2 Numerical Methods for Linear PDEs

2.1 Introduction and Examples

Consider the advection equation

$$\begin{cases} u_t + a u_x = 0, & -\infty < x < \infty, t \geq 0, \\ u(x,0) = u_0(x). \end{cases} \tag{2.1}$$

In order to numerically approximating the solution of (2.1), we first discretize the $x-t$ plane: set $h = \Delta x$ (mesh width) and $k = \Delta t$ (time step). This generates a lattice in the $x-t$ plane, i.e., equally spaced mesh points (x_j, t^n) where $x_j = jh$, $j = \ldots, -1, 0, 1, \ldots$ and $t^n = nk$, $n = 0, 1, \ldots$ With these notations, we have, e.g., $x_{j+1/2} = x_j + h/2 = (j + 1/2)\, h$.

We denote the pointwise values of the exact solution to (2.1) at the grid points, (x_j, t^n), as u_j^n, and by v_j^n we denote an approximation of u_j^n. We denote the forward differencing operator by $D_+ u_j^n = (u_{j+1}^n - u_j^n)/\Delta x$. Similarly, backward differencing and centered differencing are denoted, respectively, by $D_- u_j^n = (u_j^n - u_{j-1}^n)/\Delta x$, and $D_0 u_j^n = (u_{j+1}^n - u_{j-1}^n)/(2\Delta x)$. Often, we can find in the literature other notations for these operators, such as Δ^+ or Δ_+ for the forward difference, which may or may not include the factor $1/\Delta x$.

There are several possible strategies for approximating the solution of (2.1). One thing we can do is to replace the derivatives with finite-difference approximations. Alternatively, one can use Taylor based methods for deriving such approximations. Let's briefly explore both these possibilities.

Example 2.1 (Forward Euler)
Replace u_t in (2.1) with a forward difference approximation (in time), and replace u_x by a centered approximation (in space). Hence

$$\frac{v_j^{n+1} - v_j^n}{k} + a \left(\frac{v_{j+1}^n - v_{j-1}^n}{2h} \right) = 0,$$

i.e.,

$$v_j^{n+1} = v_j^n - \frac{\lambda a}{2} \left(v_{j+1}^n - v_{j-1}^n \right),$$

where the *mesh ratio* is defined here as $\lambda = k/h$.

Remarks.

1. If we are given the data at time step t^n then we can compute the data in the next time step t^{n+1}. Thus this is a *two-level method*.

2. This is an explicit method.

3. In matrix form, this method can be written as

$$\begin{pmatrix} v_0 \\ v_1 \\ \vdots \\ v_{N-1} \\ v_N \end{pmatrix}^{n+1} = \begin{pmatrix} 1 & -\frac{\lambda a}{2} & & & \frac{\lambda a}{2} \\ \frac{\lambda a}{2} & 1 & -\frac{\lambda a}{2} & & \\ & \ddots & \ddots & \ddots & \\ & & \frac{\lambda a}{2} & 1 & -\frac{\lambda a}{2} \\ -\frac{\lambda a}{2} & & & \frac{\lambda a}{2} & 1 \end{pmatrix} \begin{pmatrix} v_0 \\ v_1 \\ \vdots \\ v_{N-1} \\ v_N \end{pmatrix}^{n},$$

where the entries in the upper right and lower left appear due to the periodic boundary conditions. Adjustments should be made for different types of boundary conditions.

Example 2.2 (Backward Euler)
We repeat the same approximations we made in Example 2.1, only this time the spatial derivatives are evaluated at t^{n+1} :

$$\frac{v_j^{n+1} - v_j^n}{k} + a\left(\frac{v_{j+1}^{n+1} - v_{j-1}^{n+1}}{2h}\right) = 0.$$

Hence

$$v_j^{n+1} + \frac{\lambda a}{2}\left(v_{j+1}^{n+1} - v_{j-1}^{n+1}\right) = v_j^n.$$

This is an *implicit* equation which takes the matrix form

$$\begin{pmatrix} X & X & X & X & X \\ -\frac{\lambda a}{2} & 1 & \frac{\lambda a}{2} & & \\ & \ddots & \ddots & \ddots & \\ & & -\frac{\lambda a}{2} & 1 & \frac{\lambda a}{2} \\ X & X & X & X & X \end{pmatrix} \begin{pmatrix} v_0 \\ v_1 \\ \vdots \\ v_{N-1} \\ v_N \end{pmatrix}^{n+1} = \begin{pmatrix} v_0 \\ v_1 \\ \vdots \\ v_{N-1} \\ v_N \end{pmatrix}^{n}.$$

Again, the additional terms in the top and bottom rows should be determined by the boundary conditions.

The points we use in the computation of the v_j^{n+1} are the *stencil* of the method. In Example 2.1, the stencil is $\{(x_{j-1}, t^n), (x_j, t^n), (x_{j+1}, t^n), (x_j, t^{n+1})\}$. See Fig. 2.1a. In Example 2.2 the stencil is $\{(x_j, t^n), (x_{j-1}, t^{n+1}), (x_{j+1}, t^{n+1}), (x_j, t^{n+1})\}$ (see Fig 2.1b).

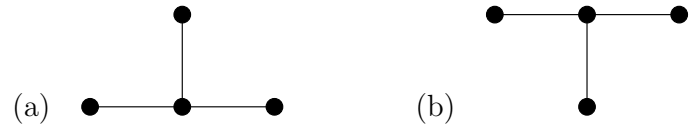

Figure 2.1: (a) The stencil of the Forward Euler method. (b) The stencil of the Backward Euler method.

Example 2.3 (Lax-Wendroff)
An example of a Taylor-based method is the Lax-Wendroff (LxW) method. We first expand

$$u(x,t+k) = u(x,t) + ku_t(x,t) + \frac{1}{2}k^2 u_{tt}(x,t) + \dots \tag{2.2}$$

Now the PDE we are solving is $u_t = -au_x$, so

$$u_{tt} = -au_{xt} = -au_{tx} = -a(-au_x)_x = a^2 u_{xx}. \tag{2.3}$$

Plugging (2.3) into (2.2), and replacing u_t with $-au_x$ yields

$$u(x,t+k) = u(x,t) - kau_x(x,t) + \frac{1}{2}k^2 a^2 u_{xx}(x,t) + \dots$$

The LxW scheme is obtained by truncating this series after the second derivative and replacing the derivatives with centered differencing:

$$v_j^{n+1} = v_j^n - \frac{ka}{2h}\left(v_{j+1}^n - v_{j-1}^n\right) + \frac{k^2a^2}{2h^2}\left(v_{j+1}^n - 2v_j^n + v_{j-1}^n\right). \tag{2.4}$$

The LxW method, (2.4), is an explicit two-level scheme with the three-point stencil $\{(x_{j-1},t^n),(x_j,t^n),(x_{j+1},t^n),(x_j,t^{n+1})\}$, which is exactly the same stencil as the one of the Forward Euler method (see Fig. 2.1a). It can be written in a matrix form as

$$\begin{pmatrix} v_0 \\ v_1 \\ \vdots \\ v_{N-1} \\ v_N \end{pmatrix}^{n+1} = A \begin{pmatrix} v_0 \\ v_1 \\ \vdots \\ v_{N-1} \\ v_N \end{pmatrix}^{n},$$

where

$$A = \begin{pmatrix} 1-\lambda^2a^2 & -\frac{\lambda a}{2}+\frac{1}{2}\lambda^2a^2 & & & X \\ \frac{\lambda a}{2}+\frac{1}{2}\lambda^2a^2 & 1-\lambda^2a^2 & -\frac{\lambda a}{2}+\frac{1}{2}\lambda^2a^2 & & \\ & \ddots & \ddots & \ddots & \\ & & \frac{\lambda a}{2}+\frac{1}{2}\lambda^2a^2 & 1-\lambda^2a^2 & -\frac{\lambda a}{2}+\frac{1}{2}\lambda^2a^2 \\ X & & & \frac{\lambda a}{2}+\frac{1}{2}\lambda^2a^2 & 1-\lambda^2a^2 \end{pmatrix}.$$

2.2 The Courant-Friedrichs-Levy (CFL) Condition

The Courant-Friedrichs-Levy (CFL) Condition comes from work done around 1928 which used finite difference methods to prove the existence of solutions of certain PDEs.

In order to heuristically demonstrate this condition, consider $u_t + au_x = 0$ and discretize it using forward differences in space and time

$$\frac{v_j^{n+1} - v_j^n}{k} + a\left(\frac{v_{j+1}^n - v_j^n}{h}\right) = 0. \tag{2.5}$$

This is a two-point scheme with a stencil $\{(x_j, t^n), (x_{j+1}, t^n), (x_j, t^{n+1})\}$. Now, if $a > 0$ the characteristics of $u_t + au_x = 0$ point in the positive x direction. Therefore the data at (x_j, t^{n+1}) depends on data to the left of x_j. See Fig 2.2. But any interval to the left of x_j is not included in the stencil of this method, so the solution computed by this method at (x_j, t^{n+1}) is based on information that does not correspond to the analytical direction of the flow of information in the problem. Hence, there is no hope that such an approximation would be correct in any sense.

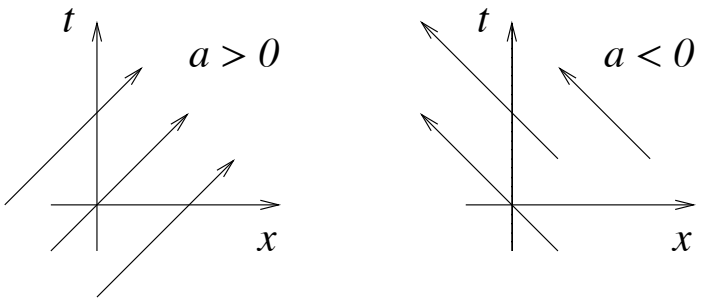

Figure 2.2: Characteristics for the advection equation

In this example, the *analytical domain of dependence* of the PDE (contained in the interval $[x_{j-1}, x_j] \times t^n$), is not contained in the *numerical domain of dependence* (determined by the stencil: in this case the interval $[x_j, x_{j+1}] \times t^n$). It is important to note that there is a hidden assumption of continuity, i.e., if we know the approximate solution at time t^n at x_j and x_{j+1}, then we can define a solution everywhere in the interval $[x_j, x_{j+1}]$.

The CFL condition states that *a necessary condition for the convergence of a numerical method is that the numerical domain of dependence contains the analytical domain of dependence*. It is important to note that the CFL condition is not a sufficient condition for the convergence of the approximate solution to the exact solution. In our example, the CFL condition requires that

$$x - ak \in (x, x + h).$$

See Fig. 2.3. Hence $0 \leq -ak \leq h$, i.e., $0 > \lambda a > -1$, which implies that $a < 0$ and $\lambda|a| < 1$. This is the CFL condition for (2.5).

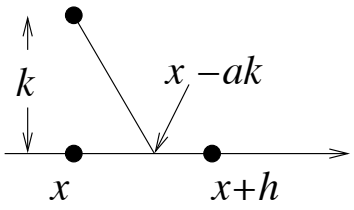

Figure 2.3: Numerical and analytical domains of dependence for $a < 0$.

The quantity λa is often called the *Courant number* and measures the "numerical speed".

2.3 Von-Neumann Stability Analysis

For a periodic function $u(x,t)$ we can write the Fourier series

$$u\left(x,t\right)=\frac{1}{\sqrt{2\pi}}\sum_{\omega=-\infty}^{\infty}\hat{u}\left(\omega,t\right)e^{i\omega x},$$

where

$$\hat{u}\left(\omega,t\right)=\frac{1}{\sqrt{2\pi}}\int_{-\pi}^{\pi}u\left(x,t\right)e^{-i\omega x}dx.$$

It is a well-known fact that a translation in space amounts to a phase shift in the Fourier space. In other words, if we define the "shift" (translation) operator as

$$E^{j}u\left(x\right)=u\left(x+jh\right),$$

then

$$\begin{aligned}\widehat{E^{j}u}(\omega) &= \frac{1}{\sqrt{2\pi}}\int_{-\pi}^{\pi}u\left(x+jh\right)e^{-i\omega x}dx\\ &= \frac{1}{\sqrt{2\pi}}\int_{-\pi+jh}^{\pi+jh}u\left(x\right)e^{-i\omega(x-jh)}dx=e^{ij\omega h}\hat{u}\left(\omega\right).\end{aligned}$$

At this point we have enough tools for studying the stability of numerical schemes for approximating solutions of linear problems with constant coefficients and with periodic boundary conditions. Consider, e.g., the Forward Euler (FE) scheme for $u_t = au_x$ (+periodic boundary conditions)

$$\frac{v\left(x,t+k\right)-v\left(x,t\right)}{k}=a\frac{v\left(x+h,t\right)-v\left(x-h,t\right)}{2h},$$

which we rewrite as

$$v\left(x,t+k\right)=v\left(x,t\right)+\frac{\lambda a}{2}\left[v\left(x+h,t\right)-v\left(x-h,t\right)\right]. \tag{2.6}$$

Fourier transforming (2.6) gives

$$\begin{aligned}\hat{v}\left(\omega,t+k\right) &= \hat{v}\left(\omega,t\right)+\frac{\lambda a}{2}\left[e^{i\omega h}-e^{-i\omega h}\right]\hat{v}\left(\omega,t\right)\\ &= \left[1+\lambda ai\sin\left(\omega h\right)\right]\hat{v}\left(\omega,t\right).\end{aligned}$$

We define the *amplification factor* (or *symbol*) $\hat{Q}$ by $\hat{v}\left(\omega,t+k\right)=\hat{Q}\hat{v}\left(\omega,t\right)$. In this example, $\hat{Q}=1+\lambda ai\sin\left(\omega h\right)$. Then if the initial data is $u\left(x,0\right)=f\left(x\right)$, then the Fourier transform of the approximate solution at time t^n equals $\hat{v}\left(\omega,t^n\right)=\hat{Q}^n\hat{v}\left(\omega,0\right)=\hat{Q}^n\hat{f}\left(\omega\right)$.

A numerical method is *stable* in the time interval $[0,T]$ for a sequence $k,h \to 0$ if for some constant $K(T)$,

$$\sup_{0\le t^n\le T;\omega,k,h} \left|\hat{Q}^n\right| \le K(T).$$

In the case of (2.6), $\hat{Q} = 1 + \lambda a i \sin(\omega h)$ so

$$\left|\hat{Q}\right| = \sqrt{1+\lambda^2 a^2 \sin^2(\omega h)}.$$

Therefore, $\left|\hat{Q}\right| > 1$ whenever $\sin^2(\omega h) \neq 0$ and $\left|\hat{Q}\right|^n$ will diverge in the limit $k,h \to 0$. We can conclude that the Forward Euler method is unconditionally unstable when the mesh ratio $\lambda = k/h$ is held fixed. It is left as a simple exercise to check that the FE methods is stable when the ratio k/h^2 is held fixed. Is there anything wrong with such stability condition? Yes. Such stability requirement forces the time-step to be too small (for a hyperbolic problem).

2.4 Artificial Viscosity

A natural question at this point is the following: given an unstable numerical method can we stabilize it? In general, for first-order linear equations, the answer to this question is positive. All we have to do is to add a sufficiently large quantity of "artificial viscosity". Numerically, this can be done by adding a dissipative term to the scheme, in the form of a second derivative. For simplicity, let's consider the case where the velocity is $a = 1$ so that the Cauchy problem becomes $u_t = u_x$, $u(x,0) = f(x)$. Consider a numerical scheme that is given by

$$\begin{cases} v_j^{n+1} = (I + kD_0)\, v_j^n + \sigma k h D_+ D_- v_j^n \\ v_j^0 = f_j \end{cases} \tag{2.7}$$

Here, σ is a constant that we still need to determined. We will see that there is a non-unique choice of σ that will result with a stable method. First, we rewrite (2.7) as

$$\frac{v_j^{n+1} - v_j^n}{k} = D_0 v_j^n + \sigma h D_+ D_- v_j^n. \tag{2.8}$$

In the limit $k \to 0$, equation (2.8) is a consistent approximation of

$$u_t = u_x + \sigma h u_{xx}.$$

As $h \to 0$ we return to $u_t = u_x$. Hence, (2.8), is a *consistent* approximation of $u_t = u_x$. Our goal is to choose σ, k, and h such that the resulting scheme is stable in the sense that $\left|\hat{Q}\right| \le 1$ (reference: Gustafsson, Kreiss and Oliger p. 44-46). A straightforward computation shows that the amplification factor of (2.7) is

$$\hat{Q} = 1 + i\lambda \sin\xi - 4\sigma\lambda \sin^2\frac{\xi}{2},$$

where $\xi = \omega h$ and $\lambda = k/h$. Hence

$$\left|\hat{Q}\right|^2 = 1 - \left(8\sigma\lambda - 4\lambda^2\right)\sin^2\frac{\xi}{2} + \left(16\sigma^2 - 4\right)\lambda^2\sin^4\frac{\xi}{2}.$$

There are two ways to proceed:

1. Suppose $2\sigma \leq 1$, then $16\sigma^2 - 4 \leq 0$. Then if $8\sigma\lambda - 4\lambda^2 \geq 0$ then $\left|\hat{Q}\right| \leq 1$. This implies that $0 < \lambda \leq 2\sigma \leq 1$. Example: take $\sigma = k/2h = \lambda/2$. With this choice of parameters we recover the Lax-Wendroff scheme:

   $$v_j^{n+1} = v_j^n + kD_0v_j^n + \frac{k^2}{2}D_+D_-v_j^n.$$

   What we just proved is that the LxW method is stable, assuming that the CFL condition $\lambda \leq 1$ is satisfied.

2. Suppose $2\sigma \geq 1$, so if $2\sigma\lambda \leq 1$ then $\left|\hat{Q}\right| \leq 1$. For example, take $\sigma = h/2k = 1/2\lambda$ so

   $$v_j^{n+1} = v_j^n + kD_0v_j^n + \frac{1}{2}h^2D_+D_-v_j^n,$$

   i.e.,

   $$v_j^{n+1} = \frac{1}{2}\left(v_{j+1}^n + v_{j-1}^n\right) + kD_0v_j^n. \tag{2.9}$$

   Equation (2.9) is the Lax-Friedrichs (LxF) scheme. The LxF scheme is also stable for $\lambda \leq 1$. Note that the only difference between the LxF scheme and the FE scheme is that the term v_j^n in (2.6) is replace with the average $\frac{1}{2}\left(v_{j+1}^n + v_{j-1}^n\right)$ in (2.9). This averaging has a stabilizing effect; the dissipation we introduced leads to stability.

2.5 Stability and Convergence

Consider the 2π-periodic initial value problem

$$\begin{cases} u_t(x,t) = P\left(x,t,\frac{\partial}{\partial x}\right)u(x,t), \\ u(x,0) = f(x), \end{cases} \tag{2.10}$$

for $h = \frac{2\pi}{N+1}$, $k > 0$. A general difference approximation of (2.10) can be written in the form

$$\begin{cases} Q_{-1}v^{n+1} = \sum_{\sigma=0}^{q} Q_\sigma v^{n-\sigma}, n = q, q+1, \ldots \\ v^\sigma = f^{(\sigma)}, \sigma = 0, 1, \ldots, q \end{cases} \tag{2.11}$$

where Q_σ are difference operators. Define the discrete solution operator S_h that evolves the solution from time t^ν to t^n through the relation

$$v^n = S_h(t^n, t^\nu)\, v^\nu.$$

Let $(u,v)_h = \sum_{j=1}^{N} \bar{u}_j v_j h$, and $\|u\|_h^2 = h\sum_{j=1}^{N}|u_j|^2$. We then have

Definition 2.4 The difference approximation (2.11) is *stable* for $0 < h \leq h_0$ if there exists constants α_S, c, κ_S such that for all h

$$\left\|Q_{-1}^{-1}\right\|_h \leq c,$$

$$\left\|S_h\left(t^n, t^\nu\right)\right\|_h \leq \kappa_S e^{\alpha_S\left(t^n - t^\nu\right)}.$$

A stable numerical scheme satisfies for all initial conditions f the following:

$$\left\|v^n\right\|_h \leq \kappa\left(t^n\right)\|f\|_h,$$

where $\kappa\left(t^n\right) = \kappa_S e^{\alpha_S t^n}$. The exponential factor is required so that we can treat approximations to differential equations with solutions that grow in time, such as $u_t = u_x + u$.

Definition 2.5 Let $u(x,t)$ be a smooth solution of (2.10). Then the *local truncation error* is defined by

$$k\tau_j^n = Q_{-1}u\left(x_j, t^{n+1}\right) - \sum_{\sigma=0}^{p} Q_\sigma u\left(x_j, t^{n-\sigma}\right).$$

The local truncation error is the extent to which the solution to the PDE fails to satisfy the difference approximation.

Definition 2.6 A difference approximation is *accurate of order* (p_1, p_2) if for all sufficiently smooth solutions $u(x,t)$ of the PDE if there is a function $L\left(t^n\right)$ such that for $h \leq h_0$

$$\left\|\tau^n\right\|_h \leq L\left(t^n\right)\left(h^{p_1} + k^{p_2}\right),$$

where $L\left(t^n\right)$ is bounded on every finite time interval. If $p_1 > 0$ and $p_2 > 0$ then the approximation is called *consistent*.

Example 2.7
The Leap-frog scheme for $u_t = u_x$ is obtained by approximating both derivatives with a centered approximation, i.e., for $\lambda = h/k$,

$$v_j^{n+1} = v_j^{n-1} + \lambda\left(v_{j+1}^n - v_{j-1}^n\right). \tag{2.12}$$

The local truncation error for the Leap-frog scheme (2.12) is given by

$$k\tau_j^n = u\left(x_j, t^{n+1}\right) - \lambda\left[u\left(x_{j+1}, t^n\right) - u\left(x_{j-1}, t^n\right)\right] - u\left(x_j, t^{n-1}\right). \tag{2.13}$$

Expanding the terms in (2.13) around (x_j, t^n), we have

$$\begin{aligned}
\tau_j^n &= \frac{1}{k}\left\{u + ku_t + \frac{k^2}{2}u_{tt} + \frac{k^3}{6}u_{ttt} + \frac{k^4}{24}u_{tttt} + O\left(k^5\right)\right. \\
&\quad - \left[u - ku_t + \frac{k^2}{2}u_{tt} - \frac{k^3}{6}u_{ttt} + \frac{k^4}{24}u_{tttt} + O\left(k^5\right)\right] \\
&\quad -\lambda\left[u + hu_x + \frac{h^2}{2}u_{xx} + \frac{h^3}{6}u_{xxx} + \frac{h^4}{24}u_{xxxx} + O\left(h^5\right)\right. \\
&\quad \left.\left. - \left(u - hu_x + \frac{h^2}{2}u_{xx} - \frac{h^3}{6}u_{xxx} + \frac{h^4}{24}u_{xxxx} + O\left(h^5\right)\right)\right]\right\} \\
&= \frac{1}{k}\left\{2ku_t + 2\frac{k^3}{6}u_{ttt} + O\left(k^5\right) - \lambda\left(2hu_x + 2\frac{h^3}{6}u_{xxx} + O\left(h^5\right)\right)\right\} \\
&= 2\left(u_t - u_x\right) + \frac{k^2}{3}u_{ttt} + \frac{h^2}{3}u_{xxx} + O\left(k^4, h^4\right) = \frac{k^2}{3}u_{ttt} + \frac{h^2}{3}u_{xxx} + O\left(k^4, h^4\right).
\end{aligned}$$

We can therefore conclude that the Leap-frog approximation is accurate of order $(2, 2)$.

Stability by itself is not sufficient to give a good approximation (for example, $v_j^n = 1$ for all j and n is a stable but wrong approximation to the solution of $u_t = u_x$). We have already seen that consistency is also not sufficient for a good approximation (forward Euler is consistent but unstable: the approximation will diverge). Fortunately, requiring both stability and consistency does guarantee convergence to the solution, as stated by

Theorem 2.8 (Lax Equivalence Theorem) *Assume that the solution of (2.10) is smooth and that the approximation (2.11) is stable. Assume that the approximation and its initial data are accurate of order* (p_1, p_2). *Then on any finite interval* $[0, T]$, *the error satisfies*

$$\begin{aligned}
\left\|v^n - u\left(\cdot, t^n\right)\right\|_h &\leq \kappa_S\left(e^{\alpha_S t^n}\left\|v^0 - u\left(\cdot, 0\right)\right\|_h + \left\|Q_{-1}^{-1}\right\|_h \varphi_h^*\left(\alpha, t^n\right) \max_{0\leq j\leq n-1}\left\|\tau^j\right\|_h\right) \\
&= O\left(h^{p_1} + k^{p_2}\right).
\end{aligned}$$

That is, the solutions of the difference approximation converge to the solution of the differential equation as $h \to 0$.

In other words, *stability and consistency implies convergence*. It is important to emphasize that we can expect such a result to hold only for linear problems. There is no such general theorem for nonlinear problems.