem for Poisson's equation $u'' = f$ on I, $u(0) = u(1) = 0$. In higher dimensions, there are circumstances where such a first-order formulation is preferable to a standard second-order form. This problem can be discretized by a Galerkin method, based on subspaces $S_h \subset H^1(I)$ and $W_h \subset L^2(I)$. However, the choice of subspaces is delicate, even in this one-dimensional context. If we partition I into subintervals and choose S_h and W_h both to be the space of continuous piecewise linears, then the resulting matrix problem is *singular*, so the method is unusable. If we choose S_h to continuous piecewise linears, and W_h to be piecewise constants, we obtain a stable method. But if we choose S_h to contain all continuous piecewise quadratic functions, and retain the space of piecewise constants for W_h, we obtain an unstable scheme. The stable and unstable methods can be compared in Figure 4. For the same problem of the Poisson equation in first-order form, but in more than one dimension, the first stable elements were discovered in 1975 [5].

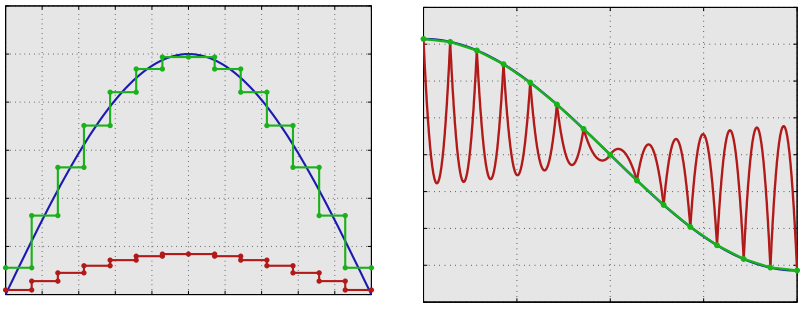

Fig. 4. Approximation of the problem (6), with $u = \cos \pi x$ shown on left, and $\sigma = u'$ on the right. The exact solution is shown in blue, and the stable finite element method, using piecewise linears for σ and piecewise constants for u, is shown in green (in the right plot, the blue curve essentially coincides with the green curve, and so is not visible). An unstable finite element method, using piecewise quadratics for σ, is shown in red.

References

[1] Charney JG, Fjørtoft R, von Neumann J (1950) Numerical integration of the barotropic vorticity equation. Tellus 2:237–254

[2] Courant R, Friedrichs K, Lewy H (1928) Über die partiellen Differenzengleichungen der mathematischen Physik. Math Ann 100(1):32–74
[3] Lax PD, Richtmyer RD (1956) Survey of the stability of linear finite difference equations. Communications on Pure and Applied Mathematics 9(2):267–293
[4] von Neumann J, Goldstine HH (1947) Numerical inverting of matrices of high order. Bull Amer Math Soc 53:1021–1099
[5] Raviart PA, Thomas JM (1977) A mixed finite element method for 2nd order elliptic problems. In: Mathematical aspects of finite element methods (Proc. Conf., Consiglio Naz. delle Ricerche (C.N.R.), Rome, 1975), Vol. 606 of Lecture Notes in Mathematics, Springer, Berlin, pp 292–315