

Controlled Ordinary Differential Equations

M403 Lecture Notes by Philip D. Loewen

We deal with systems (natural, industrial, or mathematical) described by systems of ordinary differential equations (ODE's). Introducing extra state variables allows many of these to be written as systems of *first-order* ODE's: we always assume that this has been done.

Example. Find a first-order system equivalent to $\ddot{x}(t) + x(t) = u(t)$.

Solution. Let $x_1 := x$, $x_2 := \dot{x}$ to get $\begin{cases} \dot{x}_1 = x_2, \\ \dot{x}_2 = -x_1 + u(t). \end{cases}$ ////

Elements of \mathbb{R}^n are always understood as *column vectors*, or $n \times 1$ matrices. Our shorthand $x = (x_1, x_2)$ really means $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$, and the vector-matrix form of the equation in the example above is

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u(t).$$

In dealing with a controlled differential equation of the form $\dot{x}(t) = f(t, x(t), u(t))$, choosing a specific control function $u \in PWC([a, b]; \mathbb{R}^m)$ reduces the dynamics to the (possibly nonlinear) system

$$\dot{x}(t) = F(t, x(t)) \quad t \in [a, b], \quad x(a) = \xi,$$

where $F(t, x) := f(t, x, u(t)) \quad \forall (t, x) \in [a, b] \times \mathbb{R}^n$. Our first task is to review the existence, uniqueness, and stability of solutions to such initial value problems.

Since the mapping $x \mapsto F(t, x)$ carries n -vectors to n -vectors, its derivative is the $n \times n$ "Jacobian matrix" $D_x F(t, x)$, whose rows are the gradients of the component functions of F : for the column vector $F(t, x) = (F_1(t, x), F_2(t, x), \dots, F_n(t, x))$, we have

$$D_x F(t, x)_{ij} = \left[\frac{\partial F_i}{\partial x_j} \right], \quad i = 1, \dots, n; \quad j = 1, \dots, n.$$

The approximation law below holds for vectors h in \mathbb{R}^n :

$$F(t, x + h) = F(t, x) + D_x F(t, x)h + o(|h|), \quad \text{as } h \rightarrow 0.$$

A. Existence of Solutions

We study the vector initial-value problem

$$\dot{x}(t) = F(t, x(t)) \quad \forall t \in I; \quad x(\tau) = \xi. \quad (*)$$

A.1. Theorem. Given an open set $\Omega \subseteq \mathbb{R} \times \mathbb{R}^n$ and a function $F: \Omega \rightarrow \mathbb{R}^n$, assume that the Jacobian matrix $D_x F(t, x)$ exists at every point (t, x) in Ω , and that both F

and $D_x F$ are continuous on Ω . Then for each (τ, ξ) in Ω there exist an open interval $I(\tau, \xi) = (a(\tau, \xi), b(\tau, \xi))$ containing τ and a function $x(\cdot; \tau, \xi): I(\tau, \xi) \rightarrow \mathbb{R}^n$ such that

- (i) [Existence] The choices $x(t) = x(t; \tau, \xi)$ and $I = I(\tau, \xi)$ obey $(*)$ above.
- (ii) [Uniqueness and Maximality] Whenever I is an open interval containing τ and $x: I \rightarrow \mathbb{R}^n$ is a function obeying $(*)$, one has both

$$I \subseteq I(\tau, \xi) \quad \text{and} \quad x(t) = x(t; \tau, \xi) \quad \forall t \in I.$$

- (iii) [Continuous Dependence] The set $G := \{(t, \tau, \xi) : (\tau, \xi) \in \Omega, t \in I(\tau, \xi)\}$ is open in $\mathbb{R} \times \mathbb{R} \times \mathbb{R}^n$, and $x(\cdot; \cdot, \cdot): G \rightarrow \mathbb{R}^n$ is continuous at each point of G .
- (iv) [Qualitative Properties] If F is bounded on Ω , then one has either $a(\tau, \xi) = -\infty$, or $\alpha \stackrel{\text{def}}{=} \lim_{t \downarrow a(\tau, \xi)} x(t; \tau, \xi)$ exists and $(a, \alpha) \in \text{bdy } \Omega$; and either $b(\tau, \xi) = +\infty$, or $\beta \stackrel{\text{def}}{=} \lim_{t \uparrow b(\tau, \xi)} x(t; \tau, \xi)$ exists and $(b, \beta) \in \text{bdy } \Omega$.

Proof. The following texts contain accessible proofs.

Jack K. Hale, *Ordinary Differential Equations*, New York: Wiley-Interscience, 1969.
Chapter I, Lemma 2.1, Thm. 2.1, and Thm. 3.1.

Earl A. Coddington, *An Introduction to Ordinary Differential Equations*, Englewood Cliffs: Prentice-Hall, 1961.
Chapters 5–6. ////

Sketch.

A.2. Remarks. In Thm. A.1, the assumption that $D_x F$ is continuous on Ω can be removed in favour of the weaker condition below:

for each bounded set $K \subseteq \Omega$ there is a constant M (depending on K) such that

$$|F(t, y) - F(t, x)| \leq M|y - x| \quad \forall (t, x), (t, y) \in K.$$

This is called a Lipschitz condition. It is weaker than the continuous differentiability hypothesis because that condition implies that for any compact convex subset K of Ω , the Jacobian $D_x F(t, x)$ is bounded on K . This, together with the Mean-Value Theorem, makes for an easy proof of the inequality above.

A.3. Examples. (a) $\dot{x}(t) = x(t)^2$, $x(0) = 1$.

The function $F(t, x) = x^2$ obeys the hypotheses of Thm. A.1 for any open set $\Omega \subseteq \mathbb{R}^2$. For the rectangle $\Omega = (-10, 10) \times (-1000, 1000)$, we find the unique solution $x(t; 0, 1) = 1/(1-t)$, with interval of existence $I(0, 1) = (-10, 999/1000)$. Note that (iv) holds:

$$\begin{aligned} a(0; 1) = -10 \quad \text{and} \quad \left(-10, \lim_{t \rightarrow -10^+} \frac{1}{1-t} \right) &\in \text{bdy } \Omega \\ b(0; 1) = \frac{999}{1000} \quad \text{and} \quad \left(\frac{999}{1000}, \lim_{t \rightarrow \frac{999}{1000}} \frac{1}{1-t} \right) &\in \text{bdy } \Omega. \end{aligned}$$

Draw pictures.

(b) $\dot{x}(t) = 3x(t)^{2/3}$, $x(0) = 0$.

Here $F(t, x) = x^{2/3}$ disobeys the hypotheses of Thm. A.1 on any open set Ω containing $(\tau, \xi) = (0, 0)$. Indeed, the Jacobian matrix is the 1×1 matrix (just a number) $D_x F(t, x) = \frac{2}{3}x^{-1/3}$, which is evidently not continuous at the point $x = 0$. The Lipschitz condition of Remark A.2 fails too, since

$$|F(t, x) - F(t, 0)| = 3|x|^{2/3} = \frac{3}{|x|^{1/3}}|x|$$

cannot be majorized by $M|x - 0|$ on any neighbourhood of $\xi = 0$.

Since F is jointly continuous, a deeper result known as Peano's theorem asserts that this initial-value problem has a solution. In fact, there are *two* solutions: $x(t) = 0$ and $x(t) = t^3$. Hence condition (b) is essential if unique solutions are desired. ////

A Practical Extension.

In control problems, discontinuities in the input function $u(\cdot)$ often result in a function $F(t, x) = f(t, x, u(t))$ which is not continuous, but which can be treated as follows. Assume that (a, b) is a finite interval.

A.4. Theorem. *Let D be an open subset of \mathbb{R}^n , and let $\Omega = (a, b) \times D$. Suppose the interval $[a, b]$ admits a finite partition*

$$a = a_0 < a_1 < a_2 < \cdots < a_N = b,$$

relative to which the function $F: \Omega \rightarrow \mathbb{R}^n$ has the following property: there exists some $\varepsilon > 0$ such that the definition of the function F can be extended to each cylinder

$\Omega_j(\varepsilon) := (a_{j-1} - \varepsilon, a_j + \varepsilon) \times D$ in such a way that the result obeys the hypotheses of Theorem A.1 on $\Omega_j(\varepsilon)$. Then all the conclusions of Thm. A.1 hold, except that in (*) we only get equality on $I(\tau, \xi) \setminus \{a_j\}$ because $\dot{x}(t)$ may be undefined at the partition points.

Proof. Apply Thm. A.1 N times. ////

B. Linear Systems with Constant Coefficients

ODE systems where $F(t, x) = Ax$ are significant in applications and—at least in low-dimensional cases—explicitly solvable. Thus we look in detail at equations like

$$\dot{x}(t) = Ax(t), \quad x(0) = \xi, \quad (B.1)$$

where A is a given $n \times n$ matrix with (constant) real entries, and the solution function $x(t; \xi)$ evolves in \mathbb{R}^n . It can be shown that for each initial point ξ , the function $x(\cdot; \xi)$ is defined on the whole real line. Moreover, for any $c_1, c_2 \in \mathbb{R}$ and $\xi, \eta \in \mathbb{R}^n$, the arc

$$y(t) \stackrel{\text{def}}{=} c_1 x(t; \xi) + c_2 x(t; \eta)$$

obeys $\dot{y} = Ay$ with $y(0) = c_1 \xi + c_2 \eta$, and hence deserves the notation $x(t; c_1 \xi + c_2 \eta)$. This shows that

$$x(t; c_1 \xi + c_2 \eta) = c_1 x(t; \xi) + c_2 x(t; \eta) \quad \forall c_1, c_2 \in \mathbb{R}, \quad \xi, \eta \in \mathbb{R}^n.$$

In other words, for each fixed $t \in \mathbb{R}$, the mapping $\xi \mapsto x(t; \xi)$ is a linear transformation from \mathbb{R}^n into \mathbb{R}^n . It follows that there is some $n \times n$ matrix $E(t)$ for which

$$x(t; \xi) = E(t)\xi \quad \forall \xi \in \mathbb{R}^n. \quad (B.12)$$

When $\xi = \widehat{\mathbf{e}}_k$ (the k -th standard unit vector), the product $E(t)\widehat{\mathbf{e}}_k$ selects the k -th column of $E(t)$. Hence we may write

$$E(t) = [x(t; \widehat{\mathbf{e}}_1) \mid x(t; \widehat{\mathbf{e}}_2) \mid \cdots \mid x(t; \widehat{\mathbf{e}}_n)]. \quad (B.15)$$

Clearly $E(0) = [\widehat{\mathbf{e}}_1 \mid \widehat{\mathbf{e}}_2 \mid \cdots \mid \widehat{\mathbf{e}}_n] = I$, while differentiation gives

$$\begin{aligned} \dot{E}(t) &= [\dot{x}(t; \widehat{\mathbf{e}}_1) \mid \dot{x}(t; \widehat{\mathbf{e}}_2) \mid \cdots \mid \dot{x}(t; \widehat{\mathbf{e}}_n)] \\ &= [Ax(t; \widehat{\mathbf{e}}_1) \mid Ax(t; \widehat{\mathbf{e}}_2) \mid \cdots \mid Ax(t; \widehat{\mathbf{e}}_n)] \\ &= A [x(t; \widehat{\mathbf{e}}_1) \mid x(t; \widehat{\mathbf{e}}_2) \mid \cdots \mid x(t; \widehat{\mathbf{e}}_n)] = AE(t). \end{aligned}$$

It follows that the matrix $E(t)$ assembled above is the unique solution of the following initial-value problem with square matrix unknown:

$$\frac{dE}{dt} = AE, \quad t \in \mathbb{R}; \quad E(0) = I. \quad (B.17)$$

Analogy with this problem for the scalar case (where the IVP $y' = ay$, $y(0) = 1$ has unique solution $y = e^{at}$) is just one reason for preferring the notation $e^{At} = E(t)$ for this function. (Later developments will reveal deeper reasons.)

Calculating e^{At} —Four Methods. From this point until the end of this section, all our calculations are done in the complex number system \mathbb{C} . Even when the given matrix A has only real entries, complex values may arise during the calculations that follow.

Method I—Direct, using a high-order scalar ODE. Use algebraic substitutions to transform the system of first-order equations in (B.1) into a single n -th order equation for one of the component functions. Solve that by known methods, then use the result to find the general solution for (B.1) in vector form. This will involve n arbitrary constants. For each unit vector $\hat{\mathbf{e}}_k$, find the solution $x(t; \hat{\mathbf{e}}_k)$ by choosing these constants appropriately, then use (B.15):

$$e^{At} = [x(t; \mathbf{e}_1) \mid x(t; \mathbf{e}_2) \mid \cdots \mid x(t; \mathbf{e}_n)].$$

B.1. Example. Find e^{At} for the matrix $A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$.

Solution. When $x = (x_1, x_2)$, the differential equation $\dot{x} = Ax$ encodes the system

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_2 \\ -x_1 \end{bmatrix}. \quad (*)$$

Since this equation implies $\ddot{x}_1 + x_1 = 0$, every solution has the form

$$x_1(t) = a \cos t + b \sin t, \quad x_2(t) = \dot{x}_1(t) = b \cos t - a \sin t, \quad a, b \in \mathbb{R}.$$

The initial condition $x(0) = \mathbf{e}_1$ gives $a = 1, b = 0$, and hence the solution $x(t) = (\cos t, -\sin t)$; the initial condition $x(0) = \mathbf{e}_2$ gives $a = 0, b = 1$, and hence the solution $x(t) = (\sin t, \cos t)$. These form the columns of the desired matrix exponential:

$$e^{At} = \begin{bmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{bmatrix}. \quad ////$$

B.2. Example. Find $e^{\Lambda t}$ for the diagonal matrix $\Lambda = \begin{bmatrix} \lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots & 0 \\ 0 & 0 & \lambda_3 & \cdots & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_n \end{bmatrix}$.

Solution. The system $\dot{x} = \Lambda x$ encodes n uncoupled ODE's:

$$\dot{x}_1 = \lambda_1 x_1, \quad \dot{x}_2 = \lambda_2 x_2, \quad \dots, \quad \dot{x}_n = \lambda_n x_n.$$

The general solution involves independent constants c_1, \dots, c_n :

$$x_1(t) = c_1 e^{\lambda_1 t}, \quad x_2(t) = c_2 e^{\lambda_2 t}, \quad \dots, \quad x_n(t) = c_n e^{\lambda_n t}.$$

Note that $x(0) = (x_1(0), \dots, x_n(0)) = (c_1, \dots, c_n) \stackrel{\text{def}}{=} c$, so we have

$$x(t; c) = \begin{bmatrix} e^{\lambda_1 t} & 0 & 0 & \cdots & 0 \\ 0 & e^{\lambda_2 t} & 0 & \cdots & 0 \\ 0 & 0 & e^{\lambda_3 t} & \cdots & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & e^{\lambda_n t} \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ \vdots \\ c_n \end{bmatrix}, \quad c \in \mathbb{R}^n.$$

Comparison with line (B.12) reveals that the matrix on the right is precisely e^{At} . In compressed notation,

$$\exp(t \operatorname{diag}(\lambda_1, \dots, \lambda_n)) = \operatorname{diag}(e^{\lambda_1 t}, \dots, e^{\lambda_n t}).$$

Method II—Direct, Using Eigenvalues. In a naïve approach to (B.1), based on analogy with the scalar case, one might guess that solutions have some kind of exponential time-dependence, and plug in a trial solution of the form $x(t) = e^{\lambda t} v$ for some constant λ and nonzero vector v , both to be determined. Substitution then yields the equation $Av = \lambda v$, which holds for a nonzero vector v if and only if v is an eigenvector for A with eigenvalue λ . If the matrix A has a full set of n linearly independent eigenvectors v_1, \dots, v_n , with corresponding eigenvalues $\lambda_1, \dots, \lambda_n$, then every linear combination below is a solution of (B.1):

$$x(t) = c_1 e^{\lambda_1 t} v_1 + c_2 e^{\lambda_2 t} v_2 + \dots + c_n e^{\lambda_n t} v_n, \quad c_1, \dots, c_n \in \mathbb{C}. \quad (B.2)$$

This general solution can be put to work as in Method I, or given a matrix-theoretic treatment as follows. To arrange the initial condition $x(0) = \xi$, we need only determine the coefficients c_1, \dots, c_n such that $\xi = c_1 v_1 + c_2 v_2 + \dots + c_n v_n$. Since the number n of linearly independent eigenvectors equals the dimension of the state space \mathbb{R}^n , this is possible for any ξ . Indeed, the desired equation for ξ can be written as

$$\xi = [v_1 | v_2 | \cdots | v_n] \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix} = Vc,$$

where V is a square matrix whose columns are the linearly independent eigenvectors just discussed, and c is the vector of coefficients. It follows that $c = V^{-1}\xi$, so that the solution to (B.1) with initial condition $x(0) = \xi$ can be written

$$x(t; \xi) = V \begin{bmatrix} e^{\lambda_1 t} & 0 & \cdots & 0 \\ 0 & e^{\lambda_2 t} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e^{\lambda_n t} \end{bmatrix} V^{-1}\xi. \quad (B.3)$$

Now compare line (B.12): *the square matrix in front of the vector ξ here is precisely e^{At} .* Even when the eigenvalues, eigenvectors, and coefficients turn out to be complex,

each solution generated in this way will turn out to be real-valued, and will provide one column of the desired matrix e^{At} .

It's nice to note that if $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$ is the diagonal matrix of eigenvalues for A , listed in the same order as the eigenvectors that form the columns of matrix V , then we have

$$A = V\Lambda V^{-1}, \quad e^{At} = Ve^{At}V^{-1}.$$

(See Example B.2 above, and Method IV below.)

Here is the same example considered above, solved by applying equation (B.3).

B.3. Example. Find e^{At} for the matrix $A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$.

Solution. The eigenvalues of A are $\pm i$. An eigenvector corresponding to $\lambda_1 = -i$ is $v_1 = (1, -i)$; an eigenvector corresponding to $\lambda_2 = i$ is $v_2 = (1, i)$. Thus an eigenvector matrix and its inverse are

$$V = [v_1|v_2] = \begin{bmatrix} 1 & 1 \\ -i & i \end{bmatrix}, \quad V^{-1} = \begin{bmatrix} 1/2 & i/2 \\ 1/2 & -i/2 \end{bmatrix}.$$

It follows that, with $\Lambda = \begin{bmatrix} -i & 0 \\ 0 & i \end{bmatrix}$,

$$e^{At} = Ve^{At}V^{-1} = \begin{bmatrix} 1 & 1 \\ -i & i \end{bmatrix} \begin{bmatrix} e^{it} & 0 \\ 0 & e^{-it} \end{bmatrix} \begin{bmatrix} 1/2 & i/2 \\ 1/2 & -i/2 \end{bmatrix} = \begin{bmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{bmatrix}.$$

////

Method III—Laplace Transforms. Take Laplace Transforms of both sides in (B.17), writing $\tilde{E} = \mathcal{L}(E)$:

$$s\tilde{E}(s) - E(0) = A\tilde{E}(s), \quad E(0) = I \iff [sI - A]\tilde{E}(s) = I \iff \tilde{E}(s) = [sI - A]^{-1}.$$

Recover e^{At} as the inverse Laplace transform of the matrix $(sI - A)^{-1}$:

$$e^{At} = \mathcal{L}^{-1} \{ [sI - A]^{-1} \}.$$

(One computes the inverse transform component-by-component.) This approach is not very attractive, but most of the engineering literature on control theory is dominated by calculations involving Laplace transforms, and the formulas just given provide a link to that.

Method IV—Power Series. We can extend the definitions of common single-input/single-output functions to allow for square matrix inputs and outputs. There is no trouble with polynomials, of course: given $p(x) = a_0 + a_1x + \dots + a_mx^m$, it

seems only natural to set $p(M) = a_0I + a_1M + \dots + a_mM^m$. From here it is a short step to analytic functions (“infinite polynomials”): suppose a function g is given, with the series representation

$$g(z) = a_0 + a_1z + a_2z^2 + a_3z^3 + \dots, \quad |z| < R.$$

Then for any complex $n \times n$ matrix M , all of whose eigenvalues λ satisfy $|\lambda| < R$, one can simply define

$$g(M) = a_0I + a_1M + a_2M^2 + a_3M^3 + \dots \quad (B.4)$$

(Convergence of the series is assessed componentwise.)

Computing an analytic function of a matrix argument using the definition above is quite manageable when the matrix in question is diagonalizable. Recall that when the given $n \times n$ matrix M has a full set of n linearly independent eigenvectors v_1, \dots, v_n , with corresponding eigenvalues $\lambda_1, \dots, \lambda_n$, stacking the eigenvectors side by side to form the columns of an “eigenvector matrix” $V = [v_1|v_2|\dots|v_n]$ reveals that

$$MV = [\lambda_1v_1|\lambda_2v_2|\dots|\lambda_nv_n] = [v_1|v_2|\dots|v_n] \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_n \end{bmatrix},$$

so that

$$M = V^{-1}\Lambda V,$$

where $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$ is the diagonal matrix of eigenvalues for M , *written in the same order as the eigenvectors appear in V* . One advantage of this diagonal form is that internal cancellations allow one to evaluate $M^s = V^{-1}\Lambda^kV$ for any integer $k \geq 0$, provided one interprets $M^0 = I$. Using this observation in each term of the series (B.4), we deduce that

$$g(M) = g(V^{-1}\Lambda V) = V^{-1}g(\Lambda)V = V^{-1} \begin{bmatrix} g(\lambda_1) & 0 & \dots & 0 \\ 0 & g(\lambda_2) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & g(\lambda_n) \end{bmatrix} V. \quad (B.5)$$

Thus it is quite a simple matter to compute $\sin(M)$, $\cos(M)$, $\tan^{-1}(M)$ and even $\log(I+M)$ for any diagonalizable matrix M whose eigenvalues are small enough. The most practical applications of this capability are probably first through the resolvent formula (“geometric series of matrices”)

$$(I - M)^{-1} = I + M + M^2 + M^3 + \dots,$$

valid for matrices M whose eigenvalues all obey $|\lambda| < 1$ (and used in diverse fields from functional analysis through geophysics), and in our case through the all-important exponential formula

$$e^M = \exp(M) = I + M + \frac{M^2}{2!} + \frac{M^3}{3!} + \dots, \quad (B.6)$$

valid for any square matrix M at all. This last line makes the connection between line (B.3), derived using the operational definition, and line (B.5), developed from the series method. It is important to note that line (B.6) always provides the correct matrix exponential for M , even when M is not diagonalizable and some method more complicated than (B.5) must be used to calculate it.

B.3. Proposition. *For any $n \times n$ matrix M , define e^M by (B.6). Then*

(a) e^M is invertible, with $(e^M)^{-1} = e^{-M}$; also, $e^{0M} = I$;

(b) $Me^M = e^M M$;

(c) the matrix-valued function $t \mapsto e^{Mt}$ is differentiable, entry-by-entry, and

$$\frac{d}{dt}(e^{Mt}) = Me^{Mt} = e^{Mt}M;$$

in particular, the function $x(t) = e^{At}\xi$ solves (B.1);

(d) for an $n \times n$ matrix N , one has

$$MN = NM \quad \Longrightarrow \quad e^M e^N = e^{M+N} = e^N e^M.$$

There are (non-commutative) matrix pairs for which these equations fail.

Degeneracy. Formulas (B.3) and (B.5) require that the matrix A have a full set of n linearly independent eigenvectors. Not all matrices have this property. For example, fix any real α and consider

$$J = \begin{bmatrix} \alpha & 1 & 0 \\ 0 & \alpha & 1 \\ 0 & 0 & \alpha \end{bmatrix}.$$

Clearly $\det(J - \lambda I) = (\alpha - \lambda)^3$, so $\lambda = \alpha$ is an eigenvalue of multiplicity three. However, the matrix

$$J - \alpha I = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

multiplies to give $\mathbf{0}$ only on vectors of the form $c\mathbf{i}$, $c \in \mathbb{R}$. So there aren't enough eigenvectors to build e^{Jt} by diagonalization. Start over with the differential equation: look for solutions \mathbf{x} of $\dot{\mathbf{x}} = J\mathbf{x}$ having form $\mathbf{x}(t) = e^{\alpha t}\mathbf{u}(t)$, with \mathbf{u} the new unknown. Then substitution gives

$$\alpha\mathbf{u} + \dot{\mathbf{u}} = J\mathbf{u}, \quad \text{i.e.,} \quad \dot{\mathbf{u}} = (J - \alpha I)\mathbf{u}.$$

The solution will be

$$\mathbf{u}(t) = e^{(J-\alpha I)t}\mathbf{u}(0) = e^{(J-\alpha I)t}\mathbf{x}(0),$$

and we combine

$$e^z = 1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + \cdots$$

$$(J - \alpha I)^2 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad (J - \alpha I)^k = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad \text{for } k \geq 3,$$

to get

$$\mathbf{u}(t) = \left(\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} + t \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} + \frac{t^2}{2!} \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \right) \mathbf{x}(0)$$

$$\mathbf{x}(t) = e^{\alpha t} \begin{bmatrix} 1 & t & t^2/2! \\ 0 & 1 & t \\ 0 & 0 & 1 \end{bmatrix} \mathbf{x}(0)$$

Uniqueness produces case $p = 3$ of a fact about general $p \times p$ matrices:

$$J = \begin{bmatrix} \alpha & 1 & 0 & \cdots & 0 & 0 \\ 0 & \alpha & 1 & \cdots & 0 & 0 \\ \vdots & & & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \alpha & 1 \\ 0 & 0 & 0 & \cdots & 0 & \alpha \end{bmatrix}$$

$$\implies \exp(Jt) = e^{\alpha t} \begin{bmatrix} 1 & t & t^2/2! & t^3/3! & \cdots & t^{p-1}/(p-1)! \\ 0 & 1 & t & t^2/2! & \cdots & t^{p-2}/(p-2)! \\ 0 & 0 & 1 & t & \cdots & t^{p-3}/(p-3)! \\ \vdots & & & & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & t \\ 0 & 0 & 0 & 0 & \cdots & 1 \end{bmatrix}.$$

A key theorem of linear algebra (Jordan Canonical Form) asserts that this situation is typical. Every $n \times n$ matrix A admits a square matrix V (with entries possibly complex), whose columns are “generalized eigenvectors”, such that

$$A = VJV^{-1},$$

where the matrix J is no longer simply diagonal, but rather “block diagonal”:

$$J = \begin{bmatrix} J_1 & & & & \\ & J_2 & & & \\ & & \ddots & & \\ & & & J_r & \end{bmatrix}, \quad \text{where } J_k = \begin{bmatrix} \lambda_k & 1 & 0 & \cdots & 0 \\ 0 & \lambda_k & 1 & \cdots & 0 \\ \vdots & & & \ddots & \vdots \\ 0 & \cdots & & \lambda_k & 1 \\ 0 & \cdots & & \cdots & \lambda_k \end{bmatrix},$$

and each J_k , $k = 1, \dots, r$, is a “Jordan block”—i.e., a square matrix of size $d_r \times d_r$, with the eigenvalue λ_k on the main diagonal and the number 1 on the superdiagonal.

(The same eigenvalue may contribute more than one Jordan block.) In the case where $r = n$ and $d_k = 1$ for all k , each Jordan block is simply a scalar, and we are back in the diagonalizable case covered by equation (B.5). In general, however, the matrix exponential will look like

$$e^{At} = V \begin{bmatrix} e^{J_1 t} & & & \\ & e^{J_2 t} & & \\ & & \ddots & \\ & & & e^{J_r t} \end{bmatrix} V^{-1},$$

where each $e^{J_k t}$ has the banded upper-triangular structure derived above.

Stability. The point of all this is to deal with questions of stability for the linear differential equation $\dot{\mathbf{x}} = A\mathbf{x}$, where A is an $n \times n$ matrix with real entries. Let us write $\sigma(A)$ for the set of all eigenvalues of A . This set contains at most n complex numbers, in complex-conjugate pairs. (I.e., $\lambda \in \sigma(A)$ if and only if $\bar{\lambda} \in \sigma(A)$.)

B.4. Theorem. For every initial vector ξ in \mathbb{R}^n , the IVP

$$\dot{\mathbf{x}} = A\mathbf{x}, \quad \mathbf{x}(0) = \xi \tag{*}$$

has a unique solution $x(t; \xi)$, defined for all $t \in \mathbb{R}$.

- (a) If $\Re(\lambda) < 0$ for every $\lambda \in \sigma(A)$, then $x(t; \xi) \rightarrow 0$ as $t \rightarrow \infty$ for every $\xi \in \mathbb{R}^n$.
- (b) If $\Re(\lambda) > 0$ for one or more $\lambda \in \sigma(A)$, then every open ball $\mathbb{B}(0; \varepsilon)$, $\varepsilon > 0$, contains some ξ for which $|x(t; \xi)| \rightarrow +\infty$ as $t \rightarrow \infty$.
- (c) If $\Re(\lambda) = 0$ for one or more $\lambda \in \sigma(A)$, then every open ball $\mathbb{B}(0; \varepsilon)$, $\varepsilon > 0$, contains some ξ for which $\inf_{t>0} |x(t; \xi)| > 0$.

Proof. We treat complex-valued solutions only. Deriving the stated consequences for real-valued solutions is an exercise.

(a) Every solution of (*) is a linear combination of n complex-vector-valued functions, each of the form $t^k e^{\lambda t} v$ for some integer exponent $k \in [0, n)$, constant vector v , and eigenvalue λ . The magnitude of such a term for $t > 0$ is

$$|t^k e^{\lambda t} v| = t^k e^{t \Re \lambda} |v|,$$

which converges to 0 as $t \rightarrow \infty$ because $\Re \lambda < 0$ by hypothesis. (This is a routine application of L'Hospital's rule and induction.)

(b) Let $\lambda = \sigma + i\omega$ be an eigenvalue of A with $\sigma > 0$, and with corresponding eigenvector v . Then for any $c > 0$, no matter how small, the function

$$x(t; cv) = ce^{(\sigma+i\omega)t} v$$

has $|x(t; cv)| = ce^{\sigma t} |v| \rightarrow +\infty$ as $t \rightarrow \infty$. Given $\varepsilon > 0$, we can put the initial point $\xi = cv$ into $\mathbb{B}(0; \varepsilon)$ just by choosing $c = \varepsilon/2 > 0$.

(c) Put $\sigma = 0$ in the argument of (b). Then the complex-valued solution $x(t; cv) = ce^{i\omega t} v$, has constant positive magnitude. ////

C. Linear Systems with Variable Coefficients

read@home

Now consider the linear system

$$\dot{x}(t) = A(t)x(t), \quad x(\tau) = \xi, \quad (C.1)$$

where the $n \times n$ matrix $A(\cdot)$ varies with time. Assume that $A(\cdot)$ is piecewise continuous: then (C.1) has a unique solution, denoted $x(\cdot; \tau, \xi)$. Since the differential equation in (C.1) is linear, we have

$$x(\cdot; \tau, a_1\xi_1 + a_2\xi_2) = a_1x(\cdot; \tau, \xi_1) + a_2x(\cdot; \tau, \xi_2) \quad \forall a_1, a_2 \in \mathbb{R}, \xi_1, \xi_2 \in \mathbb{R}^n.$$

It follows just as in the autonomous case that for each fixed pair t, τ , some $n \times n$ matrix $\Phi(t; \tau)$ obeys

$$x(t; \tau, \xi) = \Phi(t; \tau)\xi \quad \forall \xi \in \mathbb{R}^n. \quad (C.1a)$$

This is called *the fundamental matrix* associated with $A(\cdot)$, and by linearity, it can be constructed as

$$\Phi(t; \tau) := [x(t; \tau, \mathbf{e}_1) | x(t; \tau, \mathbf{e}_2) | \cdots | x(t; \tau, \mathbf{e}_n)].$$

Here are some of its elementary properties:

- (a) For each τ , the unique solution of this IVP for an unknown matrix E is precisely $E(t) = \Phi(t; \tau)$:

$$\frac{d}{dt}E(t) = A(t)E(t), \quad E(\tau) = I.$$

Proof: $\Phi(\tau; \tau) = [x(\tau; \tau, \mathbf{e}_1) | x(\tau; \tau, \mathbf{e}_2) | \cdots | x(\tau; \tau, \mathbf{e}_n)] = [\mathbf{e}_1 | \mathbf{e}_2 | \cdots | \mathbf{e}_n] = I$, and

$$\begin{aligned} \frac{d}{dt}\Phi(t; \tau) &= [\dot{x}(t; \tau, \mathbf{e}_1) | \cdots | \dot{x}(t; \tau, \mathbf{e}_n)] \\ &= A(t)[x(t; \tau, \mathbf{e}_1) | \cdots | x(t; \tau, \mathbf{e}_n)] = A(t)\Phi(t; \tau). \end{aligned}$$

Uniqueness follows by applying Thm. A.1 to each column separately. ////

- (b) If $A \in \mathbb{R}^{n \times n}$ is a constant matrix then $\Phi(t; \tau) = e^{A(t-\tau)}$.

Proof: Let $E(t) = e^{A(t-\tau)}$. From Section B, $E(t) = e^{A(t-\tau)} = e^{At}e^{-A\tau}$, so

$$\dot{E}(t) = Ae^{At}e^{-A\tau} = AE(t) \quad \text{and} \quad E(\tau) = e^{A\tau}e^{-A\tau} = I.$$

The result follows from (a).

- (c) $x(t; \tau, \xi) = \Phi(t; \tau)\xi$, for all t, τ, ξ .

This just restates (C.1a), but we should mention a nice interpretation: the matrix $\Phi(t; \tau)$ picks up a point ξ and carries it along the solution curve of the ODE (C.1) to the point $x(t; \tau, \xi)$ it reaches in elapsed time $t - \tau$.

- (d) $\Phi(t; s)\Phi(s; r) = \Phi(t; r)$ for all r, s , and t .

Proof: It suffices to show that the indicated matrices have identical actions on every vector $\xi \in \mathbb{R}^n$. Start on the left:

$$\begin{aligned}\Phi(t; s)\Phi(s; r)\xi &= \Phi(t; s)x(s; r, \xi) \quad \text{by (c)} \\ &= x(t; s, x(s; r, \xi)) \quad \text{by (c)} \\ &= x(t; r, \xi) \quad \text{obvious property of ODE} \\ &= \Phi(t; r)\xi \quad \text{by (c)}.\end{aligned}$$

(e) In particular, $\Phi(t; \tau)$ is invertible with $\Phi(t; \tau)^{-1} = \Phi(\tau; t)$.

(f) The fundamental matrix corresponding to $-A(t)^T$ is $\Psi(t; \tau) := \Phi(\tau; t)^T$.

Proof: Let $\Psi(t; \tau)$ be the fundamental matrix for $-A(t)^T$. Then for any vectors $v, w \in \mathbb{R}^n$ and any times $r, s \in \mathbb{R}$,

$$\begin{aligned}x(t) = \Phi(t; r)v &\iff \dot{x} = A(t)x, \quad x(r) = v, \\ p(t) = \Psi(t; s)w &\iff \dot{p} = -A(t)^T p, \quad p(s) = w.\end{aligned}$$

Observe that

$$\begin{aligned}\frac{d}{dt}(x(t) \cdot p(t)) &= \frac{d}{dt}x(t)^T p(t) \\ &= \dot{x}(t)^T p(t) + x(t)^T \dot{p}(t) \\ &= x(t)^T A(t)^T p(t) - x(t)^T A(t)^T p(t) = 0.\end{aligned}$$

Therefore the function $t \mapsto x(t)^T p(t)$ is constant: for arbitrary r, s ,

$$x(r)^T p(r) = x(s)^T p(s), \quad \text{i.e.,} \quad v^T \Psi(r; s)w = v^T \Phi(s; r)^T w.$$

Since this works for any $v, w \in \mathbb{R}^n$, it follows that $\Psi(r; s) = \Phi(s; r)^T$.

(g) If $n = 1$, $\Phi(t; \tau) = \exp\left(\int_{\tau}^t a(r) dr\right)$. For $n \geq 2$, this formula may fail.

Proof: For $n = 1$, just test that the choice above is consistent with the definition. Clearly $\Phi(\tau; \tau) = e^0 = 1$, and

$$\frac{d}{dt}\Phi(t; \tau) = a(t) \exp\left(\int_{\tau}^t a(r) dr\right) = a(t)\Phi(t; \tau) \quad \forall t.$$

For $n \geq 2$, the differentiation formula used in the previous line may not be valid: a general time-varying matrix M need not commute with its derivative, so

$$\text{when } n \geq 2, \quad \frac{d}{dt}e^{M(t)} \quad \text{often differs from} \quad e^{M(t)} \frac{dM}{dt}.$$

D. Inhomogeneous Linear Systems

D.1. Theorem. *If $\tau \in (a, b)$ and $g: (a, b) \rightarrow \mathbb{R}^n$ is integrable, the identity*

$$\frac{dx}{dt} = A(t)x + g(t), \quad a < t < b; \quad x(\tau) = \xi, \quad (D.1)$$

involving $x: (a, b) \rightarrow \mathbb{R}^n$ is equivalent to the identity

$$x(t) = \Phi(t; \tau)\xi + \int_{\tau}^t \Phi(t; r)g(r) dr, \quad a < t < b. \quad (D.2)$$

Here Φ is the fundamental matrix corresponding to $A(\cdot)$.

Proof. For simplicity, define $z: (a, b) \rightarrow \mathbb{R}^n$ using

$$z(t) = \Phi(t; \tau)^{-1}x(t), \quad \text{i.e.,} \quad x(t) = \Phi(t; \tau)z(t).$$

Differentiation gives

$$\begin{aligned} \dot{x}(t) &= A(t)x(t) + g(t) \\ \iff A(t)\Phi(t; \tau)z(t) + \Phi(t; \tau)\dot{z}(t) &= A(t)\Phi(t; \tau)z(t) + g(t). \end{aligned}$$

Cancellation occurs, and we get

$$\dot{z}(t) = \Phi(t; \tau)^{-1}g(t) = \Phi(\tau; t)g(t),$$

where we have used property (e) from Section C. Integrating this last relation gives

$$z(t) = z(\tau) + \int_{\tau}^t \Phi(\tau; r)g(r) dr.$$

To produce the desired result, we note that $z(\tau) = x(\tau) = \xi$, and consequently that

$$x(t) = \Phi(t; \tau)z(t) = \Phi(t; \tau)\xi + \Phi(t; \tau) \int_{\tau}^t \Phi(\tau; r)g(r) dr.$$

The formula given in the theorem statement comes from applying property (d) in Section C to the second term. ////

Remarks. **1.** In case $n = 1$, equation (D.2) gives a completely explicit solution for (D.1) in terms of integrals. This relies on part (g) of Section C, where we have an integral formula describing $\Phi(t; \tau)$ valid only for $n = 1$.

2. Whenever identity (D.1) holds for any function g , equation (D.2) follows. (The proof makes this clear.) There is no requirement for g to be independent of x . For example, if $n = 1$ and a is constant, any solution x of

$$\dot{x} = ax + x^3, \quad x(0) = \xi,$$

must satisfy

$$x(t) = e^{at}\xi + \int_0^t e^{a(t-r)}x(r)^3 dr.$$

This is not much help in actually calculating x (since x shows up on both the left and right sides), but it may be useful for inequalities, estimates, etc.

3. In the case where the matrix A is constant, the fundamental matrix reduces to $\Phi(t; \tau) = e^{A(t-\tau)}$, and formula (D.2) becomes

$$x(t) = e^{A(t-\tau)}\xi + \int_{\tau}^t e^{A(t-r)}g(r) dr. \quad (D.3)$$

Controlled Differential Equations. Consider a system described by

$$\frac{d\mathbf{x}}{dt} = A(t)\mathbf{x} + B(t)\mathbf{u}, \quad x(\tau) = \xi.$$

This is a standard model in applications, where matrix A is $n \times n$, matrix B is $n \times m$, and function $\mathbf{u}: \mathbb{R} \rightarrow \mathbb{R}^m$ is a “control” we are free choose to force good behaviour. The formula above gives the evolution from initial point (τ, ξ) :

$$x(t; \tau, \xi) = \Phi(t; \tau)\xi + \int_{\tau}^t \Phi(t; r)B(r)\mathbf{u}(r) dr,$$

This formula shows the evolution of the state x in response to the control input u in two pieces. The “zero-input response” $t \mapsto \Phi(t; \tau)\xi$ tells how the state would evolve from its initial point (τ, ξ) in the case $u \equiv 0$. The “zero-state response” $t \mapsto \int_{\tau}^t \Phi(t; r)B(r)\mathbf{u}(r) dr$ tells how the system would evolve if it started from the zero state $\xi = 0$ under the action of the control \mathbf{u} .

E. Linearization

Consider the general (time-varying, nonlinear) state equation

$$\dot{x}(t) = f(t, x(t), u(t)).$$

Let $\bar{u}(\cdot)$ be a *PWC* function with corresponding solution $\bar{x}(\cdot)$. If $(u(\cdot), x(\cdot))$ is a nearby solution pair, then Taylor’s theorem suggests

$$\begin{aligned} \dot{x}(t) &= f(t, x(t), u(t)) = \bar{f}(t) + \bar{f}_x(t)(x(t) - \bar{x}(t)) + \bar{f}_u(t)(u(t) - \bar{u}(t)) + \dots \\ \dot{\bar{x}}(t) &= \bar{f}(t) \end{aligned}$$

$$\therefore \dot{x}(t) - \dot{\bar{x}}(t) = \bar{f}_x(t)(x(t) - \bar{x}(t)) + \bar{f}_u(t)(u(t) - \bar{u}(t)) + \dots$$

Defining $y(t) := x(t) - \bar{x}(t)$ and $v(t) := u(t) - \bar{u}(t)$ then gives

$$\dot{y}(t) = A(t)y(t) + B(t)v(t) + \dots.$$

Simply ignoring “...” leads to a linear system which gives the approximate solutions

$$\begin{aligned} x(t) &= \bar{x}(t) + y(t), \\ u(t) &= \bar{u}(t) + v(t), \end{aligned}$$

in the form of “reference function plus linear correction”. We expect good approximations as long as the correction functions $y(\cdot)$ and $v(\cdot)$ are small.

E.1. Example (Satellite). We consider the motion of an earth satellite moving under the influence of gravity, with the capacity for thrust in the radial and tangential directions. The system is described in polar coordinates r, θ . The mass of the satellite is m ; the radial thrust is mu_2 ; the tangential thrust is mu_1 . The force of gravity at altitude r is km/r^2 , where k accounts for the mass of the earth and a constant of proportionality. The state vector is four-dimensional:

$$(x_1, x_2, x_3, x_4) := (r, \dot{r}, \theta, \dot{\theta}).$$

Newton's laws give

$$\begin{aligned} \text{(Radial)} \quad & m\ddot{r} - mr\dot{\theta}^2 = mu_2 - \frac{km}{r^2} \\ \text{(Tangential)} \quad & mr\ddot{\theta} + 2m\dot{r}\dot{\theta} = mu_1. \end{aligned}$$

We use the states above to write this as a first-order system:

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= -k/x_1^2 + x_1x_4^2 + u_2 \\ \dot{x}_3 &= x_4 \\ \dot{x}_4 &= -2\frac{x_2}{x_1}x_4 + u_1/x_1 \end{aligned}$$

For our reference solution, we consider a circular orbit of radius ρ with constant angular velocity ω and no power to the thrusters: a simple computation confirms that

$$(\bar{x}_1(t), \bar{x}_2(t), \bar{x}_3(t), \bar{x}_4(t)) = (\rho, 0, \omega t, \omega), \quad (\bar{u}_1(t), \bar{u}_2(t)) = (0, 0)$$

provide a solution if and only if $0 = -k\rho^2 + \rho\omega^2$, i.e., $\rho^3\omega^2 = k$. Linearization gives

$$\begin{aligned} \begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \\ \dot{y}_3 \\ \dot{y}_4 \end{bmatrix} &= \begin{bmatrix} 0 & 1 & 0 & 0 \\ 2k/\bar{x}_1^3 + \bar{x}_4^2 & 0 & 0 & 2\bar{x}_1\bar{x}_4 \\ 0 & 0 & 0 & 1 \\ 2\bar{x}_2\bar{x}_4/\bar{x}_1^2 - \bar{u}_1/\bar{x}_1^2 & -2\bar{x}_4/\bar{x}_1 & 0 & -2\bar{x}_2/\bar{x}_1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 0 & 0 \\ 1/\bar{x}_1 & 0 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \\ \dot{y} &= \begin{bmatrix} 0 & 1 & 0 & 0 \\ 3\omega^2 & 0 & 0 & 2\omega\rho \\ 0 & 0 & 0 & 1 \\ 0 & -2\omega/\rho & 0 & 0 \end{bmatrix} y + \begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 0 & 0 \\ 1/\rho & 0 \end{bmatrix} v. \end{aligned}$$

In future references to this example we will take $\rho = 1$.

A simple computation reveals that the eigenvalues of the linearized system matrix lie at $0, 0, i\omega$, and $-i\omega$. All of them are on the imaginary axis. The stability properties of the linearized system are not obvious—a proper understanding of the system's local behaviour requires the use of higher-order information. ////