Linear ODE

Let $I \subset \mathbb{R}$ be an interval (open or closed, finite or infinite — at either end). Suppose $A : I \to \mathbb{F}^{n\times n}$ and $b : I \to \mathbb{F}^n$ are continuous. The DE

$$(*) \qquad x' = A(t)x + b(t)$$

is called a first-order linear [system of] ODE[s] on I. Since $f(t,x) \equiv A(t)x+b(t)$ is continuous in t, x on $I \times \mathbb{F}^n$ and, for any compact subinterval $[c,d] \subset I$, f is uniformly Lipschitz in x on $[c,d] \times \mathbb{F}^n$ (with Lipschitz constant $\max_{c\le t\le d} |A(t)|$), we have global existence and uniqueness of solutions of the IVP

$$x' = A(t)x + b(t), \qquad x(t_0) = x_0$$

on all of I (where $t_0 \in I$, $x_0 \in \mathbb{F}^n$).

If $b \equiv 0$ on I, $(*)$ is called a linear homogeneous system (LH).

If $b \not\equiv 0$ on I, $(*)$ is called a linear inhomogeneous system (LI).

Fundamental Theorem for LH *The set of all solutions of (LH) $x' = A(t)x$ on I form an n-dimensional vector space over $\mathbb{F}$ (in fact, a subspace of $C^1(I,\mathbb{F}^n)$).*

Proof. Clearly $x_1' = Ax_1$ and $x_2' = Ax_2$ imply $(c_1x_1 + c_2x_2)' = A(c_1x_1 + c_2x_2)$, so the set of solutions of (LH) form a vector space over $\mathbb{F}$, which is clearly a subspace of $C^1(I,\mathbb{F}^n)$. Fix $\tau \in I$, and let $y_1, \dots, y_n$ be a basis for $\mathbb{F}^n$. For $1 \le j \le n$, let $x_j(t)$ be the solution of the IVP $x' = Ax$, $x(\tau) = y_j$. Then $x_1(t), \dots, x_n(t)$ are linearly independent in $C^1(I,\mathbb{F}^n)$; indeed,

$$\sum_{j=1}^n c_j x_j(t) = 0 \quad \text{in} \quad C^1(I,\mathbb{F}^n)$$

$$\Rightarrow$$

$$\sum_{j=1}^n c_j x_j(t) = 0 \quad \forall\, t \in I$$

$$\Rightarrow$$

$$\sum_{j=1}^n c_j y_j = \sum c_j x_j(\tau) = 0$$

$$\Rightarrow$$

$$c_j = 0 \quad j = 1, 2, \dots, n.$$

Now if $x(t)$ is any solution of (LH), there exist unique $c_1, \dots, c_n$ such that $x(\tau) = c_1y_1 + \cdots + c_ny_n$. Clearly $c_1x_1(t) + \cdots + c_nx_n(t)$ is *a* solution of the IVP

$$x' = A(t)x, \quad x(\tau) = c_1y_1 + \cdots c_ny_n,$$

so by uniqueness, $x(t) = c_1x_1(t) + \cdots + c_nx_n(t)$ for all $t \in I$. Thus $x_1(t), \dots, x_n(t)$ span the vector space of all solutions of (LH) on I. So they form a basis for the vector space of solutions to (LH), and the dimension of this vector space is n. $\square$

Remark. Define the linear operator $L : C^1(I, \mathbb{F}^n) \to C^1(I, \mathbb{F}^n)$ by $Lx = \left(\frac{d}{dt} - A(t)\right) x$, i.e., $[Lx](t) = x'(t) - A(t)x(t)$ for $x(t) \in C^1(I, \mathbb{F}^n)$. L is called a *linear differential operator*. The solution space in the previous theorem is precisely the null space of L. Thus the null space of L is finite dimensional and has dimension n.

Definition. If $\varphi_1, \ldots, \varphi_n$ are n linearly independent (as elements of $C^1(I, \mathbb{F}^n)$) solutions of (LH) $x' = Ax$, then they form a basis for the vector space of solutions to (LH). Such a basis is called a *fundamental set* of solutions of (LH). If $\Phi : I \to \mathbb{F}^{n \times n}$ is an $n \times n$ matrix function of $t \in I$ whose columns form a fundamental set of solutions of (LH), then $\Phi(t)$ is called a *fundamental matrix* for (LH) $x' = A(t)x$, in which case

$$\Phi'(t) = A(t)\Phi(t).$$

Definition. If $X : I \to \mathbb{F}^{n \times k}$ is in $C^1(I, \mathbb{F}^{n+k})$, we say that X is an $[n \times k]$ matrix solution of (LH) if $X'(t) = A(t)X(t)$. Clearly $X(t)$ is a matrix solution of (LH) if and only if each column of $X(t)$ is a solution of (LH). (We will mostly be interested in the case $k = n$.)

Theorem. Let $A : I \to \mathbb{F}^{n \times n}$ be continuous where $I \subset \mathbb{R}$ is an interval, and suppose $X : I \to \mathbb{F}^{n \times n}$ is an $n \times n$ matrix solution of (LH) $x' = A(t)x$ on I, i.e., $X'(t) = A(t)X(t)$ on I. Then

$$\det\,(X(t))' = \operatorname{tr}\,(A(t))(\det X(t)),$$

and so for all $\tau, t \in I$,

$$\det X(t) = (\det\,(X(\tau))) \exp \int_\tau^t \operatorname{tr}\,(A(s))ds.$$

Proof Sketch. Let $x_{ij}(t)$ denote the ij^{th} element of $X(t)$, and let $\hat{X}_{ij}(t)$ denote the $(n-1) \times (n-1)$ matrix obtained from $X(t)$ by deleting its ith row and jth column. The co-factor representation of the determinant gives

$$\det\,(X(t)) = \sum_{j=1}^{n} (-1)^{(i+j)} x_{ij}(t) \det\,(\hat{X}_{ij}(t)), \quad i = 1, 2, \ldots, n\ .$$

Hence

$$\frac{d}{dx_{ij}} \det\,(X(t)) = (-1)^{(i+j)} \det\,(\hat{X}_{ij}),$$

and so by the chain rule

$$(\det X(t))'$$

$$= \sum_{j=1}^{n} (-1)^{(1+j)} x'_{1j}(t) \det\,(\hat{X}_{ij}(t)) + \cdots + \sum_{j=1}^{n} (-1)^{(n+j)} x'_{nj}(t) \det\,(\hat{X}_{ij}(t)) =$$

$$\det \begin{bmatrix} x'_{11} & x'_{12} & \cdots & x'_{1n} \\ & (\text{remaining } x_{ij}) & & \end{bmatrix} + \det \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x'_{21} & x'_{22} & \cdots & x'_{2n} \\ & (\text{remaining } x_{ij}) & & \end{bmatrix} + \cdots + \det \begin{bmatrix} & (\text{remaining } x_{ij}) & \\ x'_{n1} & x'_{n2} & \cdots & x'_{nn} \end{bmatrix}.$$

Now by (LH)

$$\begin{aligned}\begin{bmatrix} x_{11}' & x_{12}' & \cdots & x_{1n}' \end{bmatrix} &= [\Sigma_k a_{1k}x_{k1} \cdots \Sigma_k a_{1k}x_{kn}] \\ &= a_{11}[x_{11}\cdots x_{1n}] + a_{12}[x_{21}\cdots x_{2n}] + \cdots + a_{1n}[x_{n1}\cdots x_{nn}].\end{aligned}$$

Subtracting $a_{12}[x_{21}\cdots x_{2n}] + \cdots + a_{1n}[x_{n1}\cdots x_{nn}]$ from the first row of the matrix in the first determinant on the RHS doesn't change that determinant. A similar argument applied to the other determinants gives

$$(\det X(t))' =$$

$$\det \begin{bmatrix} a_{11}[x_{11}\cdots x_{1n}] \\ (\text{remaining } x_{ij}) \end{bmatrix} + \det \begin{bmatrix} x_{11}\cdots x_{1n} \\ a_{22}[x_{21}\cdots x_{2n}] \\ (\text{remaining } x_{ij}) \end{bmatrix} + \cdots + \det \begin{bmatrix} (\text{remaining } x_{ij}) \\ a_{nn}[x_{n1}\cdots x_{nn}] \end{bmatrix}$$

$$= (a_{11} + a_{22} + \cdots + a_{nn}) \det X(t) = \operatorname{tr}(A(t))(\det X(t)).$$

□

Corollary. Let $X(t)$ be an $n \times n$ matrix solution of (LH) $x' = A(t)x$. Then either

$$(\forall\, t \in I) \quad \det X(t) \neq 0 \quad \text{or} \quad (\forall\, t \in I) \quad \det X(t) = 0.$$

Corollary. Let $X(t)$ be an $n \times n$ matrix solution of (LH) $x' = A(t)x$. Then the following statements are equivalent.

(1) $X(t)$ is a fundamental matrix for (LH) on I.

(2) $(\exists\, \tau \in I)$ $\det X(\tau) \neq 0$ (i.e., columns of X are linearly independent at τ)

(3) $(\forall\, t \in I)$ $\det X(t) \neq 0$ (i.e., columns of X are linearly independent at every $t \in I$).

Definition. If $X(t)$ is an $n \times n$ matrix solution of (LH) $x' = A(t)x$, then $\det(X(t))$ is often called the Wronskian [of the columns of $X(t)$].

Remark. This is not quite standard notation for general LH systems $x' = A(t)x$. It is used most commonly when $x' = A(t)x$ is the 1^{st}-order system equivalent to a scalar n^{th}-order linear homogeneous ODE.

Theorem. Suppose $\Phi(t)$ is a fundamental matrix for (LH) $x' = A(t)x$ on I.

(a) If $c \in \mathbb{F}^n$, then $x(t) = \Phi(t)c$ is a solution of (LH) on I.

(b) If $x(t) \in C^1(I, \mathbb{F}^n)$ is any solution of (LH) on I, then there exists a unique $c \in \mathbb{F}^n$ for which $x(t) = \Phi(t)c$.

Proof. The theorem just restates that the columns of $\Phi(t)$ form a basis for the set of solutions of (LH). □

Theorem. Suppose $\Phi(t)$ is a fundamental matrix (F.M.) for (LH) $x' = A(t)x$ on I.

(a) If $C \in \mathbb{F}^{n\times n}$ is invertible, then $X(t) = \Phi(t)C$ is also a F.M. for (LH) on I.

(b) If $X(t) \in C^1(I, \mathbb{F}^{n\times n})$ is any F.M. for (LH), then there exists a unique invertible $C \in \mathbb{F}^{n\times n}$ for which $X(t) = \Phi(t)C$.

Proof. For (a), observe that

$$X'(t) = \Phi'(t)C = A(t)\Phi(t)C = A(t)X(t),$$

so $X(t)$ is a matrix solution, and $\det X(t) = (\det \Phi(t))(\det C) \neq 0$.

For (b), set $\Psi(t) = \Phi(t)^{-1}X(t)$. Then $X = \Phi\Psi$, so

$$\Phi'\Psi + \Phi\Psi' = (\Phi\Psi)' = X' = AX = A\Phi\Psi = \Phi'\Psi,$$

which implies that $\Phi\Psi' = 0$. Since $\Phi(t)$ is invertible for all $t \in I$, $\Psi'(t) \equiv 0$ on I. So $\Psi(t)$ is a constant invertible matrix C. Since $C = \Psi = \Phi^{-1}X$, we have $X(t) = \Phi(t)C$. □

Remark. If $B(t) \in C^1(I, \mathbb{F}^{n\times n})$ is invertible for each $t \in I$, then

$$\frac{d}{dt}(B^{-1}(t)) = -B^{-1}(t)B'(t)B^{-1}(t).$$

Proof. $0 = \frac{d}{dt}(I) = \frac{d}{dt}(B(t)B^{-1}(t)) = B(t)\frac{d}{dt}(B^{-1}(t)) + B'(t)B^{-1}(t)$. □

Adjoint Systems

Let $\Phi(t)$ be a F.M. for (LH) $x' = A(t)x$. Then

$$(\Phi^{-1})' = -\Phi^{-1}\Phi'\Phi^{-1} = -\Phi^{-1}A\Phi\Phi^{-1} = -\Phi^{-1}A.$$

Taking conjugate transposes, $(\Phi^{-H})' = -A^H\Phi^{-H}$. So $\Phi^{-H}(t)$ is a F.M. for the *adjoint system* (LH*) $x' = -A^H(t)x$.

Theorem. If $\Phi(t)$ is a F.M. for (LH) $x' = A(t)x$ and $\Psi(t) \in C^1(I, \mathbb{F}^{n\times n})$, then $\Psi(t)$ is a F.M. for (LH*) $x' = -A^H(t)x$ if and only if $\Psi^H(t)\Phi(t) = C$, where C is a constant invertible matrix.

Proof. Suppose $\Psi(t)$ is a F.M. for (LH*). Since $\Phi^{-H}(t)$ is also a F.M. for (LH*), $\exists$ an invertible $C \in \mathbb{F}^{n\times n} \ni \Psi(t) = \Phi^{-H}(t)C^H$, i.e., $\Psi^H = C\Phi^{-1}$, $\Psi^H\Phi = C$. Conversely, if $\Psi^H(t)\Phi(t) = C$ (invertible), then $\Psi^H = C\Phi^{-1}$, $\Psi = \Phi^{-H}C$, so Ψ is a F.M. for (LH*). □

Normalized Fundamental Matrices

Definition. A F.M. $\Phi(t)$ for (LH) $x' = A(t)x$ is called *normalized at time* τ if $\Phi(\tau) = I$, the identity matrix. (Convention: if not stated otherwise, a normalized F.M. usually means normalized at time $\tau = 0$.)

Facts.

(1) For a given τ, the F.M. of (LH) normalized at τ exists and is unique. (**Proof**. The j^{th} column of $\Phi(t)$ is the solution of the IVP $x' = A(t)x$, $x(\tau) = e_j$.)

(2) If $\Phi(t)$ is the F.M. for (LH) normalized at τ, then the solution of the IVP $x' = A(t)x$, $x(\tau) = y$ is $x(t) = \Phi(t)y$. (**Proof**. $x(t) = \Phi(y)$ satisfies (LH) $x' = A(t)x$, and $x(\tau) = \Phi(\tau)y = Iy = y$.)

(3) For any fixed τ, t, the solution operator S_τ^t for (LH), mapping $x(\tau)$ into $x(t)$, is *linear* on $\mathbb{F}^n$, and its matrix is the F.M. $\Phi(t)$ for (LH) normalized at τ, evaluated at t.

(4) If $\Phi(t)$ is *any* F.M. for (LH), then for fixed τ, $\Phi(t)\Phi^{-1}(\tau)$ is the F.M. for (LH) normalized at τ. (**Proof**. It is a F.M. taking the value I at τ.) Thus (a) $\Phi(t)\Phi^{-1}(\tau)$ is the matrix of the solution operator S_τ^t for (LH); and (b) the solution of the IVP $x' = A(t)x$, $x(\tau) = y$ is $x(t) = \Phi(t)\Phi^{-1}(\tau)y$.

Inhomogeneous Linear Systems

We now want to express the solution of the IVP

$$x' = A(t)x + b(t), \quad x(t_0) = y$$

for the linear inhomogeneous system

$$\text{(LI)} \qquad x' = A(t)x + b(t)$$

in terms of a F.M. for the associated homogeneous system

$$\text{(LH)} \qquad x' = A(t)x.$$

Variation of Parameters

Let $\Phi(t)$ be any F.M. for (LH). Then, for any constant vector $c \in \mathbb{F}^n$ $\Phi(t)c$ is a solution of (LH). We will look for a solution of (LI) of the form

$$x(t) = \Phi(t)c(t)$$

(varying the "constants" — elements of c). Plugging into (LI), we want

$$(\Phi c)' = A\Phi c + b,$$

or equivalently

$$\Phi' c + \Phi c' = A\Phi c + b.$$

Since $\Phi' = A\Phi$, this gives $\Phi c' = b$, or $c' = \Phi^{-1} b$. Set

$$c(t) = c_0 + \int_{t_0}^{t} \Phi^{-1}(s) b(s) ds$$

for some constant vector $c_0 \in \mathbb{F}^n$, and let $x(t) = \Phi(t)c(t)$. These calculations show that $x(t)$ is a solution of (LI). To satisfy the initial condition $x(t_0) = y$, we take $c_0 = \Phi^{-1}(t_0)y$, and obtain

$$x(t) = \Phi(t)\Phi^{-1}(t_0)y + \int_{t_0}^{t} \Phi(t)\Phi^{-1}(s)b(s)ds.$$

In words, this equation states that

$$\left\{ \begin{array}{c} \text{soln of (LI)} \\ \text{with I.C. } x(t_0) = y \end{array} \right\} = \left\{ \begin{array}{c} \text{soln of (LH)} \\ \text{with I.C. } x(t_0) = y \end{array} \right\} + \left\{ \begin{array}{c} \text{soln of (LI)} \\ \text{with homog. I.C. } x(t_0) = 0 \end{array} \right\}.$$

Viewing y as arbitrary, we find that the *general solution of (LI)* equals the *general solution of (LH)* plus a *particular solution of (LI)* stated in terms of the solution operator.

Note: $\Phi(t)\Phi^{-1}(t_0)$ is the matrix of $S^t_{t_0}$, and $\Phi(t)\Phi^{-1}(s)$ is the matrix of S^t_s).

Duhamel's Principle. If S^t_τ is the solution operator for (LH), then the solution of the IVP $x' = A(t)x + b(t)$, $x(t_0) = y$ is $x(t) = S^t_{t_0} y + \int_{t_0}^{t} S^t_s(b(s))ds$.

Remark. So the effect of the inhomogeneous term $b(t)$ in (LI) is like adding additional IC at each time $s \in [t_0, t]$, integrating these solutions $S^t_s(b(s))$ of (LH) with respect to $s \in [t_0, t]$.

Constant Coefficient Systems

Consider the linear homogeneous constant-coefficient first-order system

$$\text{(LHC)} \qquad x' = Ax,$$

where $A \in \mathbb{F}^{n\times n}$ is a constant matrix. The F.M. of (LHC), normalized at 0, is $\Phi(t) = e^{tA}$. Recall that

$$e^B \equiv \sum_{j=0}^{\infty} \frac{1}{j!} B^j$$

where $B^0 \equiv I$, so $\Phi(0) = I$. Term by term differentiation is justified in the series for e^{tA} :

$$\begin{aligned} \Phi'(t) &= \frac{d}{dt}(e^{tA}) = \sum_{j=0}^{\infty} \frac{1}{j!} \frac{d}{dt}(tA)^j \\ &= \sum_{j=1}^{\infty} \frac{1}{(j-1)!} t^{j-1} A^j = A \sum_{k=0}^{\infty} \frac{1}{k!}(tA)^k = Ae^{tA}. \end{aligned}$$

We can express e^{tA} using the Jordan form of A: if $P^{-1}AP = J$ is in Jordan form where $P \in \mathbb{F}^{n\times n}$ is invertible (assume $\mathbb{F} = \mathbb{C}$ if A has any nonreal eigenvalues), then $A = PJP^{-1}$, so $e^{tA} = e^{tPJP^{-1}} = Pe^{tJ}P^{-1}$. If

$$J = \begin{bmatrix} J_1 & & & \bigcirc \\ & J_2 & & \\ & & \ddots & \\ \bigcirc & & & J_s \end{bmatrix}$$

where each J_k is a single Jordan block, then

$$e^{tJ} = \begin{bmatrix} e^{tJ_1} & & & \bigcirc \\ & e^{tJ_2} & & \\ & & \ddots & \\ \bigcirc & & & e^{tJ_s} \end{bmatrix}.$$

Finally, if

$$J_k = \begin{bmatrix} \lambda & 1 & & \bigcirc \\ & \lambda & \ddots & \\ & & \ddots & 1 \\ \bigcirc & & & \lambda \end{bmatrix}$$

is $l \times l$, then

$$e^{tJ_k} = \begin{bmatrix} 1 & t & \frac{t^2}{2!} & \cdots & \frac{t^{l-1}}{(l-1)!} \\ & 1 & t & \ddots & \vdots \\ & & \ddots & \ddots & \frac{t}{2!} \\ & & & & t \\ \bigcirc & & & & 1 \end{bmatrix}.$$

The solution of the inhomogeneous IVP $x' = Ax + b(t)$, $x(t_0) = y$ is

$$x(t) = e^{(t-t_0)A}y + \int_{t_0}^{t} e^{(t-s)A}b(s)ds$$

since $(e^{tA})^{-1} = e^{-tA}$ and $e^{tA}e^{-sA} = e^{(t-s)A}$.

Another viewpoint

Suppose $A \in \mathbb{C}^{n\times n}$ is a constant diagonalizable matrix with eigenvalues $\lambda_1, \ldots, \lambda_n$ and linearly independent eigenvectors $v_1, \ldots, v_n$. Then $\varphi_j(t) \equiv e^{\lambda, t}v_j$ is a solution of (LHC) $x' = Ax$ since

$$\begin{aligned} \varphi_j' &= \frac{d}{dt}(e^{\lambda_j t}v_j) = \lambda_j e^{\lambda_j t}v_j = e^{\lambda_j t}(\lambda v_j) \\ &= e^{\lambda_j t}Av_j = A(e^{\lambda_j t}v_j) = A\varphi_j. \end{aligned}$$

Clearly $\varphi_1, \ldots, \varphi_n$ are linearly independent at $t = 0$ as $\varphi_j(0) = v_j$. Thus

$$\Phi(t) = [\varphi_1(t)\varphi_2(t)\cdots\varphi_n(t)]$$

is a F.M. for (LHC). So the general solution of (LHC) (for diagonalizable A) is $\Phi(t)c = c_1 e^{\lambda_1 t}v_1 + \cdots + c_n e^{\lambda_n t}v_n$ for arbitrary scalars $c_1, \ldots, c_n$.

Remark on Exponentials

Let $B(t)$ be a C^1 $n \times n$ matrix function of t, and let $A(t) = B'(t)$. Then

$$\begin{aligned}\frac{d}{dt}(e^{B(t)}) &= \frac{d}{dt}(I + B + \frac{1}{2!}B\cdot B + \frac{1}{3!}B\cdot B\cdot B + \cdots)\\ &= A + \frac{1}{2!}(AB + BA) + \frac{1}{3!}(AB^2 + BAB + B^2A) + \cdots.\end{aligned}$$

Now, if for each t, $A(t)$ and $B(t)$ commute, then

$$\frac{d}{dt}(e^{B(t)}) = A\left(I + B + \frac{1}{2!}B^2 + \cdots\right) = B'(t)e^{B(t)}.$$

Now suppose we start with a continuous $n \times n$ matrix function $A(t)$, and for some t_0, we define $B(t) = \int_{t_0}^t A(s)ds$, so $B'(t) = A(t)$. Suppose in addition that $A(t)$ and $B(t)$ commute for all t. Then $\Phi(t) \equiv \exp\left(\int_{t_0}^t A(s)ds\right)$ is the F.M. for (LH) $x' = A(t)x$, normalized at t_0, since $\Phi(t_0) = I$ and $\Phi'(t) = A(t)\Phi(t)$ as above. A sufficient (but not necessary) condition guaranteeing that $A(t)$ and $\int_{t_0}^t A(s)ds$ commute is that $A(t)$ and $A(s)$ commute for all t, s.

*Remark.***Reduction of Order for (LH)** $x' = A(t)x$

In Coddington & Levinson it is shown that if m $(< n)$ linearly independent solutions of the $n \times n$ linear homogeneous system $x' = A(t)x$ are known, then one can derive an $(n-m) \times (n-m)$ system for obtaining $n - m$ more linearly independent solutions.

Example. $D_y S_\tau^t$ is Invertible at each y.

In this example we show how one can apply the theory of linear systems to the nonlinear solution operator. Consider the DE $x' = f(t, x)$ where f is C^1, and let S_τ^t denote the solution operator. For a fixed τ, let $x(t, y)$ denote the solution of the IVP $x' = f(t, x)$, $x(\tau) = y$. The equation of variation for the $n \times n$ Jacobian matrix $D_y x(t, y)$ is

$$\frac{d}{dt}\left(D_y x(t,y)\right) = \left(D_x f\left(t, x(t,y)\right)\right)\left(D_y x(t,y)\right),$$

and thus

$$\frac{d}{dt}\left(\det\left(D_y x(t,y)\right)\right) = \operatorname{tr}\left(D_x f\left(t, x(t,y)\right)\right)\det\left(D_y x(t,y)\right),$$

so

$$\begin{aligned}\det\left(D_y x(t,y)\right) &= \det\left(D_y x(\tau,y)\right)\exp\left(\int_\tau^t \operatorname{tr}\left(D_x f\left(s, x(s,y)\right)\right)\right)ds\\ &= \exp\left(\int_\tau^t \operatorname{tr}\left(D_x f\left(s, x(s,y)\right)\right)\right)ds,\end{aligned}$$

since

$$D_y x(\tau, y) = D_y y = I.$$

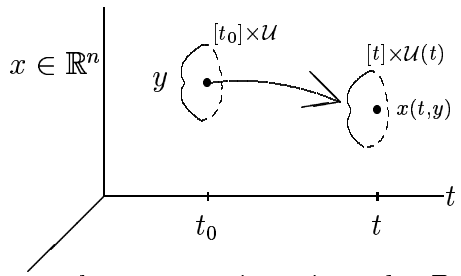
In particular, $\det\left(D_y x(t,y)\right) \neq 0$, so $D_y x(t,y)$ is invertible. For τ and t fixed, $D_y x(t,y) = D_y S_\tau^t$, so we have demonstrated again that $D_y S_\tau^t$ is invertible at each y.

Example. The Rate of Change of Volume in a Flow.

Consider an autonomous system $x' = f(x)$, where f is C^1 and $\mathbb{F} = \mathbb{R}$, so $x \in \mathbb{R}^n$. Fix t_0, and view the family of IVPs

$$x' = f(x), \quad x(t_0) = y$$

for y in an open set $\mathcal{U} \subset \mathbb{R}^n$ as a flow: at the initial time t_0, there is a particle at each point $y \in \mathcal{U}$; that particle's location at time $t \geq t_0$ is given by $x(t,y)$, where $x(t,y)$ is the solution of the IVP $x' = f(x)$, $x(t_0) = y$ (e.g., f can be thought of as a steady-state velocity field).



For $t \geq t_0$, let $\mathcal{U}(t) = \{x(t,y) : y \in \mathcal{U}\}$. Then $\mathcal{U}(t) = S_{t_0}^t[\mathcal{U}]$ and $S_{t_0}^t : \mathcal{U} \to \mathcal{U}(t)$ is (for fixed t) a C^1 diffeomorphism (i.e., for fixed t, the map $y \mapsto x(t,y)$ is a C^1 diffeomorphism on $\mathcal{U}$). In particular, $\det D_Y x(t,y)$ never vanishes. Assuming, in addition, that $\mathcal{U}$ is connected, $\det D_y x(t,y)$ must either be always positive or always negative; since $\det D_y x(t_0, y) = \det I = 1 > 0$, $\det D_y x(t,y)$ is always > 0. Now the volume $\mathrm{vol}(\mathcal{U}(t))$ satisfies

$$\mathrm{vol}(\mathcal{U}(t)) = \int_{\mathcal{U}(t)} 1\, dx = \int_{\mathcal{U}} |\det D_y x(t,y)| dy = \int_{\mathcal{U}} \det D_y x(t,y) dy.$$

Assuming differentiation under the integral sign is justified (e.g., if $\mathcal{U}$ is contained in a compact set K and $S_{t_0}^t$ can be extended to $y \in K$),

$$\begin{aligned} \frac{d}{dt}\left(\mathrm{vol}\left(\mathcal{U}(t)\right)\right) &= \int_{\mathcal{U}} \frac{d}{dt}\left(\det D_y x(t,y)\right) dy = \int_{\mathcal{U}} \mathrm{div} f\left(x(t,y)\right) \det D_y x(t,y) dy \\ &= \int_{\mathcal{U}(t)} \mathrm{div} f(x) dx, \end{aligned}$$

by the previous example, where the divergence of f is by definition

$$\mathrm{div} f(x) = \frac{\partial f_1}{\partial x_1} + \frac{\partial f_2}{\partial x_2} + \cdots + \frac{\partial f_n}{\partial x_n} = \mathrm{tr}\ \left(D_x f(x)\right).$$

This gives the rate of change of the volume of $\mathcal{U}(t)$ as the integral of the divergence of f over $\mathcal{U}(t)$. In particular, if $\mathrm{div} f(x) \equiv 0$, then $\frac{d}{dt}\left(\mathrm{vol}\left(\mathcal{U}(t)\right)\right) = 0$, and volume is conserved.

Remark. The same argument applies when $f = f(t,x)$ depends on t as well: just replace $\mathrm{div} f(x)$ by $\mathrm{div}_x f(t,x)$, the divergence of f (with respect to x):

$$\mathrm{div}_x f(t,x) = \left(\frac{\partial f_1}{\partial x_1} + \cdots + \frac{\partial f_n}{\partial x_n}\right)\Bigg|_{(t,x)}.$$

Linear Systems with Periodic Coefficients

Let $A : \mathbb{R} \to \mathbb{C}^{n\times n}$ be continuous, and periodic with period $\omega > 0$:

$$(\forall\, t \in \mathbb{R}) \qquad A(t+\omega) = A(t)..$$

Note that in this case we take the scalar field to be $\mathbb{F} = \mathbb{C}$. Consider the periodic linear homogeneous system

$$\text{(PLH)} \qquad x' = A(t)x, \quad t \in \mathbb{R}.$$

All solutions exist $(\forall\, t \in \mathbb{R})$ since $f(t,x) \equiv A(t)x$ is uniformly Lipschitz in x on $\mathbb{R} \times \mathbb{C}^n$, since, by continuity, there exists $M > 0$ such that

$$|A(t)| \le M \quad \forall\, t \in \mathbb{R}.$$

M is a uniform Lipschitz constant for $f(t,x) = A(t)x$.

Lemma. If $\Phi(t)$ is a F.M. for (PLH), then so also is $\Psi(t) \equiv \Phi(t+\omega)$.

Proof. For each t, $\Psi(t)$ is invertible. Also, $\Psi'(t) = \Phi'(t+\omega) = A(t+\omega)\Phi(t+\omega) = A(t)\Psi(t)$, so $\Psi(t)$ is a matrix solution of (PLH). □

Theorem. To each F.M. $\Phi(t)$ for (PLH), there exists an invertible periodic C^1 matrix function $P : \mathbb{R} \to \mathbb{C}^{n\times n}$ and a *constant* matrix $R \in \mathbb{C}^{n\times n}$ for which $\Phi(t) = P(t)e^{tR}$.

Proof. By the lemma, there is an invertible matrix $C \in \mathbb{C}^{n\times n}$ such that $\Phi(t+\omega) = \Phi(t)C$. Since C is invertible, it has a logarithm, i.e. there exists a matrix $W \in \mathbb{C}^{n\times n}$ such that $e^W = C$. Let $R = \frac{1}{\omega}W$. Then $C = e^{\omega R}$. Define $P(t) = \Phi(t)e^{-tR}$. Then $P(t)$ is invertible for all t, $P(t)$ is C^1, and $\Phi(t) = P(t)e^{tR}$. Finally,

$$\begin{aligned} P(t+\omega) &= \Phi(t+\omega)e^{-(t+\omega)R} \\ &= \Phi(t)Ce^{-\omega R}e^{-tR} = \Phi(t)e^{-tR} = P(t), \end{aligned}$$

so $P(t)$ is periodic. □

Linear Scalar n^{th}-order ODEs

Let $I \equiv [a,b]$ be an interval in $\mathbb{R}$, and suppose $a_j(t)$ are in $C(I,\mathbb{F})$ for $j = 0,1,\ldots,n$, with $a_n(t) \neq 0\, \forall\, t \in I$. Consider the n^{th}-order linear differential operator $L : C^n(I) \to C(I)$ given by

$$Lu = a_n(t)\frac{d^n u}{dt^n} + \cdots + a_1(t)\frac{du}{dt} + a_0(t)u,$$

and the n^{th}-order homogeneous equation (nLH) $Lu = 0$, $t \in I$. Consider the equivalent $n \times n$ first-order system (LH) $x' = A(t)x$, $t \in I$, where

$$A(t) = \begin{bmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \\ \frac{-a_0}{a_n} & \cdots & & -\frac{a_{n-1}}{a_n} \end{bmatrix} \qquad \text{and} \qquad x = \begin{bmatrix} u \\ u' \\ u'' \\ \vdots \\ u^{(n-1)} \end{bmatrix} \in \mathbb{F}^n.$$

Fix $t_0 \in I$. Appropriate initial conditions for (nLH) are

$$\begin{bmatrix} u(t_0) \\ u'(t_0) \\ \vdots \\ u^{(n-1)}(t_0) \end{bmatrix} = x(t_0) = \zeta \equiv \begin{bmatrix} \zeta_1 \\ \zeta_2 \\ \vdots \\ \zeta_n \end{bmatrix}.$$

Recall that u is a C^n solution of (nLH) if and only if x is a C^1 solution of (LH), with a similar equivalence between associated IVP's. If $\Phi(t)$ is a F.M. for (LH), with $A(t)$ as given above, then $\Phi(t)$ has the form

$$\Phi = \begin{bmatrix} \varphi_1 & \varphi_2 & \cdots & \varphi_n \\ \varphi_1' & \varphi_2' & \cdots & \varphi_n' \\ \vdots & \vdots & & \vdots \\ \varphi_1^{(n-1)} & \varphi_2^{(n-1)} & \cdots & \varphi_n^{(n-1)} \end{bmatrix},$$

where each $\varphi_j(t)$ satisfies (nLH).

Definition. If $\varphi_1(t), \ldots, \varphi_n(t)$ are solutions of (nLH), then the *Wronskian* of $\varphi_1, \ldots, \varphi_n$ (a scalar function of t) is defined to be

$$W(\varphi_1, \ldots, \varphi_n)(t) = \det \begin{bmatrix} \varphi_1(t) & & \varphi_n(t) \\ \varphi_1'(t) & \cdots & \varphi_n'(t) \\ \vdots & & \vdots \\ \varphi_1^{(n-1)}(t) & & \varphi_n^{(n-1)}(t) \end{bmatrix} (= \det \Phi(t)).$$

Since $\Phi(t)$ is a matrix solution of (LH), we know

$$\det(\Phi(t)) = \det(\Phi(t_0)) \exp \int_{t_0}^{t} \operatorname{tr}(A(s)) ds,$$

so

$$W(\varphi_1, \ldots, \varphi_n)(t) = W(\varphi_1, \ldots, \varphi_n)(t_0) \exp \int_{t_0}^{t} -\frac{a_{n-1}(s)}{a_n(s)} ds.$$

In particular, for solutions $\varphi_1, \ldots, \varphi_n$ of (nLH),

$$\text{either} \quad W(\varphi_1, \ldots, \varphi_n)(t) \equiv 0 \text{ on } I, \text{ or } (\forall\, t \in I) \quad W(\varphi_1, \ldots, \varphi_n)(t) \neq 0.$$

Theorem. Let $\varphi_1, \ldots, \varphi_n$ be n solutions of (nLH) $Lu = 0$. Then they are linearly independent on I (i.e., as elements of $C^n(I)$) if and only if $W(\varphi_1, \ldots, \varphi_n)(t) \neq 0$ on I.

Proof. If $\varphi_1, \ldots, \varphi_n$ are linearly dependent in $C^n(I)$, then there exist scalars $c_1, \ldots, c_n$ such that

$$c_1\varphi_1(t) + \cdots + c_n\varphi_n(t) \equiv 0 \text{ on } I, \text{ with } c \equiv \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix} \neq 0;$$

thus $\Phi(t)c = 0$ on I, so $W(\varphi_1, \ldots, \varphi_n)(t) = \det \Phi(t) = 0$ on I. Conversely, if $\det \Phi(t) = 0$ on I, then the solutions

$$\begin{bmatrix} \varphi_1 \\ \vdots \\ \varphi_1^{(n-1)} \end{bmatrix}, \cdots, \begin{bmatrix} \varphi_n \\ \vdots \\ \varphi_n^{(n-1)} \end{bmatrix}$$

of (LH) are linearly dependent (as elements of $C^1(I, \mathbb{F}^n)$), so there exist scalars $c_1, \ldots, c_n$ such that

$$c_1 \begin{bmatrix} \varphi_1(t) \\ \vdots \end{bmatrix} + \cdots + c_n \begin{bmatrix} \varphi_n(t) \\ \vdots \end{bmatrix} \equiv 0 \text{ on } I,$$

where not all $c_j = 0$. In particular, $c_1\varphi_1(t) + \cdots + c_n\varphi_n(t) \equiv 0$ on I, so $\varphi_1, \ldots, \varphi_n$ are linearly dependent in $C^n(I)$. □

Corollary. The dimension of the vector space of solutions of (nLH) (a subspace of $C^n(I)$) is n, i.e., $\dim \mathcal{N}(L) = n$, where $\mathcal{N}(L)$ denotes the null space of $L : C^n(I) \to C(I)$.

The differential operator L (normalized so that $a_n(t) \equiv 1$) is itself determined by n linearly independent solutions of (nLH) $Lu = 0$.

Fact. Suppose $\varphi_1(t), \ldots, \varphi_n(t) \in C^n(I)$ with $W(\varphi_1, \ldots, \varphi_n)(t) \neq 0$ $(\forall\, t \in I)$. Then there exists a unique n^{th} order linear differential operator

$$L = \frac{d^n}{dt^n} + a_{n-1}(t)\frac{d^{n-1}}{dt^{n-1}} + \cdots + a_1(t)\frac{d}{dt} + a_0(t)$$

(with $a_n(t) \equiv 1$ and each $a_j(t) \in C(I)$) for which $\varphi_1, \ldots, \varphi_n$ form a fundamental set of solutions of (nLH) $Lu = 0$, namely,

$$Lu = \frac{W(\varphi_1, \ldots, \varphi_n, u)}{W(\varphi_1, \ldots, \varphi_n)}$$

where

$$W(\varphi_1, \ldots, \varphi_n, u) = \det \begin{bmatrix} \varphi_1 & \cdots & \varphi_n & u \\ \varphi_1' & & \varphi_n' & u' \\ \vdots & & \vdots & \vdots \\ \varphi_1^{(n)} & \cdots & \varphi_n^{(n)} & u^{(n)} \end{bmatrix}.$$

Sketch. In this formula for Lu, expanding the determinant in the last column shows that L is an n^{th} order linear differential operator with continuous coefficients $a_j(t)$ and $a_n(t) \equiv 1$. Clearly $\varphi_1, \ldots, \varphi_n$ are solutions of $Lu = 0$. For uniqueness (with $a_n(t) \equiv 1$), note that if $\varphi_1, \ldots, \varphi_n$ are linearly independent solutions of $Lu = 0$ for any L, then

$$\Phi^T(t) \begin{bmatrix} a_0(t) \\ a_1(t) \\ \vdots \\ a_{n-1}(t) \end{bmatrix} = - \begin{bmatrix} \varphi_1^{(n)}(t) \\ \vdots \\ \varphi_n^{(n)}(t) \end{bmatrix}.$$

Since $W(\varphi_1, \ldots, \varphi_n)(t) \neq 0$ $(\forall\, t \in I)$, $\Phi(t)$ is invertible $\forall\, t \in I$, so

$$\begin{bmatrix} a_1(t) \\ \vdots \\ a_{n-1}(t) \end{bmatrix} = -\Phi^{-T}(t) \begin{bmatrix} \varphi^{(n)}(t) \\ \vdots \\ \varphi_n^{(n)}(t) \end{bmatrix}$$

is uniquely determined by $\varphi_1, \ldots, \varphi_n$.

Remark. A first-order system (LH) $x' = A(t)x$ is uniquely determined by any F.M. $\Phi(t)$. Since $\Phi'(t) = A(t)\Phi(t)$, $A(t) = \Phi'(t)\Phi^{-1}(t)$.

Linear Inhomogeneous n^{th}-order scalar equations

For simplicity, normalize the coefficients $a_j(t)$ so that $a_n(t) \equiv 1$ in L. Consider (nLI)

$$Lu = u^{(n)} + a_{n-1}(t)u^{(n-1)} + \cdots + a_0(t)u = \beta(t).$$

Let

$$x = \begin{bmatrix} u \\ u' \\ \vdots \\ u^{(n-1)} \end{bmatrix}, \quad b(t) = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \beta(t) \end{bmatrix}, \quad \text{and} \quad A(t) = \begin{bmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \\ -a_0 & \cdots & & -a_{n-1} \end{bmatrix},$$

then $x(t)$ satisfies (LI) $x' = A(t)x + b(t)$. We can apply our results for (LI) to obtain expressions for solutions of (nLI).

Theorem. If $\varphi_1, \ldots, \varphi_n$ is a fundamental set of solutions of (nLH) $Lu = 0$, then the solution $\psi(t)$ of (nLI) $Lu = \beta(t)$ with initial condition $u^{(k)}(t_0) = \zeta_{k+1}$ $(k = 0, \ldots, n-1)$ is

$$\psi(t) = \varphi(t) + \sum_{k=1}^{n} \varphi_k(t) \int_{t_0}^{t} \frac{W_k(\varphi_1, \ldots, \varphi_n)(s)}{W(\varphi_1, \ldots, \varphi_n)(s)} \beta(s)ds$$

where $\varphi(t)$ is the solution of (nLH) with the same initial condition at t_0, and W_k is the determinant of the matrix function obtained from

$$\Phi(t) = \begin{bmatrix} \varphi_1 & \cdots & \varphi_n \\ \varphi_1' & \cdots & \varphi_n' \\ \vdots & & \vdots \\ \varphi_1^{(n-1)} & \cdots & \varphi_n^{(n-1)} \end{bmatrix}$$

by replacing the k^{th} column of $\Phi(t)$ replaced by the nth unit coordinate vector e_n.

Proof. We know

$$x(t) = \Phi(t)\Phi^{-1}(t_0)x_0 + \Phi(t) \int_{t_0}^{t} \Phi^{-1}(s)b(s)ds,$$

where $x_0 = [\zeta_1, \ \cdots \ \zeta_n]^T$ and $b(s) = [0 \ \cdots \ \beta(s)]^T$, solves the IVP $x' = A(t)x, \ x(t_0) = x_0$. The first component of $x(t)$ is $\psi(t)$, and the first component of $\Phi(t)\Phi^{-1}(t_0)x_0$ is the solution $\varphi(t)$ of (nLH) described above. By Cramer's Rule,

$$\text{the } k^{\text{th}} \text{ component of } \Phi^{-1}(s)e_n \text{ is } \frac{W_k(\varphi_1, \ldots, \varphi_n)(s)}{W(\varphi_1, \ldots, \varphi_n)(s)}.$$

Thus the first component of $\Phi(t) \int_{t_0}^t \Phi^{-1}(s)b(s)ds$ is

$$[\varphi_1(t) \cdots \varphi_n(t)] \int_{t_0}^t \Phi^{-1}(s)e_n\beta(s)d = \sum_{k=1}^n \varphi_k(t) \int_{t_0}^t \frac{W_k(\varphi_1, \ldots, \varphi_n)(s)}{W(\varphi_1, \ldots, \varphi_n)(s)}\beta(s)ds.$$

□

Linear n^{th}-order scalar equations with constant coefficients

For simplicity, take $a_n = 1$ and $\mathbb{F} = \mathbb{C}$. Consider

$$Lu = u^{(n)} + a_{n-1}u^{(n-1)} + \cdots + a_0u,$$

where $a_0, \ldots, a_{n-1}$ are constants. Then

$$A = \begin{bmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \\ -a_0 & & \cdots & -a_{n-1} \end{bmatrix}$$

has characteristic polynomial

$$p(\lambda) = \lambda^n + a_{n-1}\lambda^{n-1} + \cdots + a_1\lambda + a_0.$$

Moreover, since A is a companion matrix it is nonderogatory, i.e., each distinct eigenvalue of A has only one Jordan block in the Jordan form of A. Indeed, recall that for any λ,

$$A - \lambda I = \begin{bmatrix} -\lambda & 1 & & 0 \\ & \ddots & \ddots & \\ -a_0 & & & (-a_{n-1} - \lambda) \end{bmatrix}$$

has rank $\geq n - 1$, so the geometric multiplicity of each eigenvalue is $1 = \dim(\mathcal{N}(A - \lambda I))$.

Now if λ_k is a root of $p(\lambda)$ having multiplicity m_k (as a root of $p(\lambda)$), then terms of the form $t^j e^{\lambda_k t}$ for $0 \leq j \leq m_k - 1$ appear in elements of e^{tJ} (where $P^{-1}AP = J$ is in Jordan form), and thus also appear in $e^{tA} = Pe^{tJ}P^{-1}$ the F.M. for (LH) $x' = Ax$, normalized at 0. This explains the well-known result:

Theorem. Let $\lambda_1, \ldots, \lambda_s$ be the distinct roots of $p(\lambda) = \lambda^n + a_{n-1}\lambda^{n-1} + \cdots + a_0 = 0$, and suppose λ_k has multiplicity m_k for $1 \leq k \leq s$. Then a fundamental set of solutions of

$$Lu = u^{(n)} + a_{n-1}u^{(n-1)} + \cdots + a_0u = 0,$$

where $a_k \in \mathbb{C}$, is

$$\{t^j e^{\lambda_k t} : 1 \leq k \leq s, 0 \leq j \leq m_k - 1\}.$$

Standard proof: Show that the functions are linear indepent and then plug in and verify they are solutions: write

$$L = \left(\frac{d}{dt} - \lambda_1\right)^{m_1} \cdots \left(\frac{d}{dt} - \lambda_s\right)^{m_s},$$

and use

$$\left(\frac{d}{dt} - \lambda_k\right)^{m_k} \left(t^j e^{\lambda_k t}\right) = 0 \quad \text{for} \quad 0 \leq j \leq m_k - 1.$$

Introduction to the Numerical Solution of IVP for ODE

Consider the IVP: DE $x' = f(t,x)$, IC $x(a) = x_a$. For simplicity, we will assume here that $x(t) \in \mathbb{R}^n$ (so $\mathbb{F} = \mathbb{R}$), and that $f(t,x)$ is continuous in t, x and uniformly Lipschitz in x (with Lipschitz constant L) on $[a,b] \times \mathbb{R}^n$. So we have global existence and uniqueness for the IVP above on $[a,b]$.

Moreover, the solution of the IVP $x' = f(t,x)$, $x(a) = x_a$ depends continuously on the initial values $x_a \in \mathbb{R}^n$. This IVP is an example of a *well-posed problem*: for each choice of the "data" (here, the initial values x_a), we have:

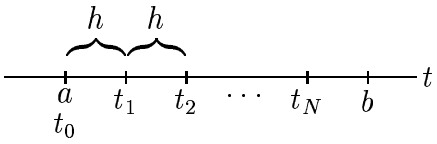
(1) **Existence.** There exists a solution of the IVP on $[a,b]$.

(2) **Uniqueness.** The solution, for each given x_a, is unique.

(3) **Continuous Dependence.** The solution depends continuously on the data.

Here, e.g., the map $x_a \mapsto x(t,x_a)$ is continuous from $\mathbb{R}^n$ into $(C([a,b]), \|\cdot\|_\infty)$. A well-posed problem is a reasonable problem to approximate numerically.

Grid Functions



Choose a mesh width h (with $0 < h \le b-a$, and let $N = \left[\frac{b-a}{h}\right]$ (greatest integer $\le (b-a)/h$). Let $t_i = a + ih$ $(i = 0, 1, \ldots, N)$ be the grid points in t (note: $t_0 = a$), and let x_i denote the approximation to $x(t_i)$. Note that t_i and x_i depend on h, but we will usually suppress this dependence in our notation.

Explicit One-Step Methods

Form of method: start with x_0 (presumably $x_0 \approx x_a$). Recursively compute $x_1, \ldots, x_N$ by

$$x_{i+1} = x_i + h\psi(h, t_i, x_i), \qquad i = 0, \ldots, N-1.$$

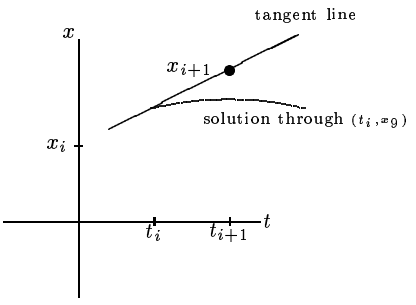
Here, $\psi(h,t,x)$ is a function defined for $0 \le h \le b-a$, $a \le t \le b$, $x \in \mathbb{R}^n$, and ψ is associated with the given function $f(t,x)$.

Examples.

Euler's Method.

$$x_{i+1} = x_i + hf(t_i, x_i)$$

Here, $\psi(h,t,x) = f(t,x)$.



Taylor Methods. To see how the Taylor Method of order p (p as in integer ≥ 1) is constructed, consider the Taylor expansion of a C^{p+1} solution $x(t)$ of $x' = f(t,x)$:

$$x(t+h) = x(t) + hx'(t) + \cdots + \frac{h^p}{p!}x^{(p)}(t) + \underbrace{\frac{h^{p+1}}{(p+1)!}x^{(p+1)}(\theta)}_{\text{remainder term}}$$

In the remainder term, θ is between t and $t+h$, so the remainder term is $\mathcal{O}(h^{p+1})$, that is the magnitude of the remainder term is bounded above by a constant multiple of h^{p+1} for all $h > 0$ sufficiently small. Here we can take the constant to be $\frac{1}{(p_1)!}\max_{a\leq t\leq b}|x^{(p+1)}(t)|$. In the approximation, we will neglect the remainder term, and use the DE $x' = f(t,x)$ to replace $x'(t), x''(t), \ldots$ by expressions involving f and its derivatives:

$$\begin{aligned} x'(t) &= f(t,x(t)) \\ x''(t) &= \frac{d}{dt}(f(t,x(t))) = \overset{(n\times 1)}{D_t f}\Big|_{(t,x(t))} + \overset{(n\times n)}{D_x f}\Big|_{(t,x(t))} \overset{(n\times 1)}{\frac{dx}{dt}} \\ &= (D_t f + (D_x f)f)\Big|_{(t,x(t))} \qquad \text{(for } n=1\text{, this is } f_t + f_x f\text{).} \end{aligned}$$

For higher derivatives, inductively differentiate the expression for the previous derivative, and replace any occurrence of $\frac{dx}{dt}$ by $f(t,x(t))$. These expansions lead us to define the Taylor methods of order p:

$$p = 1 : x_{i+1} = x_i + hf(t_i, x_i) \qquad \text{(Euler's method, } \psi(h,t,x) = f(t,x))$$

$$p = 2 : x_{i+1} = x_i + hf(t_k, x_i) + \frac{h^2}{2}(D_t f + (D_x f)f)\Big|_{(t_i,x_i)}$$

For the case $p = 2$, we have

$$\psi(h,t,x) = T_2(h,t,x) \equiv \left(f + \frac{h}{2}(D_t f + (D_x f)f)\right)\Big|_{(t,x)}.$$

We will use the notation $T_p(h,t,x)$ to denote the $\psi(h,t,x)$ function for the Taylor method of order p.

Remark. Taylor methods of order ≥ 2 are rarely used computationally. They require derivatives of f to be programmed and evaluated. They are, however, of theoretical interest in determining the order of a method.

Remark. A "one-step method" is actually an association of a function $\psi(h,t,x)$ (defined for $0 \leq h \leq b-a, a \leq t \leq b, x \in \mathbb{R}^n$) to each function $f(t,x)$ (which is continuous in t,x and

Lipschitz in x on $[a,b] \times \mathbb{R}^n$). We study "methods" looking at one function f at a time. Many methods (e.g., Taylor methods of order $p \geq 2$) require more smoothness of f, either for their definition, or to guarantee that the solution $x(t)$ is sufficiently smooth. Recall that if $f \in C^p$ (in t and x), then the solution $x(t)$ of the IVP $x' = f(t,x)$, $x(a) = x_a$ is in $C^{p+1}[a,b]$. For "higher-order" methods, this smoothness is essential in getting the error to be higher order in h. We will assume from here on (usually tacitly) that f is sufficiently smooth when needed.

Examples.
Modified Euler's Method

$$\begin{aligned} x_{i+1} &= x_i + hf\left(t_i + \frac{h}{2}, x_i + \frac{h}{2}f(t_i, x_i)\right) \\ (\text{so } \psi(h,t,x) &= f\left(t + \frac{h}{2}, x + \frac{h}{2}f(t,x)\right)). \end{aligned}$$

Here $\psi(h,t,x)$ tries to approximate

$$x'\left(t + \frac{h}{2}\right) = f\left(t + \frac{h}{2}, x\left(t + \frac{h}{2}\right)\right),$$

using the Euler approximation to $x\left(t + \frac{h}{2}\right)$ $\left(\approx x(t) + \frac{h}{2}f(t,x(t))\right)$.

Improved Euler's Method (or Heun's Method)

$$\begin{aligned} x_{i+1} &= x_i + \frac{h}{2}\left(f(t_i,x_i) + f\left(t_{i+1}, x_i + hf(t_i,x_i)\right)\right) \\ (\text{so } \psi(h,t,x) &= \frac{1}{2}\left(f(t,x) + f\left(t+h, x + hf(t,x)\right)\right)). \end{aligned}$$

Here again $\psi(h,t,x)$ tries to approximate

$$x'\left(t + \frac{h}{2}\right) \approx \frac{x'(t) + x'(t+h)}{2}.$$

Or $\psi(h,t,x)$ can be viewed as an approximation to the trapezoid rule applied to

$$\frac{1}{h}\left(x(t+h) - x(t)\right) = \frac{1}{h}\int_t^{t+h} x' \approx \frac{1}{2}x'(t) + \frac{1}{2}x'(t+h).$$

Modified Euler and Improved Euler are examples of 2^{nd} order two-stage Runge-Kutta methods. Notice that no derivatives of f need be evaluated, but f needs to be evaluated *twice* in each step (from x_i to x_{i+1}).

Before stating the convergence theorem, we introduce the concept of *accuracy*.

Local Truncation Error

Let $x_{i+1} = x_i + h\psi(h, t_i, x_i)$ be a one-step method, and let $x(t)$ be a solution of the DE $x' = f(t, x)$. The *local truncation error* (LTE) for $x(t)$ is defined to be

$$l(h, t) \equiv x(t + h) - (x(t) + h\psi(h, t, x(t))),$$

that is, the local truncation error is *the amount by which the true solution of the DE fails to satisfy the numerical scheme.*

Given h, define

$$\tau(h, t) = \frac{l(h, t)}{h} \quad \text{for } h > 0,$$

and set $\tau_i(h) = \tau(h, t_i)$. Also, set

$$\tau(h) = \max_{a \le t \le b} |\tau(h, t)| \quad \text{for } h > 0.$$

Note that

$$l(h, t_i) = x(t_{i+1}) - (x(t_i) + h\psi(h, t_i, x(t_i))),$$

explicitly showing the dependence of l on h, t_i, and $x(t)$.

Definition. A one-step method is called [formally] *accurate of order p* (for a positive integer p) if for any solution $x(t)$ of the DE $x' = f(t, x)$ which is C^{p+1}, we have $l(h, t) = \mathcal{O}(h^{p+1})$.

Definition. A one-step method is called *consistent* if $\psi(0, t, x) = f(t, x)$. Consistency is essentially minimal accuracy:

Proposition. A one-step method

$$x_{i+1} = x_i + h\psi(h, t_i, x_i),$$

where $\psi(h, t, x)$ is continuous for $0 \le h \le h_0$, $a \le t \le b$, $x \in \mathbb{R}^n$ for some $h_0 \in (0, b - a]$, is consistent for the DE $x' = f(t, x)$ if and only if $\tau(h) \to 0$ as $h \to 0^+$.

Proof. Suppose the method is consistent. Fix a solution $x(t)$. For $0 < h \le h_0$, let

$$Z(h) = \max_{\{(k,t)|\, a \le t \le b, a \le t+k \le b, 0 \le k \le h,\}} |\psi(k, t, x(t)) - \psi(0, t + k, x(t + k))|.$$

By uniform continuity, $Z(h) \to 0$ as $h \to 0^+$. For some $\theta \in (0, h)$ and $a \le t < b$,

$$\begin{aligned} x(t + h) &= x(t + h) = x(t) + hx'(t + \theta) \\ &= x(t) + hf(t + \theta, x(t + \theta)) \\ &= x(t) + h\psi(0, t + \theta, x(t + \theta)). \end{aligned}$$

Combining this with the definiton of $l(h, t)$ gives

$$|l(h, t)| = h|\psi(0, t + \theta, x(t + \theta)) - \psi(h, t, x(t))| \le hZ(h),$$

so $\tau(h) \le Z(h) \to 0$. Conversely, if $\tau(h) \to 0$, then for any $t \in [a, b)$ and any $h \in (0, b - t]$,

$$\frac{x(t + h) - x(t)}{h} = \psi(h, t, x(t)) + \tau(h, t).$$

Taking the limit as $h \downarrow 0$ gives $f(t, x) = x'(t) = \psi(0, t, x(t))$. □

Convergence Theorem for One-Step Methods

Theorem. Let $f(t,x)$ be a continuous mapping from $[a,b] \times \mathbb{F}^n$ into $\mathbb{F}^n$, and assume that f is uniformly Lipschitz in x on $[a,b] \times \mathbb{R}^n$. Let $x(t)$ be the solution of the IVP $x' = f(t,x)$, $x(a) = x_a$ on $[a,b]$. Suppose that the function $\psi(h,t,x)$ in the one step method satisfies the following two conditions

1. (*Stability*) $\psi(h,t,x)$ is continuous in h,t,x and uniformly Lipschitz in x (with Lipschitz constant K) on $0 \le h \le h_0$, $a \le t \le b$, $x \in \mathbb{R}^n$ for some $h_0 > 0$ with $h_0 \le b-a$, and

2. (*Consistency*) $\psi(0,t,x) = f(t,x)$.

Given $h \in (0, b-a]$, recursively define

$$x_{i+1}(h) = x_i(h) + h\psi(h, t_i, x_i(h)) \quad \text{for } 0 \le i \le \frac{b-a}{h},$$

as in the one-step method. Define

$$t_k(h) = a + kh \quad \text{for } 0 \le k \le \frac{b-a}{h},$$

and set

$$e_k(h) = x(t_k(h)) - x_k(h) \quad \text{for } 0 \le k \le \frac{b-a}{h}.$$

The vector $e_k(h)$ is the error in estimating the true solution to the IVP at $a+kh$, namely $x(a+kh)$, by the approximation given by the one-step method, $x_k(h)$. In particular, $e_0(h) = x_a - x_0(h)$ is the error in the initial value $x_0(h)$. With these definitions, we have

$$|e_i(h)| \le e^{K(t_i(h)-a)}|e_0(h)| + \tau(h)\left(\frac{e^{K(t_i(h)-a)}-1}{K}\right),$$

so

$$|e_i(h)| \le e^{K(b-a)}|e_0(h)| + \frac{e^{K(b-a)}-1}{K}\tau(h).$$

Moreover, $\tau(h) \to 0$ as $h \to 0$. Therefore, if $e_0(h) \to 0$ as $h \to 0$, then

$$\max_{0 \le i \le \frac{b-a}{h}} |e_i(h)| \ \to 0 \quad \text{as} \quad h \to 0,$$

which implies the uniform convergence of the iterates on the grid.

Proof. Hold $h > 0$ fixed, and ignore rounding error. Subtracting

$$x_{i+1} = x_i + h\psi(h, t_i, x_i)$$

from

$$x(t_{i+1}) = x(t_i) + h\psi(h, t_i, x(t_i)) + h\tau_i,$$

gives

$$\begin{aligned} |e_{i+1}| &\leq |e_i| + h|\psi(h, t_i, x(t_i)) - \psi(h, t_i, x_i)| + h|\tau_i| \\ &\leq |e_i| + hK|e_i| + h\tau(h). \end{aligned}$$

So

$$\begin{aligned} |e_1| &\leq (1 + hK)|e_0| + h\tau(h), \quad \text{and} \\ |e_2| &\leq (1 + hK)|e_1| + h\tau(h) \\ &\leq (1 + hK)^2|e_0| + h\tau(h)(1 + (1 + hK)). \end{aligned}$$

By induction,

$$\begin{aligned} |e_i| &\leq (1 + hK)^i|e_0| + h\tau(h)(1 + (1 + hK) + (1 + hK)^2 + \cdots + (1 + hK)^{i-1}) \\ &= (1 + hK)^i|e_0| + h\tau(h)\frac{(1 + hK)^i - 1}{(1 + hK) - 1} \\ &= (1 + hK)^i|e_0| + \tau(h)\frac{(1 + hK)^i - 1}{K} \end{aligned}$$

Now use $(1 + hK)^{\frac{1}{h}} \uparrow e^K$ as $h \to 0^+$ (for $K > 0$), and $i = \frac{t_i - a}{h}$ to obtain

$$|e_i| \leq e^{K(t_i - a)}|e_0| + \tau(h)\frac{e^{K(t_i - a)} - 1}{K}$$

since

$$(1 + hK)^j = (1 + hK)^{\frac{t_j - a}{n}} \leq e^{K(t_i - a)}.$$

The preceding proposition shows $\tau(h) \to 0$, and the theorem follows. □

The theorem implies that if a one-step method is accurate of order p and stable [i.e. ψ is Lipschitz in x], then for sufficiently smooth f, $x(t) \in C^{p+1}$, so

$$l(h, t) = \mathcal{O}(h^{p+1}) \quad \text{and thus} \quad \tau(h) = \mathcal{O}(h^p).$$

If, in addition, $e_0(h) = \mathcal{O}(h^p)$, then

$$\max_i |e_i(h)| = \mathcal{O}(h^p),$$

i.e. p^{th} order convergence.

Example. The "Taylor method of order p" is accurate of order p. If $f \in C^p$, then $x \in C^{p+1}$, and

$$l(h, t) = x(t + h) - \left(x(t) + hx'(t) + \cdots + \frac{h^p}{p!}x^{(p)}(t)\right) = x^{(p+1)}(\theta)\frac{h^{p+1}}{(p + 1)!}.$$

So

$$|l(h, t)| \leq M_{p+1}\frac{h^{p+1}}{(p + 1)!} \quad \text{where} \quad M_{p+1} = \max_{a \leq t \leq b} |x^{(p+1)}(t)|.$$

Fact. A one-step method $x_{i+1} = x_i + h\psi(h, t_i, x_i)$ is accurate or order p if and only if

$$\psi(h, t, x) = T_p(h, t, x) + \mathcal{O}(h^p),$$

where T_p is the "ψ" for the Taylor method of order p.

Proof. Since

$$x(t + h) - x(t) = hT_p(h, t, x(t)) + \mathcal{O}(h^{p+1}),$$

we have for any given one-step method that

$$\begin{aligned} l(h, t) &= x(t + h) - x(t) - h\psi(h, t, x(t)) \\ &= hT_p(h, t, x(t)) + \mathcal{O}(h^{p+1}) - h\psi(h, t, x(t)) \\ &= h(T_p(h, t, x(t)) - \psi(h, t, x(t))) + \mathcal{O}(h^{p+1}). \end{aligned}$$

So $l(h, t) = \mathcal{O}(h^{p+1})$ iff $h(T_p(h, t, x(t)) - \psi(h, t, x(t))) = \mathcal{O}(h^{p+1})$ iff $\psi = T_p + \mathcal{O}(h^p)$. □

Remark. The controlled growth of the effect of the local truncation error (LTE) from previous steps in the proof of the convergence theorem (a consequence of the Lipschitz continuity of ψ in x) is called *stability*. The theorem states:

Stability + Consistency (minimal accuracy) ⇒ Convergence.

In fact, here, the converse is also true.

Explicit Runge-Kutta methods

One of the problems with Taylor methods is the need to evaluate higher derivatives of f. Runge-Kutta (RK) methods replace this with the much more reasonable need to evaluate f more than once to go from x_i to x_{i+1}. An m-stage (explicit) RK method is of the form

$$x_{i+1} = x_i + h\psi(h, t_i, x_i),$$

with

$$\psi(h, t, x) = \sum_{j=1}^{m} a_j k_j(h, t, x),$$

where $a_1, \ldots, a_m$ are given constants,

$$k_1(h, t, x) = f(t, x)$$

and for $2 \leq j \leq m$,

$$k_j(h, t, x) = f(t + \alpha_j h, x + h \sum_{r=1}^{j-1} \beta_{jr} k_r(h, t, x))$$

with $\alpha_2, \ldots, \alpha_m$ and β_{jr} $(1 \le r < j \le m)$ given constants. We usually choose $0 < \alpha_j \le 1$, and for accuracy reasons we choose

$$(*) \qquad \alpha_j = \sum_{r=1}^{j-1} \beta_{jr} \qquad (2 \le j \le m).$$

Example. $m = 2$

$$x_{i+1} = x_i + h(a_1 k_1(h, t_i, x_i) + a_2 k_2(h, t_i, x_i))$$

where

$$\begin{aligned} k_1(h, t_i, x_i) &= f(t_i, x_i) \\ k_2(h, t_i, x_i) &= f(t_i + \alpha_2 h, x_i + h\beta_2 k_1(h, t_i, x_i)). \end{aligned}$$

For simplicity, write α for α_2 and β for β_2, and set $\alpha = \beta$ as in (*). Expanding in h,

$$\begin{aligned} k_2(h, t, x) &= f(t + \alpha h, x + h\beta f(t, x)) \\ &= f(t, x) + \alpha h D_t f(t, x) + (D_x f(t, x))(h\beta f(t, x)) + \mathcal{O}(h^2) \\ &= [f + h(\alpha D_t f + \beta (D_x f) f)]\,(t, x) + \mathcal{O}(h^2). \end{aligned}$$

So

$$\psi(h, t, x) = (a_1 + a_2) f + h(a_2 \alpha D_t f + a_2 \beta (D_x f) f) + \mathcal{O}(h^2).$$

Recalling that

$$T_2 = f + \frac{h}{2}(D_t f + (D_x f) f),$$

and that the method is accurate of order two if and only if

$$\psi = T_2 + \mathcal{O}(h^2),$$

we obtain the following necessary and sufficient condition on a two-stage (explicit) RK method to be accurate of order two:

$$a_1 + a_2 = 1, \quad a_2 \alpha = \frac{1}{2}, \quad \text{and} \quad a_2 \beta = \frac{1}{2}.$$

Since we have already chosen $\alpha = \beta$ (we now see why), these conditions become:

$$\boxed{a_1 + a_2 = 1, \qquad a_2 \alpha = \tfrac{1}{2}}.$$

Therefore, there is a one-parameter family (e.g., parameterized by α) of 2^{nd} order, two-stage $(m = 2)$ explicit RK methods.

Examples.

(1) Setting $\alpha = \frac{1}{2}$ gives $a_2 = 1$, $a_1 = 0$, which is the Modified Euler method.

(2) Choosing $\alpha = 1$ gives $a_2 = \frac{1}{2}$, $a_1 = \frac{1}{2}$, which is the Improved Euler method, or Heum's method.

Remark. Note that an m-stage explicit RK method requires m function evaluations (i.e., evaluations of f) in each step (x_i to $x_{i=1}$).

Attainable Orders of Accuracy for Explicit RK methods

# of stages (m)	highest order attainable	
1	1	← Euler's method
2	2	
3	3	
4	4	
5	4	
6	5	
7	6	
8	7	

Explicit RK methods are *always* stable: ψ inherits its Lipschitz continuity from f.

Example.
Modified Euler. Let L be the Lipschitz constant for f, and suppose $h \le h_0$ (for some $h_0 \le b-1$).

$$\begin{aligned} x_{i+1} &= x_i + hf\left(t_i + \frac{h}{2}, x_i + \frac{h}{2}f(t_i, x_i)\right) \\ \psi(t,h,x) &= f\left(t + \frac{h}{2}, x + \frac{h}{2}f(t,x)\right) \end{aligned}$$

So

$$\begin{aligned} |\psi(h,t,x) - \psi(h,t,y)| &\le L\left|\left(x + \tfrac{h}{2}f(t,x)\right) - \left(y + \tfrac{h}{2}f(t,y)\right)\right| \\ &\le L|x-y| + \tfrac{h}{2}L|f(t,x) - f(t,y)| \\ &\le L|x-y| + \tfrac{h}{2}L^2|x-y| \\ &\le K|x-y| \end{aligned}$$

where $K = L + \frac{h_0}{2}L^2$ is thus the Lipschitz constant for ψ.

Example. A popular 4^{th} order four-stage RK method is

$$x_{i+1} = x_i + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4)$$

where

$$\begin{aligned} k_1 &= f(t_i, x_i) \\ k_2 &= f\left(t_i + \frac{h}{2}, x_i + \frac{h}{2}k_1\right) \\ k_3 &= f\left(t_i + \frac{h}{2}, x_i + \frac{h}{2}k_2\right) \\ k_4 &= f(t_i + h, x_i + hk_3). \end{aligned}$$

The same argument as above shows this method is stable.

Remark. RK methods require multiple function evaluations per step (going from x_i to x_{i+1}). One-step methods discard information from previous steps (e.g., x_{i-1} is not used to get x_{i+1} — except in its influence on x_i). We next study a class of multi-step methods. But first, we make a few observations about linear difference equations.

Linear Difference Equations (Constant Coefficients)

In this discussion, x_i will be a (scalar) sequence defined for $i \geq 0$. Consider the linear difference equation (k-step)

$$\text{(LDE)} \qquad x_{i+k} + \alpha_{k-1}x_{i+k-1} + \cdots + \alpha_0 x_i = b_i \qquad\qquad (i \geq 0).$$

If $b_i \equiv 0$, the linear difference equation (LDE) is said to be homogeneous, in this case we will refer to it as (lh). If $b_i \neq 0$ for some $i \geq 0$, the linear difference equation (LDE) is said to be inhomogeneous , in this case we refer to it as (li).

Initial Value Problem (IVP): Given x_i for $i = 0, \ldots, k-1$, determine x_i satisfying (LDE) for $i \geq 0$.

Theorem. There exists a unique solution of (IVP) for (lh) or (li).

Proof. An obvious induction on i. The equation for $i = 0$ determines x_k, etc. □

Theorem. The solution set of (lh) is a k-dimensional vector space (a subspace of the set of all sequences $\{x_i\}_{i\geq 0}$).

Proof Sketch. For $j = 1, 2, \ldots k$, initialize the (LDE) sequence by setting

$$[x_0,\ x_1,\ \ldots,\ x_{k-1}]^T = e_j \in \mathbb{R}^k.$$

Then solving (lh) for each $j = 1, 2, \ldots k$ gives basis of the solution space of (lh). □

In (LDE) we may assume with no loss in generality that $a_0 \neq 0$. Indeed, if $\alpha_0 = 0$, (LDE) isn't really a k-step difference equation since we can shift indices and treat it as a $\widetilde{k}$-step difference equation for a $\widetilde{k} < k$, namely $\widetilde{k} = k - \nu$, where ν is the smallest index with $\alpha_\nu \neq 0$. Thus, henceforth we assume that $a_0 \neq 0$.

Let $r_1, \ldots, r_s$ be the distinct zeroes of p, with multiplicities $m_1, \ldots, m_s$ (note: each $r_j \neq 0$ since $\alpha_0 \neq 0$, and $m_1 + \cdots + m_s = k$.

Define the *characteristic polynomial* of (lh) to be

$$p(r) = r^k + \alpha_{k-1}r^{k-1} + \cdots + \alpha_0.$$

Let us assume that $\alpha_0 \neq 0$. Let $r_1, \ldots, r_s$ be the distinct zeroes of p, with multiplicities $m_1, \ldots, m_s$. Note that each $r_j \neq 0$ since $\alpha_0 \neq 0$, and $m_1 + \cdots + m_s = k$. These zeros generate the following basis of solutions of (lh):

$$\left\{ \{i^l r_j^i\}_{i=0}^{\infty} \,:\, 1 \leq j \leq s,\ 0 \leq l \leq m_j - 1 \right\}.$$

Example. Fibonacci Sequence:

$$F_{i+2} - F_{i+1} - F_i = 0, \quad F_0 = 0, \quad F_1 = 1.$$

The associated characteristic polynomial $r^2 - r - 1 = 0$ has roots

$$r_\pm = \frac{1 \pm \sqrt{5}}{2} \qquad (r_+ \approx 1.6, r_- \approx -0.6).$$

The general solution of (lh) is

$$F_i = C_+ \left(\frac{1+\sqrt{5}}{2}\right)^i + C_- \left(\frac{1-\sqrt{5}}{2}\right)^i.$$

Since $|r_-| < 1$, we have

$$C_- \left(\frac{1-\sqrt{5}}{2}\right)^i \to 0 \text{ as } i \to \infty.$$

The initial conditions $F_0 = 0$ and $F_1 = 1$ imply that $C_+ = \frac{1}{\sqrt{5}}$ and $C_- = -\frac{1}{\sqrt{5}}$. Hence, asymptotically, the Fibonacci sequence behaves like the sequence $\frac{1}{\sqrt{5}}\left(\frac{1+\sqrt{5}}{2}\right)^i$.

Remark. If $\alpha_0 = \alpha_1 = \cdots = \alpha_{\nu-1} = 0$ and $\alpha_\nu \neq 0$ (i.e., 0 is a root of multiplicity ν), then $x_0, x_1, \ldots, x_{\nu-1}$ are completely independent of x_i for $i \geq \nu$. So $x_{i+k} + \cdots + a_{i+\nu} = b_i$ for $i \geq 0$ with x_i given for $i \geq \nu$ behaves like a $(k-\nu)$-step difference equation.

Remark. Define $\widetilde{x}_i = [x_i,\ x_{i+1},\ \ldots,\ x_{i+k-1}]^T$. Then $\widetilde{x}_{i+1} = A\widetilde{x}_i$ for $i \geq 0$, where

$$A = \begin{bmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \\ -\alpha_0 & \ldots & & -\alpha_{k-1} \end{bmatrix},$$

and $\widetilde{x}_0 = \begin{bmatrix} x_0,\ x_1,\ \vdots,\ x_{k-1} \end{bmatrix}^T$ is given by the I.C. So (lh) is equivalent to the one-step vector difference equation

$$\widetilde{x}_{i+1} = A\widetilde{x}_i, \quad i \geq 0,$$

whose solution is $\widetilde{x}_i = A^i \widetilde{x}_0$. The characteristic polynomial of (lh) is the characteristic polynomial of A. Because A is a companion matrix, each distinct eigenvalue has only one Jordan block. If $A = PJP^{-1}$ is the Jordan decomposition of A (J in Jordan form, P invertible), then

$$\widetilde{x}_i = PJ^iP^{-1}\widetilde{x}_0.$$

Let J_j be the $m_j \times m_j$ block corresponding to r_j (for $1 \leq j \leq s$), so $J_j = r_j I + Z_j$, where Z_j denotes the $m_j \times m_j$ shift matrix:

$$Z_j = \begin{bmatrix} 0 & 1 & \\ \ddots & \ddots & 1 \\ & & 0 \end{bmatrix}.$$

Then

$$J_j^i = (r_j I + Z_j)^i = \sum_{l=0}^{i} \binom{i}{l} r_j^{i-1} Z_j^l.$$

Since $\binom{i}{l}$ is a polynomial in i of degree l and $Z_j^{m_j} = 0$, we see entries of the form (constant) $i^l r_j^i$ for $0 \le l \le m_j - 1$.

Remark. (li) becomes

$$\widetilde{x}_{i+1} = A\widetilde{x}_i + \widetilde{b}_i, \quad i \ge 0,$$

where $\widetilde{b}_i = [0, \ldots, 0, b_i]^T$. This leads to a discrete version of Duhamel's principle (exercise).

Remark. All solutions $\{x_i\}_{i\ge 0}$ of (lh) stay bounded (i.e. are elements of l^∞)

$\Leftrightarrow$ the matrix A is power bounded (i.e., $\exists\, M$ so that $\|A^i\| \le M$ for all $i \ge 0$)

$\Leftrightarrow$ the Jordan blocks $J_1, \ldots, J_s$ are all power bounded

$$\Leftrightarrow \begin{cases} \text{(a)} & \text{each } |r_j| \le 1 & (\text{for } 1 \le j \le s) \\ \text{and} \quad \text{(b)} & \text{for any } j \text{ with } m_j > 1 \text{ (multiple roots)}, & |r_j| < 1 \end{cases}.$$

If (a) and (b) are satisfied, then the map $\widetilde{x}_0 \mapsto \{x_i\}_{i\ge 0}$ is a bounded linear operator from $\mathbb{R}^k$ (or $\mathbb{C}^k$) into l^∞ (exercise).

Linear Multistep Methods (LMM)

A LMM is a method of the form

$$\sum_{j=0}^{k} \alpha_j x_{i+j} = h \sum_{j=0}^{k} \beta_j f_{i+j}, \quad i \ge 0$$

for the approximate solution of an ODE IVP

$$x' = f(t, x), \quad x(a) = x_a \ .$$

Here we want to approximate the solution $x(t)$ of this IVP for $a \le t \le b$ at the points $t_i = a + ih$ (where h is the time step), $0 \le j \le \frac{b-a}{h}$. The term x_i denotes the approximation to a solution of the IVP at t_i, $x(t_i)$. Similarly, f_{i+j} denotes $f(t_{i+j}, x_{i+j})$. We normalize the coefficients so that $\alpha_k = 1$. The above is called a *k-step LMM* whenever at least one of the coefficients α_0 and β_0 is non-zero. One can view the equation above as a difference equation, solving for x_{i+k} from $x_i, x_{i+1}, \ldots, x_{i+k-1}$. Assume as usual that f is continuous in (t, x) and uniformly Lipschitz in x. For simplicity of notation, assume that $x(t)$ is real and scalar; the anaysis that follows applies to $x(t) \in \mathbb{R}^n$ or $x(t) \in \mathbb{C}^n$ (viewed as $\mathbb{R}^{2n}$ for differentiability) with minor adjustments.

Example. (Midpoint rule) (explicit)

$$x(t_i + 2) - x(t_i) = \int_{t_i}^{t_{i+2}} x'(s)ds \approx 2hx'(t_{i+1}) = 2hf(t_{i+1}, x(t_{i+1})).$$

This approximate relationship suggests the LMM

$$\textit{Midpoint rule}: \qquad x_{i+2} - x_i = 2hf_{i+1} \ .$$

Example. (Trapezoid rule) (implicit)

The approximation

$$x(t_{i+1}) - x(t_i) = \int_{t_i}^{t_{i+1}} x'(s)ds \approx \frac{h}{2}(x'(t_{i+1}) + x'(t_i))$$

suggests suggests the LMM

$$\textit{Trapezoid rule}: \qquad x_{i+1} - x_i = \frac{h}{2}(f_{i+1} + f_i) \ .$$

Explicit vs Implicit.
If $\beta_k = 0$, the LMM is called *explicit*: once we know $x_i, x_{i+1}, \ldots, x_{i+k-1}$, we compute immediately

$$x_{i+k} = \sum_{j=0}^{k-1}(h\beta_j f_{i+j} - \alpha_j x_{i+j}) \ .$$

On the other hand, if $\beta_k \neq 0$, the LMM is called *implicit*: knowing $x_k, x_{i+1}, \ldots, x_{i+k-1}$, we need to

$$\text{solve} \quad x_{i+k} = h\beta_k f(t_{i+k}, x_{i+k}) - \sum_{j=0}^{k-1}(\alpha_j x_{i+j} - h\beta_j f(i+j) \quad \text{for} \quad x_{i+k} \ .$$

Remark. If h is sufficiently small, implicit LMM methods also have unique solutions given h and $x_0, x_1, \ldots, x_{k-1}$. To see this let L be the Lipschitz constant for f. Given $x_i, \ldots, x_{i+k-1}$, the value for x_{i+k} is obtained by solving the equation

$$x_{i+k} = h\beta_k f(t_{i+k}, x_{i+k}) + g_i,$$

where

$$g_i = \sum_{j=0}^{k-1}(h\beta_j f_{i+j} - \alpha_j x_{i+j})$$

is constant as far as x_{i_k} is concerned. That is, we are looking for a fixed point of

$$\Phi(x) = h\beta_k f(t_{i+k}, x) + g_i \ .$$

Note that if $h|\beta_k|L < 1$, then Φ is a contraction:

$$|\Phi(x) - \Phi(y)| \leq h|\beta_k| \, |f(t_{i+k}, x) - f(t_{i+k}, y)| \leq h|\beta_h|L|x - y|.$$

So by the Contraction Mapping Fixed Point Theorem, Φ has a unique fixed point. Any initial guess for x_{i+k} leads to a sequence converging to the fixed point using functional iteration

$$x_{i+k}^{(l+1)} = h\beta_k f(t_{i+k}, x_{i+k}^{(l)}) + g_i$$

which is initiated at some initial point $x_{i+k}^{(0)}$. In practice, one chooses to either

(1) iterate to convergence, or

(2) a fixed number of iterations.

In both approaches one typically uses an *explicit* method to get the initial guess $x_{i+k}^{(0)}$. This pairing is often called a Predictor-Corrector Method.

Function Evaluations. One FE means evaluating f once.

Explicit LMM: 1 FE per step (after initial start)

Implicit LMM: ? FEs per step if one iterates to convergence, and

usually 2 FE per step for a Predictor-Corrector Method.

Initial Values. To start a k-step LMM, we need $x_0, x_1, \ldots, x_{k-1}$. We usually take $x_0 = x_a$, and approximate $x_1, \ldots, x_{k-1}$ using a one-step method (e.g., a Runge-Kutta method).

Local Truncation Error. For a true solution $x(t)$ to $x' = f(t, x)$, define the LTE to be

$$l(h,t) = \sum_{j=0}^{k} \alpha_j x(t+jh) - h\sum_{j=0}^{k} \beta_j x'(t+jh).$$

If $x \in C^{p+1}$, then

$$\begin{aligned} x(t+jh) &= x(t) + jhx'(t) + \cdots + \frac{(jh)^p}{p!}x^{(p)}(t) + \mathcal{O}(h^{p+1}) \quad \text{and} \\ hx'(t+jh) &= hx'(t) + jh^2x''(t) + \cdots + \frac{j^{p-1}h^p}{(p-1)!}x^{(p)}(t) + \mathcal{O}(h^{p+1}) \end{aligned}$$

and so

$$l(h,t) = C_0x(t) + C_1hx'(t) + \cdots + C_ph^px^{(p)}(t) + \mathcal{O}(h^{p+1}),$$

where

$$\begin{aligned} C_0 &= \alpha_0 + \cdots + \alpha_k \\ C_1 &= (\alpha_1 + 2\alpha_2 + \cdots + k\alpha_k) - (\beta_0 + \cdots + \beta_k) \\ &\vdots \\ C_q &= \frac{1}{q!}(\alpha_1 + 2^q\alpha_2 + \cdots + k^q\alpha_k) - \frac{1}{(q-1)!}(\beta_1 + 2^{q-1}\beta_2 + \cdots + k^{q-1}\beta_k). \end{aligned}$$

Definition. A LMM is called *accurate of order* p if $l(h,t) = \mathcal{O}(h^{p+1})$ for any solution of $x' = f(t,x)$ which is C^{p+1}.

Fact. A LMM is accurate of order at least p iff $C_0 = C_1 = \cdots = C_p = 0$.

Remarks.

(i) For the LTE of a method to be $\mathcal{O}(h)$ for all f's, we must have $C_0 = C_1 = 0$: for any f which is C^1, all solutions $x(t)$ are C^2, so

$$l(h,t) = C_0 x(t) + C_1 h x'(t) + \mathcal{O}(h^2) \text{ is } \mathcal{O}(h) \quad \text{iff} \quad C_0 = C_1 = 0 \ .$$

(ii) Note that $C_0, C_1, \ldots$ depend only on $\alpha_0, \ldots, \alpha_k, \beta_0, \ldots, \beta_k$, not on f. So here, "minimal accuracy" is 1^{st}-order.

Definition. A LMM is called *consistent* if $C_0 = C_1 = 0$ (i.e., at least first-order accurate).

Remark. If a LMM is consistent, then any solution $x(t)$ for any f (continuous in (t,x), Lipschitz in x) has $l(h,t) = \mathcal{O}(h)$: since $x \in C^1$,

$$x(t+jh) = x(t) + jhx'(t) + \mathcal{O}(h) \quad \text{and} \quad hx'(t+jh) = hx'(t) + \mathcal{O}(h),$$

so

$$l(h,t) = C_0 x(t) + C_1 h x'(t) + \mathcal{O}(h).$$

Exercise: Verify that the $\mathcal{O}(h)$ terms converge to 0 uniformly in t (after dividing by h) as $h \to 0$: use the uniform continuity of $x'(t)$ on $[a,b]$.

Definition. A k-step LMM

$$\sum \alpha_j x_{i+j} = h \sum \beta_j f_{i+j}$$

is called *convergent* if for each IVP $x' = f(t,x)$, $x(a) = x_a$ on $[a,b]$ ($f \in (C, \text{Lip})$) and for any choice of $x_0(h), \ldots, x_{k-1}(h)$ for which

$$\lim_{h \to 0} |x(t_i(h)) - x_i(h)| = 0 \quad \text{for} \quad i = 0, \ldots, k-1,$$

we have

$$\lim_{h \to 0} \max_{\{i : a \le t_i(h) \le b\}} |x(t_i(h)) - x_i(h)| = 0 \ .$$

Remarks.

(i) This asks for *uniform* decrease in the error on grid as $h \to 0$.

(ii) By continuity of $x(t)$, the condition on the initial values is equivalent to $x_0(h) \to x_a$.

Fact. If a LMM is convergent, then the zeroes of the (first) characteristic polynomial of the method $p(r) = \alpha_k r^k + \cdots + \alpha_0$ satisfy the *Dahlquist root condition*:

(a) all zeroes r satisfy $|r| \leq 1$, and

(b) multiple zeroes r satisfy $|r| < 1$.

Example. (Zero Stability) Consider the IVP $x' = 0$, $0 \leq t \leq 1$, $x(0) = 0$, so $f \equiv 0$, and the LMM:

$$\sum \alpha_j x_{i+j} = 0 \ .$$

(1) Let r be any zero of $p(r)$. Then the solution with initial conditions

$$x_i = hr^i \quad \text{for} \quad 0 \leq i \leq k-1$$

is

$$x_i = hr^i \quad \text{for} \quad 0 \leq i \leq \frac{b-a}{h} \ .$$

Suppose $h = \frac{b-a}{m}$ for some $m \in \mathbb{Z}$. Then convergence implies that

$$x_m(h) \approx x(1) = 0.$$

But

$$x_m(h) = hr^m = \frac{b-a}{m} r^m.$$

So

$$|x_m(h) - x(1)| = \frac{b-a}{m}|r^m| \to 0 \quad \text{as} \quad m \to \infty$$

(i.e., $h \to 0$) iff $|r| \leq 1$.

(2) Similarly if r is a multiple zero of $p(r)$, taking $x_i(h) = hir^i$ for $0 \leq i \leq k-1$ gives

$$x_i(h) = hir^i \quad 0 \leq i \leq \frac{b-a}{h} \ .$$

So if $h = \frac{b-a}{m}$,

$$x_m(h) = \frac{b-a}{m} m r^m = (b-a)r^m,$$

so $x_m(h) \to 0$ as $h \to 0$ iff $|r| < 1$.

Definition. A LMM is called *zero-stable* if it satisfies the Dahlquist root condition.

Recall from our discussion of linear difference equations that zero-stability is equivalent to requiring that all solutions of $(lh) \quad \sum_{j=0}^{k} \alpha_j x_{i+j} = 0$ for $i \geq 0$ stay bounded as $i \to \infty$.

Remark. A consistent *one-step* LMM (i.e., $k = 1$) is always zero-stable. Indeed, consistency implies that $C_0 = C_1 = 0$ which in turn implies that $p(1) = \alpha_0 + \alpha_1 = C_0 = 0$ and so $r = 1$ is *the* zero of $p(r)$. Thus, in particular, $\alpha_1 = 1, \alpha_0 = -1$. That is $p(r) = r - 1$, and so LMM is zero-stable.

Exercise: Show that if an LMM is convergent, then it is consistent.

Theorem [LMM Convergence]
An LMM is convergent if and only if it is zero-stable and consistent. Moreover, for zero-stable methods, we get an error estimate of the form

$$\max_{a\le t_i(h)\le b} |x(t_i(h)) - x_i(h)| \le K_1 \underbrace{\max_{0\le i\le k-1} |x(t_i(h)) - x_i(h)|}_{\text{initial error}} + K_2 \underbrace{\frac{\max_i |l(h, t_i(h))|}{h}}_{\substack{\text{“growth of error”} \\ \text{controlled by} \\ \text{zero-stability}}}$$

Remark. If $x \in C^{p+1}$ and the LMM is accurate of order p, then

$$|LTE|/h = \mathcal{O}(h^p).$$

To get p^{th}-order convergence (i.e., $LHE = \mathcal{O}(h^p)$), we need

$$x_i(h) = x(t_i(h)) + \mathcal{O}(h^p) \quad \text{for} \quad i = 0, \ldots, k-1.$$

This can be done using $k-1$ steps of a RK method of order $\ge p-1$.

Lemma. Consider

$$(li) \qquad \sum_{j=0}^{k} \alpha_j x_{i+j} = b_i \quad \text{for} \quad i \ge 0 \qquad (\text{where } \alpha_k = 1),$$

with the initial values $x_0, \ldots, x_{k-1}$ given, and suppose that the characteristic polynomial $p(r) = \sum_{j=0}^{k} \alpha_j r^j$ satisfies the Dahlquist root condition. Then there is an $M > 0$ such that for $i \ge 0$,

$$|x_{i+k}| \le M \left(\max\{|x_0|, \ldots, |x_{k-1}|\} + \sum_{\nu=0}^{i} |b_\nu| \right).$$

Remark. Recall that the Dahlquist root condition implies that there is an $M > 0$ for which $\|A^k\|_\infty \le M$ for all $i \ge 0$, where

$$A = \begin{bmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \\ -\alpha_0 & & \vdots & -\alpha_{k-1} \end{bmatrix}$$

is the companion matrix for $p(r)$, and $\|\cdot\|_\infty$ is the operator norm induced by the vector norm $\|\cdot\|_\infty$.

Proof. Let $\widetilde{x}_i = [x_i,\ x_{i+1},\ \ldots,\ x_{i+k-1}]^T$ and $\widetilde{b}_i = [0,\ \ldots,\ 0,\ b_i]^T$. Then $\widetilde{x}_{i+1} = A\widetilde{x}_i + \widetilde{b}_i$, so by induction

$$\widetilde{x}_{i+1} = A^{i+1}\widetilde{x}_0 + \sum_{\nu=0}^{i} A^{i-\nu}\widetilde{b}_\nu.$$

Thus

$$\begin{aligned} |x_{i+k}| &\leq \|\widetilde{x}_{i+1}\|_\infty \\ &\leq \|A^{i+1}\|_\infty \|\widetilde{x}_0\|_\infty + \sum_{\nu=0}^{i} \|A^{i-\nu}\|_\infty \|\widetilde{b}_\nu\|_\infty \\ &\leq M(\|\widetilde{x}_0\|_\infty + \sum_{\nu=0}^{i} |b_\nu|). \end{aligned}$$

□

Proof of the LMM Convergence Theorem. The fact that convergence implies zero-stability and consistency has already been discussed. Next suppose a LMM is zero-stable and consistent. Let $x(t)$ be the true solution of the IVP $x' = f(t,x)$, $x(a) = x_a$ on $[a,b]$, let L be the Lipschitz constant for f, and

$$\beta = \max_{0\leq j\leq k} |\beta_j|.$$

Hold h fixed, and set

$$\begin{aligned} e_i(h) &= x(t_i(h)) - x_i(h), & E &= \max\{|e_0|, \ldots, |e_{k-1}|\}, \\ l_i(h) &= l(h, t_i(h)), & \lambda(h) &= \max_{0\leq i\leq \frac{b-a}{h}} |l_i(h)|. \end{aligned}$$

Step 1. The first step is to derive a "difference inequality" for $|e_i|$. The difference inequality refered to here is a discrete form of the integral inequality leading to Gronwall's inequality. For $i \in \mathcal{I}$, we have

$$\begin{aligned} \sum_{j=0}^{k} \alpha_j x(t_{i+j}) &= h \sum_{j=0}^{k} \beta_j f(t_{i+j}, x(t_{i+j})) + l_i \\ \sum_{j=0}^{k} \alpha_j x_{i+j} &= h \sum_{j=0}^{k} \beta_j f_{i+j}. \end{aligned}$$

Subtraction gives

$$\sum_{j=0}^{k} \alpha_j e_{i+j} = b_i,$$

where

$$b_i \equiv h \sum_{j=0}^{k} \beta_j (f(t_{i+j}, x(t_{i+j})) - f(t_{i+j}, x_{i+j})) + l_i.$$

Then

$$|b_i| \leq h \sum_{j=0}^{k} |\beta_j| L |e_{i+j}| + |l_i|.$$

So, by the preceeding Lemma with x_{i+k} replaced by e_{i+k}, for $i = 1, 2, \ldots,$

$$\begin{aligned}
|e_{i+k}| &\leq M[E + \sum_{\nu=0}^{i} |b_\nu|] \\
&\leq M\left[E + \sum_{\nu=0}^{i}\left(h\sum_{j=0}^{k}|\beta_j|L|e_{\nu+j}|\right) + |l_\nu|\right] \\
&= M[E + hL\beta\sum_{j=0}^{k}|e_{i+j}| + hL\beta\sum_{\nu=0}^{i-1}\sum_{j=0}^{k}|e_{\nu+j}| + \sum_{\nu=0}^{i}|l_\nu|\ ] \\
&\leq M\left[E + hL\beta|e_{i+k}| + hL\beta\sum_{\nu=0}^{k+i-1}|e_\nu| + \frac{b-a}{h}\lambda\right],
\end{aligned}$$

where the final inequality follows from the fact that $0 \leq i \leq \frac{(b-a)}{h}$. From here on, assume h is small enough to satisfy

$$Mh\beta L \leq \frac{1}{2}.$$

Since $\left\{h \leq b-a : Mh\beta L \geq \frac{1}{2}\right\}$ is a compact subset of $(0, b-a]$, the estimate in the Theorem is clearly true for those values of h. Moving $Mh\beta L|e_{i+k}|$ to the LHS gives

$$\begin{aligned}
|e_{i+k}| &\leq 2ME + 2M(b-a)\frac{\lambda}{h} + h(2ML\beta)\sum_{\nu=0}^{i+k-1}|e_\nu| \\
&= hM_1\sum_{\nu=0}^{i+k-1}|e_\nu| + (M_2E + M_3\lambda/h) \qquad\qquad i \in \mathcal{I}\ .
\end{aligned}$$

where $M_1 = 2ML\beta$, $M_2 = 2M$, and $M_3 = 2M(b-a)$. (Note: For explicit methods, $\beta_k = 0$, so we would not have to limit h, and the factor 2 can be dropped.)

Step 2. We now develop a discrete "comparison" argument to bound $|e_i|$. Let y_i be the solution of

$$(*) \qquad y_{i+k} = hM_1\sum_{\nu=0}^{i+k-1} y_\nu + (M_2E + M_3\lambda/h) \quad \text{for } i \in \mathcal{I},$$

with initial values $y_j = |e_j|$ for $0 \leq j \leq k-1$. Then clearly by induction $|e_{i+k}| \leq y_{i+k}$ for $i = 1, 2, \ldots$. Now

$$y_k \leq hM_1kE + (M_2E + M_3\ \lambda/h) \leq M_4E + M_3\lambda/h,$$

where $M_4 = (b-a)M_1k + M_2$. Subtracting $(*)$ for i from $(*)$ for $i+1$ gives

$$y_{i+k+1} - y_{i+k} = hM_1y_{i+k}, \quad \text{and so} \quad y_{i+k+1} = (1 + hM_1)y_{i+k}.$$

Therefore, by induction on $i \in \mathcal{I}$,

$$\begin{aligned} y_{i+k} &= (1 + hM_1)^i y_k \\ &\leq (1 + hM_1)^{(b-a)/h} y_k \\ &\leq e^{M_1(b-a)} y_k \\ &\leq K_1 E + K_2 \lambda / h, \end{aligned}$$

where $K_1 = e^{M_1(b-a)} M_4$ and $K_2 = e^{M_1(b-a)} M_3$. Thus, for $i \in \mathcal{I}$,

$$|e_{i+k}| \leq K_1 E + K_2 \lambda / h;$$

since $K_1 \geq M_4 \geq M_2 \geq M \geq 1$, also $|e_j| \leq E \leq K_1 E + K_2 \lambda / h$ for $0 \leq j \leq k-1$. Since consistency implies $\lambda = \mathcal{O}(h)$, we are done. □

Remarks.

(1) Note that K_1 and K_2 depend only on a, b, L, k, the α_j's and β_j's, and M.

(2) The estimate can be refined — we did not try to get the best constants K_1, K_2. For example, $e^{M_1(b-c)}$ could be replaced by $e^{M_1(t_i-a)}$ in both K_1 and K_2, yielding more precise estimates depending on i, similar to the estimate for one-step methods.