Matrix Stability Analysis

Consider the initial boundary value problem (IBVP)

\[ u_t = \sigma u_{xx}, \quad 0 < x < 1, t > 0 \]  (1)
\[ u(0, t) = g(t), \quad u(1, t) = h(t) \]  (2)
\[ u(x, 0) = f(x) \]  (3)

Equation (1) can be written as

\[ u_t = Lu, \]  (4)

where \( L \) is a linear differential operator.

We have seen three different numerical schemes to approximate the solution of IBVP (1)-(3). They are

1. **Forward in time–Centered in space**

   \[ U_i^{n+1} = rU_{i-1}^n + (1 - 2r)U_i^n + rU_{i+1}^n, \quad i = 1, \ldots, m \]  (5)

   where \( r = \sigma \Delta t/\Delta x^2 \). This scheme is \( O(\Delta t) + O(\Delta x^2) \). The linear system that results from (5) can be represented by

   \[ U^{n+1} = L^F U^n \]  (6)

2. **Backward in time–Centered in space**

   \[ -rU_{i-1}^{n+1} + (1 + 2r)U_i^{n+1} - rU_{i+1}^{n+1} = U_i^n, \quad i = 1, \ldots, m \]  (7)

   This scheme is \( O(\Delta t) + O(\Delta x^2) \) The linear system that results from (7) can be represented by

   \[ L^B U^{n+1} = U^n \]  (8)

3. **Crank–Nicholson**

   \[ \frac{-r}{2} U_{i-1}^{n+1} + (1 + r)U_i^{n+1} - \frac{r}{2} U_{i+1}^{n+1} = \frac{r}{2} U_{i-1}^n + (1 - r)U_i^n + \frac{r}{2} U_{i+1}^n, \quad i = 1, \ldots, m \]  (9)

   This scheme is \( O(\Delta t^2) + O(\Delta x^2) \) The linear system that results from (9) can be represented by

   \[ L^S U^{n+1} = L^G U^n + C^n \]  (10)
0.1 Definition 1: Stability of Linear Finite Difference Methods

A linear finite difference method (FDM) of the form

$$U^{n+1} = L_{\Delta} U^{n}$$

(11)

corresponding to an IBVP of (4) (such as (1)-(3)) is stable if there exists $C > 0$, independent of the mesh spacing and the initial data, such that

$$||U^{n}|| \leq C ||U^{0}||, \quad n \to \infty, \quad \Delta t \to 0, \quad \Delta x \to 0, \quad n\Delta t \leq T$$

(12)

0.2 Theorem 1: Equivalent Condition

The FDM (11) is stable if and only if there exists a constant $C > 0$ independent of $\Delta x$ and $\Delta t$ such that

$$||L_{\Delta}^{n}|| \leq C, \quad n \to \infty, \quad \Delta t \to 0, \quad \Delta x \to 0, \quad n\Delta t \leq T$$

(13)

Remark: Notice that $C$ may be greater than 1.

Proof.

Notice that

$$U^{n} = L_{\Delta} U^{n-1} = L_{\Delta} (L_{\Delta} U^{n-2}) = L_{\Delta}^{2} U^{n-2} = \cdots = L_{\Delta}^{n} U^{0}$$

Therefore, for $U^{0} \neq 0$

$$||U^{n}|| \leq C ||U^{0}|| \iff ||L_{\Delta}^{n} U^{0}|| \leq C ||U^{0}|| \iff \frac{||L_{\Delta}^{n} U^{0}||}{||U^{0}||} \leq C \iff ||(L_{\Delta})^{n}|| \leq C$$

(14)

0.3 Corollary 1: Practical Condition

If the discrete operator $L_{\Delta}$ of the FDM (11) satisfies

$$||L_{\Delta}|| \leq 1,$$

then the FDM (11) is stable.

Proof.

Notice that $||L_{\Delta}^{n}|| \leq ||L_{\Delta}||^{n}$. Therefore, if

$$||L_{\Delta}|| \leq 1 \Rightarrow ||(L_{\Delta})^{n}|| \leq ||L_{\Delta}||^{n} \leq 1$$
The stability follows from Theorem 1.

Remark: Apply this condition to the explicit FDM FT-CS using the infinity norm.

0.4 Corollary 2: More General Sufficient Condition

If there is a \( c > 0 \) independent of \( \Delta x \) and \( \Delta t \) such that the discrete operator \( L_\Delta \) of the FDM (11) satisfies

\[
||L_\Delta|| \leq 1 + c\Delta t,
\]

for \( \Delta t < \Delta t^* \), then the FDM (11) is stable.

Proof. Notice that \( n\Delta t \leq T \) and \( 1 + c\Delta t \leq e^{c\Delta t} \), then \( 1 + c\Delta t \leq e^{cT/n} \). Therefore,

\[
|| (L_\Delta)^n || \leq ||L_\Delta||^n \leq (1 + c\Delta t)^n \leq e^{cT} = e^\tilde{c} = C
\]

0.5 Definition 2: Spectral Radius

The spectral radius \( \rho(L_\Delta) \) of the FDM matrix \( L_\Delta \) is the absolute value of its largest eigenvalue. Assuming that \( \lambda_i, i = 1, \ldots, N \) are the eigenvalues of \( L_\Delta \), then

\[
\rho(L_\Delta) = \max_{1 \leq i \leq N} |\lambda_i|
\]

0.6 Theorem 2: Relationship Between Spectral Radius and Norm of \( L_\Delta \)

If \( \rho(L_\Delta) \) and \( L_\Delta \) are the spectral radius and the vector-induced norm of \( L_\Delta \) then,

\[
\rho(L_\Delta) \leq ||L_\Delta||
\]

Proof. For any eigenvector \( x_i \), it holds \( ||L_\Delta x_i|| = ||\lambda_i x_i|| \), for \( i = 1, 2, \ldots, N \). Therefore,

\[
|\lambda_i| = \frac{||L_\Delta x_i||}{||x_i||} \leq \max_{x \neq 0} \frac{||L_\Delta x||}{||x||} \leq ||L_\Delta|| \Rightarrow \rho(L_\Delta) \leq ||L_\Delta||
0.7 Corollary 3: Necessary Condition

The condition

\[ \rho^n(L_\Delta) \leq C, \]

for a constant \( C > 0 \) independent of \( \Delta x \) and \( \Delta t \) is a necessary condition for the stability of the FDM (11).

Proof.
Notice that \( \rho^n(L_\Delta) = \rho((L_\Delta)^n) \leq \|(L_\Delta)^n\| \). Therefore, if \( \rho^n(L_\Delta) \) is not bounded then \( \|(L_\Delta)^n\| \) is also not bounded and the FDM is not stable.

0.8 Corollary 4: A More Practical Condition (special matrices)

If \( L_\Delta \) of the FDM (11) is symmetric or similar to a symmetric matrix, then

\[ \rho(L_\Delta) \leq 1, \]

for any \( \Delta x \) and \( \Delta t \), is also a sufficient condition for stability in the Euclidean norm.

Proof.
If \( L_\Delta \) is a symmetric matrix then the Euclidean norm \( \|L_\Delta\|_2 = \sqrt{\rho(L_\Delta L_\Delta^T)} = \rho(L_\Delta) \).
Therefore,

\[ \rho(L_\Delta) \leq 1 \Rightarrow \|L_\Delta\|_2 \leq 1 \]

and the stability follows from Corollary 1.

Remark: Apply this condition to show stability of FT-CS and BT-CS FDM for IBVP (1)-(3) with homogeneous boundary conditions.

0.9 Definition 4: Convergence

A finite difference approximation \( U^n \) converges to the solution \( u^n \) (the restriction of the exact solution \( u(x,t_n) \) to the mesh) on \( 0 < t \leq T \) in a particular vector norm if

\[ \|u^n - U^n\| \to 0, \quad n \to \infty, \quad \Delta x \to 0, \quad \Delta t \to 0, \quad n\Delta t \leq T \]  \( (15) \)

\textit{Why do we want to prove stability for FDM such as (11) approximating certain PDE problems modelled by (4)?} The answer to this question is found in the next theorem
0.10 Theorem 3: Lax-Equivalence Theorem

A consistent linear FDM such as (11) is convergent if and only if it is stable.

In many problems of practical interest, we would like to study stability when $t \to \infty$. To analyze stability for these problems, we need an alternative stability definition.

0.11 Definition 3: Absolute Stability

A FDM such as (11) is absolutely stable for a given mesh (of size $\Delta x$ and $\Delta t$) if

\[ ||U^n|| \leq ||U^0||, \quad n > 0 \]  \hspace{1cm} (16)

0.12 Definition 4: Unconditional Stability

A FDM such as (11) is unconditionally stable if it is absolutely stable for all choices of mesh spacing $\Delta x$ and $\Delta t$. 
LAX Equivalence Theorem.

Definition: The IVP for the first-order (in time) PDE
\[ u_t = L u \] (L differential operator) is well-posed if for any time \( T > 0 \), there is a constant \( C_T \) such that any solution \( U(x,t) \) satisfies
\[
\int_0^\infty |U(x,t)|^2 \, dx \leq C_T \int_0^\infty |U(x,0)|^2 \, dx.
\]
for \( 0 \leq t \leq T \).

Theorem: A consistent finite difference scheme for a PDE for which the IVP is well-posed is convergent if and only if it is stable.

Proof: \( \leftarrow \) Stability \Rightarrow \) Convergence.

Consider the numerical scheme
\[
\tilde{U}^{n+1} = L_A \tilde{U}^n
\] (1)
for example, FT-CS heat conduction
\[
\begin{bmatrix}
\tilde{U}_1^{n+1} \\
\tilde{U}_2^{n+1} \\
\vdots \\
\tilde{U}_{j-1}^{n+1}
\end{bmatrix} =
\begin{bmatrix}
1 - 2r & r & 0 & \cdots & 0 \\
r & 1 - 2r & r & \cdots & 0 \\
0 & r & 1 - 2r & r & \cdots \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & 1 - 2r
\end{bmatrix}
\begin{bmatrix}
\tilde{U}_1^n \\
\tilde{U}_2^n \\
\vdots \\
\tilde{U}_{j-1}^n
\end{bmatrix}
\]
Proof.- (Lax Equivalence theorem).

If \( U^n \) is a solution of the partial diff. eqn.
then
\[
\dot{U}^{n+1} = L_\Delta \dot{U}^n + \Delta t \mathbf{\tilde{r}}^n
\]

For example, Heat cond.

\[
\dot{U}^{n+1}_j = \Delta U^n_j - 2 \Delta \dot{U}^n_j + \Delta \ddot{U}^n_j + \Delta t \mathbf{\tilde{r}}^n_j
\]

where
\[
\mathbf{\tilde{r}}^n_j = -\frac{\Delta t}{2} (U^n_{j+1})^{n+1} + \frac{\Delta x^2}{12} (U^n_{j+3} - 2U^n_j + U^n_{j-1})
\]

or
\[
\dot{U}^{n+1} = L_\Delta \dot{U}^n + \Delta t \mathbf{\tilde{r}}^n. \tag{2}
\]

The difference of the vector solution \( \tilde{U}^n \) of the PDE.
And the vector solution \( \tilde{U}^n \) of the discrete approx.

is called \( \tilde{e}^n = \tilde{U}^n - \tilde{U}^n \) (global discretization error).

Subtracting (1) from (2)

\[
\dot{\tilde{e}}^{n+1} = \dot{U}^{n+1} - \dot{\tilde{U}}^{n+1} = L_\Delta (\tilde{U}^n - \tilde{U}^n) + \Delta t \mathbf{\tilde{r}}^n = L_\Delta \tilde{e}^n + \Delta t \mathbf{\tilde{r}}^n.
\]

\[
\Rightarrow \quad \dot{\tilde{e}}^{n+1} = L_\Delta \dot{\tilde{e}}^n + \Delta t \mathbf{\tilde{r}}^n. \tag{3}
\]
If $L_\Delta$ is independent of $n$ by iterating on (3)

$$\hat{\varepsilon}^n = L_\Delta \hat{\varepsilon}^{n-1} + \Delta t \hat{\tau}^{n-1} = L_\Delta (L_\Delta \hat{\varepsilon}^{n-2} + \Delta t \hat{\tau}^{n-2}) + \Delta t \hat{\tau}^{n-1}$$

$$= L_\Delta^2 \hat{\varepsilon}^{n-2} + \Delta t [L_\Delta \hat{\tau}^{n-2} + \hat{\tau}^{n-1}] =$$

$$= L_\Delta^3 \hat{\varepsilon}^{n-3} + \Delta t [L_\Delta^2 \hat{\tau}^{n-3} + L_\Delta \hat{\tau}^{n-2} + \hat{\tau}^{n-1}] =$$

$$\cdots \cdots = L_\Delta^n \hat{\varepsilon}^0 + \Delta t \left[ L_\Delta^{n-1} \hat{\tau}^0 + L_\Delta^{n-2} \hat{\tau}^1 + \cdots + L_\Delta \hat{\tau}^{n-1} \right]$$

or

$$\hat{\varepsilon}^n = L_\Delta^n \hat{\varepsilon}^0 + \Delta t \left[ L_\Delta^{n-1} \hat{\tau}^0 + L_\Delta^{n-2} \hat{\tau}^1 + \cdots + \hat{\tau}^{n-1} \right]$$

$$\Rightarrow \|\hat{\varepsilon}^n\| \leq \Delta t \left[ \| L_\Delta^{n-1} \| \| \hat{\tau}^0 \| + \| L_\Delta^{n-2} \| \| \hat{\tau}^1 \| + \cdots + \| \hat{\tau}^{n-1} \| \right]$$ (4)

Let's choose a time of interest $T$ and an arbitrary $\varepsilon > 0$. Since numerical scheme is consistent there exists $S > 0$ such that $\|\hat{\tau}^k\| \leq \varepsilon$ if $\Delta t, \Delta x \leq S_1$ for all $k$ such that $k \Delta t \leq T$. 
Also, using the hypothesis that the scheme is stable and the previous theorem about stability we conclude that there exist $C$ and $\delta > 0$ such that

$$\text{if } \Delta x, \Delta t < \delta \Rightarrow \|(la)^k\| \leq C \quad \text{for all } k \text{ such that } k\Delta t \leq T.$$ 

Therefore, choosing $S = \min (\delta, \delta)$, for $\Delta x, \Delta t < S$

(4) reduces to

$$\|\tilde{e}^n\| \leq \Delta t \left( (n-1)\epsilon + \epsilon \right) \begin{cases} \leq \Delta t n \epsilon \epsilon, & \text{if } C \geq 1. \\ \leq \Delta t n \epsilon, & \text{if } C < 1 \end{cases}$$

Since $n \Delta t \leq T$

$$\Rightarrow \|\tilde{e}^n\| \leq \begin{cases} T \epsilon, & \text{if } C \geq 1 \text{ or } C < 1. \\ T \epsilon, & \text{if } C \geq 1. \end{cases}$$

In both cases, the scheme converges.