Control of Autonomous Linear Systems

M403 Lecture Notes by Philip D. Loewen

We study systems governed by the controlled differential equation

$$\dot{x}(t) = Ax(t) + Bu(t), \qquad x(0) = \xi, \tag{$*$}$$

where x evolves in $\mathbb{R}^n$, u evolves in $\mathbb{R}^m$, and the coefficient matrices $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$ are given. We will focus primarily on the class of *piecewise continuous* control functions u: a function $u\colon [0,+\infty) \to \mathbb{R}^m$ is called piecewise continuous if every finite interval $[0,b]$ admits a finite partition $0 = a_0 < a_1 < \ldots < a_N = b$ with the property that u is continuous on each open interval (a_{k-1}, a_k) with finite one-sided limits $u(a_{k-1}^+)$ and $u(a_k^-)$ for each $k = 1, 2, \ldots, N$. The collection of all piecewise continuous functions is denoted PWC. Recall that for a fixed control u in PWC, the unique solution of $(*)$ is given by

$$x(t) = e^{tA}\xi + \int_0^t e^{(t-r)A}Bu(r)\,dr.$$

(The right-hand side of this formula actually makes sense for any function u that is integrable on every finite real interval.)

We will look at three different approaches to the choice of u in $(*)$, each with different criteria and outcomes: linear feedback and eigenvalue assignment, linear-quadratic regulator design, and time-optimal control.

A. Controllability

Introduction. Which states can be joined by a piecewise continuous control function? To answer this basic question, consider the *attainable set*

$$\mathcal{A}(t;\xi) = \left\{ \eta \in \mathbb{R}^n \,:\, \eta = e^{At}\xi + \int_0^t e^{A(t-r)}Bu(r)\,dr \text{ for some } u \in PWC\left([0,t];\mathbb{R}^m\right) \right\}.$$

This set contains a point η if and only some piecewise continuous control $u\colon [0,t] \to \mathbb{R}^m$ drives the state to the terminal point η at time t, along a trajectory x for the system equation $(*)$ starting from $x(0) = \xi$. In general, $\mathcal{A}(t;\xi)$ is a time-dependent subset of $\mathbb{R}^n$. At least three of its properties are comparatively obvious:

(i) $\mathcal{A}(t;\xi) = e^{At}\xi + \mathcal{A}(t;0)$, for all $t \geq 0$, $\xi \in \mathbb{R}^n$;

(ii) $\mathcal{A}(t;0)$ is a linear subspace of $\mathbb{R}^n$ for every $t \geq 0$;

(iii) $\mathcal{A}(t;\xi)$ is an affine subspace of $\mathbb{R}^n$ for every $t \geq 0$.

We can characterize $\mathcal{A}(t;\xi)$ with the aid of line (i) and the famous *controllability matrix*

$$\mathcal{C} := \left[B \,|\, AB \,|\, A^2B \,|\, \cdots \,|\, A^{n-1}B\right].$$

This is an $n \times mn$ matrix formed by stacking n matrices of size $n \times m$ side by side; examples appear below.

A.1. Theorem. *For each $t > 0$, one has*

(a) $\mathcal{A}(t;0) = \mathrm{Im}(\mathcal{C}) = \{\mathcal{C}w \,:\, w \in \mathbb{R}^{nm}\}$;

(b) $\mathcal{A}(t;0)^{\perp} = \ker(\mathcal{C}^T) = \left\{y \in \mathbb{R}^n \,:\, \mathcal{C}^T y = 0\right\}$; *and*

(c) $\dim \mathcal{A}(t;0) = \mathrm{rank}(\mathcal{C})$.

Proof. For any matrix $\mathcal{C}$, standard results from linear algebra assert that $\mathrm{Im}(\mathcal{C})^{\perp} = \ker(\mathcal{C}^T)$ and $\mathrm{rank}(\mathcal{C}) = \dim(\mathrm{Im}(\mathcal{C}))$. Consequently (a) implies both (b) and (c). But since one has $(S^{\perp})^{\perp} = S$ for every subspace S of $\mathbb{R}^n$, it is also true that (b) implies (a). Thus the theorem can be established simply by proving (b). This we now do.

(b)($\supseteq$) If y is a vector in $\ker(\mathcal{C}^T)$, then $\mathcal{C}^T y = 0$, so

$$0^T = y^T\mathcal{C} = \left[y^TB \,|\, y^TAB \,|\, y^TA^2B \,|\, \cdots \,|\, y^TA^{n-1}B\right].$$

(Both sides here are row vectors of size mn.) Now by the Cayley Hamilton theorem, A satisfies its own characteristic polynomial. That is, for the coefficients $a_0, \ldots, a_{n-1}$ defined by

$$p(s) = \det(sI - A) = s^n - a_{n-1}s^{n-1} - \ldots - a_1 s - a_0,$$

the following matrix equation holds:

$$0 = p(A) = A^n - a_{n-1}A^{n-1} - \ldots - a_1 A - a_0 I.$$

In particular, one has $A^n = a_{n-1}A^{n-1} + \ldots + a_1 A + a_0 I$, and it follows that

$$y^TA^nB = y^T\left(\sum_{i=0}^{n-1} a_i A^i\right)B = \sum_{i=0}^{n-1} a_i\left(y^TA^iB\right) = 0.$$

Likewise, $y^TA^kB = 0$ for all $k \geq 0$. It follows that for any $r \in [0,t]$, one has

$$\begin{aligned} y^Te^{(t-r)A}B &= y^T\left(\lim_{N\to\infty}\sum_{k=0}^{N}\frac{(t-r)^k}{k!}A^k\right)B \\ &= \lim_{N\to\infty}\sum_{k=0}^{N}\left[\frac{(t-r)^k}{k!}(y^TA^kB)\right] = \lim_{N\to\infty}\sum_{k=0}^{N} 0 = 0. \end{aligned}$$

It follows that for any control u in PWC, one has

$$0 = y^T\int_0^t e^{(t-r)A}Bu(r)\,dr,$$

which shows that $0 = y^T\eta$ for every $\eta \in \mathcal{A}(t;0)$. This is the definition of $y \in \mathcal{A}(t,0)^{\perp}$.

(b)($\subseteq$) If y is a vector in $\mathcal{A}(t;0)^{\perp}$, then $0 = y^T\eta$ for every $\eta \in \mathcal{A}(t;0)$. In other words, for any control u in PWC, one has

$$0 = y^T\int_0^t e^{A(t-r)}Bu(r)\,dr,$$

Choosing the control $u(r) = B^T e^{A^T(t-r)} y$ reveals

$$0 = \int_0^t \left| y^T e^{(t-r)A} B \right|^2 dr,$$

from which we deduce that $y^T e^{A(t-r)} B = 0$ for all r in $[0,t]$. Alternately letting $r \to t$ and differentiating in this relation, we have

$$\begin{aligned}
& y^T e^{A(t-r)} B = 0 && \forall r \in (0,t) && \left(\Rightarrow y^T B = 0 \right) \\
\frac{\partial}{\partial r}\text{:} \quad & y^T e^{A(t-r)} (-A) B = 0 && \forall r \in (0,t) && \left(\Rightarrow y^T AB = 0 \right) \\
\frac{\partial}{\partial r}\text{:} \quad & y^T e^{A(t-r)} (-A)^2 B = 0 && \forall r \in (0,t) && \left(\Rightarrow y^T A^2 B = 0 \right) \\
& \qquad \vdots \\
\frac{\partial}{\partial r}\text{:} \quad & y^T e^{A(t-r)} (-A)^{n-1} B = 0 && \forall r \in (0,t) && \left(\Rightarrow y^T A^{n-1} B = 0 \right)
\end{aligned}$$

This reveals that $0^T = \left[y^T B \,|\, y^T AB \,|\, \cdots \,|\, y^T A^{n-1} B \right] = y^T \mathcal{C}$, i.e., $\mathcal{C}^T y = 0$. ////

A.2. Corollary. *For a given pair of matrices $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$, the following assertions are equivalent:*

(a) $\operatorname{rank}(\mathcal{C}) = n$, *where* $\mathcal{C} = \left[B \,|\, AB \,|\, \cdots \,|\, A^{n-1} B \right]$;

(b) $\mathcal{A}(t;0) = \mathbb{R}^n$ *for some* $t > 0$.

(c) $\mathcal{A}(t;0) = \mathbb{R}^n$ *for each* $t > 0$.

(d) $\mathcal{A}(t;\xi) = \mathbb{R}^n$ *for each* $t > 0$ *and each* $\xi \in \mathbb{R}^n$.

In this case we say the matrix pair (A,B) is controllable.

A.3. Remarks. (a) If the pair (A,B) is uncontrollable, then any trajectory of $(*)$ starting from the origin is trapped in the fixed subspace $\mathcal{A}(t;0) = \operatorname{Im}(\mathcal{C})$ of $\mathbb{R}^n$ for all time. Any trajectory starting from some initial point ξ can never leave the moving affine subspace $\mathcal{A}(t;\xi) = e^{At}\xi + \operatorname{Im}(\mathcal{C})$.

(b) In situations where the main goal is to stabilize the system, a more intuitive notion is **null-controllability:** one defines $\mathcal{N}(t) = \{\xi \in \mathbb{R}^n : 0 \in \mathcal{A}(t;\xi)\}$ as the set of all initial points that can be steered to the origin by some control acting on the interval $[0,t]$. But we note that

$$\begin{aligned}
\xi \in \mathcal{N}(t) &\Leftrightarrow 0 \in e^{At}\xi + \mathcal{A}(t;0) \\
&\Leftrightarrow \xi \in -e^{-At}\mathcal{A}(t;0).
\end{aligned}$$

It follows that $\mathcal{N}(t) = -e^{-At}\mathcal{A}(t;0)$ is a subspace of $\mathbb{R}^n$ for each $t > 0$. In fact, it is a constant subspace, equal to $\mathcal{A}(t;0) = \operatorname{Im}(\mathcal{C})$ for all $t > 0$. (This is a nice practice problem.)

(c) Theorem A.1 remains valid for various families of control functions u beyond the collection of piecewise continuous controls: we could use only continuous controls, or we could permit only piecewise constant controls. The key to the proofs above is to have a vector space $\mathcal{U}$ of control functions sufficiently large that $y^T\eta = 0$ for all $\eta \in \mathcal{A}(t;0)$ will imply $y^T e^{A(t-r)}B = 0$ for all $r \in [0,t]$. Any one of the three just mentioned has this property.

(d) The equation $\mathcal{A}(t;0) = \mathbb{R}^n$ for all $t > 0$ is mathematically correct, but clearly unrealizable in practice. This inconsistency reveals a flaw in the modelling behind $(*)$, which allows for arbitrarily large values for the control input vector u. Later we will see how putting constraints on u changes the picture.

(e) If $m < n$ you can save time by computing $A(AB)$ instead of $(A^2)B$, even though the two results are equal.

A.4. Examples. (a) A system trapped in the plane $z = 0$.

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad B = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \qquad \Longrightarrow \qquad \mathcal{C} = \left[\begin{array}{c|c|c} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 0 & 0 \end{array}\right].$$

This system is not controllable, because the columns of $\mathcal{C}$ span only a 2-dimensional subspace of $\mathbb{R}^3$:

$$\mathcal{A}(t;\mathbf{0}) = \mathrm{Im}(\mathcal{C}) = \left\{ \begin{bmatrix} r \\ s \\ 0 \end{bmatrix} \,:\, r, s \in \mathbb{R} \right\}.$$

(b) Linearized satellite problem. Here

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 3\omega^2 & 0 & 0 & 2\omega \\ 0 & 0 & 0 & 1 \\ 0 & -2\omega & 0 & 0 \end{bmatrix}, \qquad B = \begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 0 & 0 \\ 1 & 0 \end{bmatrix}.$$

Computation gives

$$\mathcal{C} = \left[\begin{array}{cc|cc|cc|cc} 0 & 0 & 0 & 1 & 2\omega & 0 & 0 & -\omega^2 \\ 0 & 1 & 2\omega & 0 & 0 & -\omega^2 & -2\omega^3 & 0 \\ 0 & 0 & 1 & 0 & 0 & -2\omega & -4\omega^2 & 0 \\ 1 & 0 & 0 & -2\omega & -4\omega^2 & 0 & 0 & 2\omega^3 \end{array}\right].$$

The first 4×4 block has full rank, so the system is controllable.

Loss of radial thrust: If motor failure forces $u_2 \equiv 0$, this is the same as setting column 2 of matrix B to zero. This will zero columns 2,4,6,8 of matrix $\mathcal{C}$, but the rank will still be 4, so the system remains controllable.

Loss of tangential thrust: If instead we set column 1 of matrix B to zero, this zeros columns 1,3,5,7 of matric $\mathcal{C}$ and reduces its rank to 3. (row 4 is a multiple of row 1). Controllability is broken. ////

B. Linear State Feedback

Consider this linear time-invariant (LTI) system in which a scalar input u drives a scalar response variable x through a differential equation of order n:

$$x^{(n)} + a_{n-1}x^{(n-1)} + a_{n-2}x^{(n-2)} + \ldots + a_2\ddot{x} + a_1\dot{x} + a_0x = u. \qquad (*)$$

There are two approaches to controlling it. The first is **open-loop control,** where we use prior knowledge of the initial system configuration to design a particular function $u = u(t)$ for use in $(*)$. Different initial conditions will typically require a completely different function u. In the second approach, **closed-loop control**, we design a function $U\colon \mathbb{R}^n \to \mathbb{R}$ and use it to generate the input signal in $(*)$ by enforcing the identity

$$u(t) = U(x(t), \dot{x}(t), \ldots, x^{(n-1)}(t)).$$

This is mathematically equivalent to replacing $(*)$ with the autonomous dynamical system

$$x^{(n)} + a_{n-1}x^{(n-1)} + \ldots + a_2\ddot{x} + a_1\dot{x} + a_0x = U(x, \dot{x}, \ldots, x^{(n-1)}). \qquad (\dagger)$$

Ideally, the feedback law U could be chosen so that the closed-loop system has good dynamics independent of the choice of initial data. The simplest type of feedback function U is linear: we choose constants $f_0, f_1, \ldots, f_n$ and specify

$$U(x_1, x_2, \ldots, x_n) = f_0x_1 + f_1x_2 + \cdots + f_{n-1}x_n.$$

With this choice, system $(\dagger)$ becomes

$$\begin{aligned} x^{(n)} + a_{n-1}x^{(n-1)} + \ldots + a_2\ddot{x} + a_1\dot{x} + a_0x &= f_0x + f_1\dot{x} + f_2\ddot{x} + \ldots + f_{n-1}x^{(n-1)}, \\ x^{(n)} + (a_{n-1} - f_{n-1})\,x^{(n-1)} + \ldots + (a_2 - f_2)\,\ddot{x} + (a_1 - f_1)\,\dot{x} + (a_0 - f_0)\,x &= 0. \quad (\ddagger) \end{aligned}$$

To solve $(\ddagger)$ would involve "guessing" $x = e^{st}$ and plugging in, then observing that we have a solution if and only if the complex constant s obeys

$$0 = p(s) \stackrel{\text{def}}{=} s^n + (a_{n-1} - f_{n-1})\,s^{n-1} + \ldots + (a_2 - f_2)\,s^2 + (a_1 - f_1)\,s + (a_0 - f_0)\,.$$

Since the constants f_k are our design parameters, we can influence every coefficient in this polynomial, and hence arrange any collection of zeros we want for the characteristic polynomial. (The only restriction is that since p is a polynomial with real coefficients, its zeros must occur in complex-conjugate pairs.) In detail, given any set of n complex numbers $\lambda_1, \lambda_2, \ldots, \lambda_n$ with reflection symmetry across the real axis of the complex plane, a polynomial with these numbers as its zeros is given by

$$P(s) = (s - \lambda_1)(s - \lambda_2)\cdots(s - \lambda_n) = s^n + \alpha_{n-1}s^{n-1} + \cdots + \alpha_2s^2 + \alpha_1s + \alpha_0$$

for suitable real constants $\alpha_0, \ldots, \alpha_n$. (Knowing the λ_k's lets us compute the α_k's.) We can arrange for the zeros of p to fall at precisely these points by forcing $p \equiv P$, i.e., by choosing

$$a_k - f_k = \alpha_k, \quad \text{or} \quad f_k = a_k - \alpha_k,\ k = 0, 1, \ldots, n-1.$$

Practically-minded people (e.g., engineers) have pretty clear ideas of what regions in the complex plane make good homes for the eigenvalues for a well-behaved system. We explore these next.

Eigenvalue Assignment Criteria. We use the simplest possible 2×2 system to justify some standard design criteria used in applied science. Suppose we choose the feedback matrix F and apply it, to obtain an autonomous system with coefficient matrix

$$A + BF = \begin{bmatrix} 0 & 1 \\ -\omega^2 & -2\zeta\omega \end{bmatrix}.$$

(Here $\zeta \in [0, 1)$ is the "damping factor" and $\omega > 0$ is the "undamped natural frequency", both constant.) Study desirable properties of solutions to $\dot{\mathbf{x}} = (A + BF)\mathbf{x}$, starting with an expansion of $\mathbf{x}(t) = (y(t), v(t))$:

$$\dot{y} = v, \quad \dot{v} = -\omega^2 y - 2\zeta\omega v \qquad \Longrightarrow \qquad \ddot{y} + 2\zeta\omega\dot{y} + \omega^2 y = 0.$$

Guess $y = e^{st}$ for constant s and get solutions iff

$$0 = s^2 + 2\zeta\omega s + \omega^2, \quad \text{i.e.,} \quad s = -\zeta\omega \pm \sqrt{\zeta^2\omega^2 - \omega^2} = -\zeta\omega \pm i\omega\sqrt{1 - \zeta^2}.$$

(Or, look for eigenvalues of $A + BF$ using its characteristic polynomial, which can be read out of the matrix directly, thanks to its special form (treated last class):

$$p(\lambda) = \lambda^2 + 2\zeta\omega\lambda + \omega^2.$$

The same complex numbers result.)

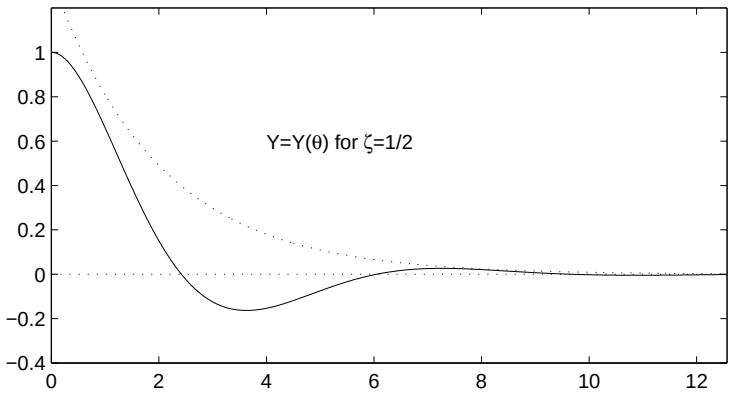
Running the system $\dot{\mathbf{x}} = (A+BF)\mathbf{x}$ from the initial state $\mathbf{x}(0) = (1, 0)$ produces a trajectory whose first component is

$$y(t) = e^{-\zeta\omega t}\left[\cos(\sqrt{1-\zeta^2}\omega t) + \frac{\zeta}{\sqrt{1-\zeta^2}}\sin(\sqrt{1-\zeta^2}\omega t)\right]. \tag{1}$$

We study y using the change of variables $\theta = \omega t$, writing

$$Y(\theta) = y(\theta/\omega) = e^{-\zeta\theta}\left[\cos(\sqrt{1-\zeta^2}\theta) + \frac{\zeta}{\sqrt{1-\zeta^2}}\sin(\sqrt{1-\zeta^2}\theta)\right].$$

The sketch below shows Y for $\zeta = \frac{1}{2}$.



Standard methods reveal that

$$Y(\theta) = \left(\frac{1}{1-\zeta^2}\right) e^{-\zeta\theta} \cos(\sqrt{1-\zeta^2}\theta - \phi) \quad \text{for} \quad \phi = \tan^{-1}\left(\frac{\zeta}{\sqrt{1-\zeta^2}}\right). \qquad (2)$$

The decaying exponential factor $\left(\frac{1}{1-\zeta^2}\right) e^{-\zeta\theta}$ is also shown in the sketch above. Three desirable properties can be quantified with reference to this figure.

Response Time. Write t_R for the time required for the disturbance applied above to make the transition from 90% of its initial height to 10%. That is, $\omega t_R = \theta_2 - \theta_1 =: \theta_R$ for the smallest positive solutions of

$$Y(\theta_1) = 0.90, \qquad Y(\theta_2) = 0.10.$$

Numerical values of θ_R for various choices of ζ are tabulated below. They are easy to find numerically using Maple or Matlab:

ζ	θ_R
0.10	1.10
0.20	1.20
0.30	1.32
0.40	1.46
0.50	1.64
0.60	1.85
0.70	2.13
0.80	2.47
0.90	2.88

A (rather poor) constant approximation for θ_R based on these observations is $\theta_R \approx 1.8$, and it leads to the approximate relation

$$t_R \approx \frac{1.8}{\omega}.$$

Requesting a faster response time means making t_R smaller, so making ω larger.

Settling Time. Write t_S for the time required for the amplitude factor in (2) to drop to 1% of its initial value, so that $\theta_S = \omega t_S$ satisfies

$$e^{-\zeta\theta_S} = \frac{1}{100}, \qquad \text{i.e.,} \qquad \theta_S = \frac{\ln(100)}{\zeta}, \qquad \text{i.e.,} \qquad t_S = \frac{\ln(100)}{\zeta\omega}.$$

Requesting faster settling time means making t_S smaller, so making $\zeta\omega$ larger.

Peak Overshoot. The minimum value of y matches the minimum value of Y. It occurs at the smallest positive solution of $Y'(\theta) = 0$, namely,

$$\theta_{\text{peak}} = \frac{\pi}{\sqrt{1-\zeta^2}}, \quad \text{where} \quad Y(\theta_{\text{peak}}) = -e^{-\pi\zeta/\sqrt{1-\zeta^2}}.$$

Requesting smaller values for $|Y_{\text{peak}}| = \exp\left(\frac{\pi\zeta}{\sqrt{1-\zeta^2}}\right)$ means making $\zeta/\sqrt{1-\zeta^2}$ larger.

Eigenvalue Connections. For the linear system under study, the eigenvalues are

$$\lambda = -\zeta\omega \pm i\omega\sqrt{1-\zeta^2},$$

giving

$$\begin{aligned}
&\Re e(\lambda) = -\zeta\omega,\\
&\Im m(\lambda) = \pm\omega\sqrt{1-\zeta^2},\\
&\frac{\Im m(\lambda)}{\Re e(\lambda)} = \pm\frac{\sqrt{1-\zeta^2}}{\zeta},\\
&|\lambda|^2 = \zeta^2\omega^2 + \omega^2\left(1-\zeta^2\right) = \omega^2.
\end{aligned}$$

The three parameters of physical interest can be expressed as

$$\begin{aligned}
t_R &\approx \frac{1.8}{|\lambda|},\\
t_S &= \frac{\ln(100)}{\zeta\omega} \approx -\frac{4.6}{\Re e(\lambda)},\\
\frac{1}{\pi}\ln\left(\frac{1}{|Y_{\text{peak}}|}\right) &= \frac{\zeta}{\sqrt{1-\zeta^2}} = \left|\frac{\Im m(\lambda)}{\Re e(\lambda)}\right|^{-1}.
\end{aligned}$$

So for small response time, we want $|\lambda|$ reasonably large. (Sketch: λ outside a disk around 0 in $\mathbb{C}$.) For small settling time, we want $\Re e(\lambda) < 0$ with large magnitude. (Sketch: λ left of some vertical line the the left of the imaginary axis.) For small overshoot, we want $\Im m(\lambda)/\Re e(\lambda)$ small in magnitude. (Sketch: λ in some wedge with slope not too big having vertex at origin of $\mathbb{C}$.) The composite criterion gives an interesting shape in $\mathbb{C}$ where eigenvalues should be placed.

Matrix Version. What does the development above look like in the standard vector-matrix setup? Recall the basic controlled differential equation:

$$x^{(n)} + a_{n-1}x^{(n-1)} + a_{n-2}x^{(n-2)} + \ldots + a_2\ddot{x} + a_1\dot{x} + a_0 x = u. \qquad (*)$$

The standard choices $x_1 = x$, $x_2 = \dot{x}$, …, $x_n = x^{(n-1)}$ lead to a first-order system whose vector-matrix form is

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \vdots \\ \dot{x}_{n-1} \\ \dot{x}_n \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -a_0 & -a_1 & -a_2 & \cdots & -a_{n-1} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} u. \qquad (**)$$

It is easy to see that this linear system is controllable (home practice). What is more, the characteristic polynomial of the state coefficient matrix (call it A) is precisely the same polynomial we used in the scalar approach, namely

$$p(s) = |sI - A| = s^n + a_{n-1}s^{n-1} + \ldots + a_2 s^2 + a_1 s + a_0.$$

Proof. Compute $\det(sI - A)$ recursively using expansion along the last column. Note

$$sI - A = \begin{bmatrix} s & -1 & 0 & \cdots & 0 & 0 \\ 0 & s & -1 & \cdots & 0 & 0 \\ 0 & 0 & s & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & s & -1 \\ a_0 & a_1 & a_2 & \cdots & a_{n-2} & s + a_{n-1} \end{bmatrix}.$$

After the first stage, use the notation

$$p_k(s) = \begin{vmatrix} s & -1 & 0 & \cdots & 0 & 0 \\ 0 & s & -1 & \cdots & 0 & 0 \\ 0 & 0 & s & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & s & -1 \\ a_0 & a_1 & a_2 & \cdots & a_{k-2} & a_{k-1} \end{vmatrix}, \qquad k = 1, 2, \ldots, n-1.$$

The indicated expansion gives

$$\begin{aligned} \det(sI - A) &= (s + a_{n-1})s^{n-1} - (-1)\begin{vmatrix} \cdots \end{vmatrix} = s^n + a_{n-1}s^{n-1} + p_{n-1}(s) \\ &= s^n + a_{n-1}s^{n-1} + \left[a_{n-2}s^{n-2} + p_{n-2}(s)\right] \\ &\vdots \\ &= s^n + a_{n-1}s^{n-1} + a_{n-2}s^{n-2} + \cdots + a_1 s + a_0. \end{aligned}$$ ////

In the vector-matrix setup, the general feedback law $U = U(x)$ transforms the system into the autonomous $\dot{x} = Ax + BU(x)$. The linear choice considered above can be expressed as

$$U(x) = f_0 x_1 + f_1 x_2 + \cdots + f_{n-1} x_n = [f_0 \ f_1 \ \cdots f_{n-1}]\, x \stackrel{\text{def}}{=} Fx$$

where F is the $1 \times n$ matrix of constants appearing above. This choice replaces the controlled linear system $(**)$ with the linear system

$$\dot{x} = Ax + B(Fx) = (A + BF)x.$$

The product BF is a square matrix whose first $n-1$ rows are zero and whose last row is an exact copy of F. Thus the feedback system has dynamics given by

$$A + BF = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ (f_0 - a_0) & (f_1 - a_1) & (f_2 - a_2) & \cdots & (f_{n-1} - a_{n-1}) \end{bmatrix}.$$

This matrix has the same form as the matrix A in $(**)$: the work above shows that its characteristic polynomial must be

$$P(s) = |sI - (A + BF)| = s^n + (a_{n-1} - f_{n-1})s^{n-1} + \ldots + (a_1 - f_1)s + (a_0 - f_0).$$

So choosing the coefficients in the feedback "matrix" F is equivalent to specifying each of the coefficients in the characteristic polynomial of the feedback system. As noted earlier, the freedom to select these coefficients is equivalent to the power to select the zeros of P: this explains the term "eigenvalue assignment" for the task of choosing F.

Equivalent Systems. To treat systems of the form $\dot{x} = Ax + Bu$ where u is scalar but the matrices A and B do not have the special structure above, we make a change of coordinates. The early parts of the theory apply even to systems with vector-valued inputs.

Given a controlled linear system with coefficient matrices A (size $n \times n$) and B (size $n \times m$), one might imagine expressing the state vector x in terms of different coordinates. Suppose, for definiteness, that the new coordinates form an n-vector y, with the old and new coordinate schemes related by an invertible square matrix P:

$$x = Py, \qquad y = P^{-1}x.$$

In this case we can use the given system dynamics to find a differential equation satisfied by the new coordinate vector y:

$$\dot{y} = P^{-1}\dot{x} = P^{-1}\left[Ax + Bu\right] = P^{-1}A\left(Py\right) + P^{-1}Bu.$$

This has the same form as the original equation, but new coefficient matrices we might call $\widetilde{A}$ and $\widetilde{B}$. As a general rule, we will say that two matrix pairs (A, B) and $(\widetilde{A}, \widetilde{B})$ are *equivalent* if there exists some nonsingular matrix P for which

$$\widetilde{A} = P^{-1}AP, \qquad \widetilde{B} = P^{-1}B.$$

It is an easy exercise to show that this "equivalence" satisfies the three basic properties of an abstract mathematical "equivalence relation":

(R) Reflexivity: If (A, B) is equivalent to $(\widetilde{A}, \widetilde{B})$, then $(\widetilde{A}, \widetilde{B})$ is equivalent to (A, B);

(S) Symmetry: (A, B) is equivalent to (A, B).

(T) Transitivity: If (A, B) is equivalent to $(\widetilde{A}, \widetilde{B})$, which is in turn equivalent to $(\overline{A}, \overline{B})$, then (A, B) is equivalent to $(\overline{A}, \overline{B})$.

The intuitive idea to keep in mind, however, is that equivalent systems describe the same geometrical object in terms of different coordinate schemes. In particular, the choice of coordinates should not make any difference when testing the controllability of a system. To see that this is the case, consider the controllability matrix in new coordinates:

$$\begin{aligned}
\widetilde{\mathcal{C}} &= \left[\, \widetilde{B} \,|\, \widetilde{A}\widetilde{B} \,|\, \widetilde{A}^2\widetilde{B} \,|\, \cdots \,|\, \widetilde{A}^{n-1}\widetilde{B} \,\right] \\
&= \left[\, P^{-1}B \,|\, \left(P^{-1}AP\right)\left(P^{-1}B\right) \,|\, \left(P^{-1}AP\right)^2\left(P^{-1}B\right) \,|\, \cdots \,|\, \left(P^{-1}AP\right)^{n-1}\left(P^{-1}B\right) \,\right] \\
&= P^{-1}\left[\, B \,|\, AB \,|\, A^2B \,|\, \cdots \,|\, A^{n-1}B \,\right] = P^{-1}\mathcal{C}.
\end{aligned}$$

Since the matrix P^{-1} is nonsingular, the two matrices $\mathcal{C}$ and $\widetilde{\mathcal{C}}$ have the same rank. If the rank is n, then both systems are controllable. If the rank is less than n, both systems are trapped in affine subspaces of $\mathbb{R}^n$: in the original coordinates, the subspace in question is $\mathrm{Im}(\mathcal{C})$; in the new coordinates, the same subspace is described as $\mathrm{Im}(\widetilde{\mathcal{C}}) = P^{-1}\,\mathrm{Im}(\mathcal{C})$.

For equivalent systems, the state coefficient matrices have the same characteristic polynomial, and hence the same eigenvalues. To see this, suppose $\widetilde{A} = P^{-1}AP$. Then

$$\left|\lambda I - \widetilde{A}\right| = \left|P^{-1}\left[\lambda I - A\right]P\right| = |P|^{-1}\,|\lambda I - A|\,|P| = |\lambda I - A|.$$

We now prove that any controllable single-input system (i.e., $m = 1$) can be given an equivalent representation in the form of $(**)$ above. This explains the name of the form $(**)$ as the *controllable canonical form* for the single-input case, and has good consequences for linear feedback.

B.1. Theorem. *Suppose (A, B) is a matrix pair, with $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{m\times n}$, and $m = 1$. Then the following are equivalent:*

(a) *The pair (A, B) is controllable, and the characteristic polynomial of A is $p(\lambda) \stackrel{\text{def}}{=} \det(\lambda I - A) = \lambda^n + a_{n-1}\lambda^{n-1} + \ldots + a_1\lambda + a_0$.*

(b) *There exists a nonsingular "coordinate change" matrix P such that the matrices $\widetilde{A} = P^{-1}AP$ and $\widetilde{B} = P^{-1}B$ are given by*

$$\widetilde{A} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -a_0 & -a_1 & -a_2 & \cdots & -a_{n-1} \end{bmatrix}, \qquad \widetilde{B} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}.$$

Proof. (b)⇒(a) Assume (b). It is straightforward to show that the matrix pair $(\widetilde{A}, \widetilde{B})$ is controllable. According to our discussion of equivalent systems, it follows that (A, B) is controllable. Moreover, since $A = P\widetilde{A}P^{-1}$, the characteristic polynomial of A is the same as the characteristic polynomial of $\widetilde{A}$. As we have already seen, this is precisely the polynomial written out in statement (a).

(a)⇒(b)Define the n columns in a matrix P as follows:

$$[v_1|v_2|\cdots|v_n] = P = \mathcal{C}\begin{bmatrix} a_1 & \cdots & a_{n-2} & a_{n-1} & 1 \\ a_2 & \cdots & a_{n-1} & 1 & 0 \\ a_3 & \cdots & 1 & 0 & 0 \\ \vdots & \ddots & & & \\ 1 & \cdots & 0 & 0 & 0 \end{bmatrix}.$$

Since both matrices on the RHS are invertible ($\mathcal{C}$ by hypothesis, the other by inspection), so is P. We'll show that P does the job.

Work from right to left, expressing the columns of the matrix shown above in terms of standard unit vectors:

$$\begin{aligned}
v_n &= \mathcal{C}\widehat{\mathbf{e}}_1 = B \\
v_{n-1} &= \mathcal{C}\left[a_{n-1}\widehat{\mathbf{e}}_1 + \widehat{\mathbf{e}}_2\right] = a_{n-1}B + AB = a_{n-1}v_n + Av_n \\
v_{n-2} &= \mathcal{C}\left[a_{n-2}\widehat{\mathbf{e}}_1 + a_{n-1}\widehat{\mathbf{e}}_2 + \widehat{\mathbf{e}}_3\right] = a_{n-2}B + a_{n-1}AB + A^2B = a_{n-2}v_n + Av_{n-1} \\
&\vdots \\
v_1 &= \mathcal{C}\left[a_1\widehat{\mathbf{e}}_1 + a_2\widehat{\mathbf{e}}_2 + \cdots + \alpha_{n-1}\widehat{\mathbf{e}}_{n-1} + \widehat{\mathbf{e}}_n\right] \\
&= a_1B + a_2AB + \cdots + a_{n-1}A^{n-2}B + A^{n-1}B = a_1v_n + Av_2.
\end{aligned}$$

Observation 1: For $2 \le k \le n$, $v_{k-1} = a_{k-1}v_n + Av_k$, so

$$Av_k = v_{k-1} - a_{k-1}v_n \qquad \text{for } k = 2, \ldots, n.$$

Observation 2: By the Cayley-Hamilton Theorem, $p(A) = 0$. Hence

$$Av_1 = \left[a_1A + a_2A^2 + \cdots + a_{n-1}A^{n-1} + A^n\right]B = \left[-a_0I\right]B = -a_0v_n.$$

Observation 3: Column-by-column expansion of $I = P^{-1}P$ gives

$$\left[\widehat{\mathbf{e}}_1 \,|\, \widehat{\mathbf{e}}_2 \,|\, \cdots \,|\, \widehat{\mathbf{e}}_n\right] = P^{-1}\left[v_1 \,|\, v_2 \,|\, \cdots \,|\, v_n\right], \qquad \text{i.e.,} \qquad \widehat{\mathbf{e}}_j = P^{-1}v_j \ \forall j = 1, \ldots, n.$$

In combination with Observation 1 (rearranged), this lets us calculate column k of the matrix $P^{-1}AP$:

$$\begin{aligned}
P^{-1}Av_1 &= P^{-1}(-a_0v_n) = -a_0\widehat{\mathbf{e}}_n \\
P^{-1}Av_k &= P^{-1}\left[v_{k-1} - a_{k-1}v_n\right] = \widehat{\mathbf{e}}_{k-1} - a_{k-1}\widehat{\mathbf{e}}_n, \qquad k = 2, 3, \ldots, n;
\end{aligned}$$

These equations confirm the desired result. To see this, write them out column-by-column:

$$P^{-1}AP = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -a_0 & -a_1 & -a_2 & \cdots & -a_{n-1} \end{bmatrix}, \quad P^{-1}B = P^{-1}v_n = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}. \quad ////$$

B.2. Theorem. *For constant matrices $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$, the following are equivalent:*

(a) *The matrix pair (A, B) is controllable.*

(b) *Any list of n complex numbers $\lambda_1, \ldots, \lambda_n$ that is closed under conjugation can be obtained as the list of eigenvalues of $A+BF$ for some feedback matrix $F \in \mathbb{R}^{m\times n}$.*

Proof. (b⇒a) Choose any n distinct real numbers $\lambda_k \notin \sigma(A)$ and let $\Sigma = \{\lambda_1, \ldots, \lambda_n\}$. Use (b) to produce a matrix $F \in \mathbb{R}^{m\times n}$ whose eigenvalues are precisely $\lambda_1, \ldots, \lambda_n$, and let x_k denote the eigenvector of $A+BF$ whose eigenvalue is λ_k. Notice that the set of eigenvectors $\{x_1, \ldots, x_n\}$ must be linearly independent, because each eigenvector corresponds to a different eigenvalue.

Now by definition, we have for each k that

$$(A + BF)x_k = \lambda_k x_k, \qquad \text{i.e., } x_k = (\lambda_k I - A)^{-1}BFx_k.$$

(The indicated matrix is invertible precisely because $\lambda_k \notin \sigma(A)$.) But we know that $(\lambda I - A)^{-1}$ has a representation as

$$(\lambda I - A)^{-1} = \sum_{j=1}^{n} r_j(\lambda)A^{j-1} \qquad \forall \lambda \in \mathbb{C} \setminus \sigma(A)$$

for some rational functions $r_1, \ldots, r_n$ defined in $\mathbb{C} \setminus \sigma(A)$. It follows that for each k,

$$\begin{aligned} x_k &= \sum_{j=1}^{n} r_j(\lambda_k)A^{j-1}BFx_k \\ &= \left[\, B \,|\, AB \,|\, \cdots \,|\, A^{n-1}B \,\right] \begin{bmatrix} r_1(\lambda_k)Fx_k \\ r_2(\lambda_k)Fx_k \\ \vdots \\ r_n(\lambda_k)Fx_k \end{bmatrix}. \end{aligned}$$

In particular, each x_k lies in $\mathrm{Im}(\mathcal{C}) = \mathcal{A}(t;0)$. But since the vectors $x_1, \ldots, x_n$ are linearly independent in $\mathbb{R}^n$, it follows that $\mathcal{A}(t;0) = \mathbb{R}^n$, so the system is controllable.

(a⇒b) [Case $m = 1$ only.] Let p be the characteristic polynomial of A, i.e.,

$$p(\lambda) = \det(\lambda I - A) = \lambda^n + a_{n-1}\lambda^{n-1} + \ldots + a_1\lambda + a_0.$$

Then use the desired eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ to build the characteristic polynomial you would like to have instead, namely,

$$q(\lambda) \stackrel{\text{def}}{=} (\lambda - \lambda_1)(\lambda - \lambda_2)\cdots(\lambda - \lambda_n) = \lambda^n + \alpha_{n-1}\lambda^{n-1} + \ldots + \alpha_1\lambda + \alpha_0.$$

To swap the a_k for the α_k by feedback, first use Theorem A.6 to choose an invertible "coordinate change" matrix P such that the revised coefficients $\widetilde{A} = P^{-1}AP$, $\widetilde{B} = P^{-1}B$ have the controllable canonical form

$$\widetilde{A} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -a_0 & -a_1 & -a_2 & \cdots & -a_{n-1} \end{bmatrix}, \qquad \widetilde{B} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}.$$

Then notice that for any $1 \times n$ feedback matrix $\widetilde{F} = [f_0\ f_1\ \cdots f_{n-1}]$, the product $\widetilde{B}\widetilde{F}$ is a square matrix whose first $n-1$ rows are zero and whose last row agrees with $\widetilde{F}$. Thus in the transformed coordinates, the feedback system has dynamics given by

$$\widetilde{A} + \widetilde{B}\widetilde{F} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ f_0 - a_0 & f_1 - a_1 & f_2 - a_2 & \cdots & f_{n-1} - a_{n-1} \end{bmatrix}.$$

This is precisely the form of the matrix in (**): in particular, its characteristic polynomial is

$$\left|\lambda I - (\widetilde{A} + \widetilde{B}\widetilde{F})\right| = \lambda^n + (a_{n-1} - f_{n-1})\lambda^{n-1} + \ldots + (a_1 - f_1)\lambda + (a_0 - f_0).$$

To give $\widetilde{A} + \widetilde{B}\widetilde{F}$ the preassigned eigenvalues $\lambda_1, \ldots, \lambda_n$, we want this polynomial to match q identically. That is, we want

$$f_k = a_k - \alpha_k, \qquad k = 0, \ldots, n-1.$$

This choice will position the eigenvalues of the transformed system correctly. To treat the original system, simply define the feedback matrix $F = \widetilde{F}P^{-1}$. Then one has

$$A + BF = P\widetilde{A}P^{-1} + (P\widetilde{B})(\widetilde{F}P^{-1}) = P(\widetilde{A} + \widetilde{B}\widetilde{F})P^{-1},$$

so the feedback systems have the dynamic coefficient matrices that are equivalent in the sense discussed above. In particular, the eigenvalues of $A + BF$ are the same as those of $\widetilde{A} + \widetilde{B}\widetilde{F}$, namely, the given numbers $\lambda_1, \ldots, \lambda_n$. ////

Note that when $m = 1$, positioning the n eigenvalues of the feedback system uses up all n degrees of freedom available in the choice of the feedback matrix F. When $m > 1$, however, there are some unused degrees of freedom which may be used to arrange other aspects of good behaviour. Just what those are is an advanced topic we will not pursue here.

Moving the other direction, suppose the given system is uncontrollable. Deciding which eigenvalues can be shifted by linear feedback, and in particular if it is possible to move all eigenvalues to the left half of the complex plane, is an advanced topic beyond the scope of these notes. See Wonham for details.

C. Linear Output Feedback

In real-life control systems, not all states are available for feedback. A more accurate model is provided by

$$\begin{aligned} \dot{\mathbf{x}} &= A\mathbf{x} + B\mathbf{u}, \\ \mathbf{y} &= C\mathbf{x}, \end{aligned} \tag{$\dagger$}$$

where A is the $n \times n$ system matrix, B is the $n \times m$ matrix of control coefficients, and C is a new ingredient: an "observation matrix" of size $p \times n$. The vector $\mathbf{y}$ holds p numbers, based on the n components of the system state $\mathbf{x}$: of course, $p \leq n$, and strict inequality is typical. (In case $p = n$, $C = I$, we have our previous situation with $\mathbf{y} \equiv \mathbf{x}$.)

State Estimation. If we can't measure $\mathbf{x}$ directly, we have to guess it, using all the information at our disposal ... namely, the controls we have used, and the observations in $\mathbf{y}$. Try to "estimate" $\mathbf{x}(t)$ using $\mathbf{z}(t)$, where $\mathbf{z}$ is generated by solving

$$\dot{\mathbf{z}}(t) = A\mathbf{z} + B\mathbf{u} + L(C\mathbf{z} - \mathbf{y}), \qquad \mathbf{z}(0) = \zeta. \tag{‡}$$

So the dynamics of $\mathbf{z}$ mimic those of $\mathbf{x}$, but a new term appears: the difference

$$C\mathbf{z}(t) - \mathbf{y}(t) = C\left(\mathbf{z}(t) - \mathbf{x}(t)\right)$$

between the predicted output $C\mathbf{z}$ and the actual system output $\mathbf{y} = C\mathbf{x}$ comes in through some coefficient matrix L. This difference is a linear function of the *estimation error*

$$\mathbf{e}(t) = \mathbf{z}(t) - \mathbf{x}(t).$$

If we can't observe all components of $\mathbf{x}$, we will never be able to check all components of $\mathbf{e}$. However, we can use algebra: the estimation error will evolve according to

$$\dot{\mathbf{e}}(t) = \dot{\mathbf{z}}(t) - \dot{\mathbf{x}}(t) = A\left(\mathbf{z} - \mathbf{x}\right) + LC\left(\mathbf{z} - \mathbf{x}\right) = \left[A + LC\right]\mathbf{e}.$$

Eigenvalue Assignment. To make the estimation error converge to 0, we want to choose the coefficient matrix L in (‡) so that $\Re e(\lambda) < 0$ for each eigenvalue λ of $A + LC$. These eigenvalues are the same as the eigenvalues of

$$(A + LC)^T = A^T + C^T L^T,$$

and we know that eigenvalues on the right can be positioned arbitrarily using a suitable L if and only if the matrix pair (A^T, C^T) is controllable. Engineers call the matrix pair (A, C) observable under conditions detailed on HW04. Here we see that

$$(A, C) \text{ observable} \qquad \Longleftrightarrow \qquad (A^T, C^T) \text{ controllable.}$$

Linear Feedback with State Estimation. Feedback connection using an estimator: $\mathbf{u} = F\mathbf{z}$ gives

$$\begin{aligned} \dot{\mathbf{x}} &= A\mathbf{x} + BF\mathbf{z}, \\ \dot{\mathbf{z}} &= A\mathbf{z} + BF\mathbf{z} + LC(\mathbf{z} - \mathbf{x}). \end{aligned}$$

Now this $2n \times 2n$ system is equivalent under coordinate transformation

$$P = \begin{bmatrix} I & 0 \\ I & -I \end{bmatrix} \qquad \text{(note that } P^2 = I, \text{ so } P^{-1} = P\text{)}$$

to a nice block-triangular one whose eigenvalues are precisely

$$\sigma(A + BF) \cup \sigma(A + LC).$$

This is rather splendid: you design the feedback matrix F to give certain eigenvalues to the original system assuming access to all the states; you design the estimator coefficients in L independently; and when you combine these in the simplest possible way, the eigenvalues you originally chose are perfectly preserved. In particular, if all the eigenvalues in both stages lie in the open left half of the complex plane, the system above will have the origin for a stable equilibrium point that attracts all trajectories.

C. Convexity

C.1. Definition. Let X be a real vector space. A subset S of X is called *convex* if, for every pair of points s_0 and s_1 in S, one has

$$s_\alpha \stackrel{\text{def}}{=} (1-\alpha)s_0 + \alpha s_1 \in S \qquad \forall \alpha \in (0,1).$$

C.2. Theorem. *Let $S \subseteq \mathbb{R}^n$ be a convex set. Then*

(a) *If $s_0 \in \operatorname{int} S$ and $s_1 \in \operatorname{cl} S$, then $(1-\alpha)s_0 + \alpha s_1 \in \operatorname{int} S$ for all $\alpha \in (0,1)$;*

(b) *$\operatorname{int} S$ is convex; $\operatorname{int} S = \operatorname{int}(\operatorname{cl} S)$; and if $\operatorname{int} S \neq \emptyset$, then $\operatorname{cl} S = \operatorname{cl}(\operatorname{int} S)$;*

(c) *a point x lies outside $\operatorname{cl} S$ if and only if there exists a nonzero vector v in $\mathbb{R}^n$ such that*
$$\langle v\,,\, s\rangle \leq \langle v\,,\, x\rangle - 1 \qquad \forall s \in S.$$

(d) *a point x lies outside $\operatorname{int} S$ if and only if there exists a nonzero vector v in $\mathbb{R}^n$ such that*
$$\langle v\,,\, s\rangle \leq \langle v\,,\, x\rangle \qquad \forall s \in S.$$

Proof [(c) *and* (d) *only*]. **"(c)⇒"** Suppose $\widehat{x}$ is a point outside $\operatorname{cl} S$. Consider the problem of minimum distance from $\widehat{x}$ to the set $\operatorname{cl} S$:

$$\min\{|s - \widehat{x}| \,:\, s \in \operatorname{cl} S\}.$$

Since the constraint set $\operatorname{cl} S$ is closed and the objective function $s \mapsto |s - \widehat{x}|$ is continuous, with compact sublevel sets, a minimizing s-value certainly exists: call it $\widehat{s}$. ($\widehat{s}$ is "a nearest point" in $\operatorname{cl} S$ to $\widehat{x}$.) Observe that for any point s' in S,

$$\begin{aligned} |\widehat{s} - \widehat{x}|^2 \leq |s' - \widehat{x}|^2 &= |s' - \widehat{s}|^2 + 2\langle s' - \widehat{s}\,,\, \widehat{s} - \widehat{x}\rangle + |\widehat{s} - \widehat{x}|^2 \\ \iff \quad \langle \widehat{x} - \widehat{s}\,,\, s' - \widehat{s}\rangle &\leq \tfrac{1}{2}|s' - \widehat{s}|^2. \end{aligned} \tag{$*$}$$

Now for any fixed s in S, we can substitute $s' = \widehat{s} + \alpha(s - \widehat{s})$ in $(*)$ for any $\alpha \in (0,1)$, thanks to the convexity of the set S. This gives

$$\langle \widehat{x} - \widehat{s}\,,\, \alpha(s - \widehat{s})\rangle \leq \tfrac{1}{2}\alpha^2 |s - \widehat{s}|^2 \quad \forall \alpha \in (0,1).$$

Dividing the result by α and taking the limit as $\alpha \to 0^+$, we find

$$\langle \widehat{x} - \widehat{s}\,,\, s - \widehat{s}\rangle \leq 0.$$

This works for any $s \in S$, and it implies that the vector $v_0 = \widehat{x} - \widehat{s}$ satisfies

$$\begin{aligned} \langle v_0\,,\, s - \widehat{x}\rangle &= \langle v_0\,,\, s - \widehat{s}\rangle + \langle v_0\,,\, \widehat{s} - \widehat{x}\rangle \\ &= \langle v_0\,,\, s - \widehat{s}\rangle - |v_0|^2 \leq 0 - |v_0|^2. \end{aligned}$$

In particular, the vector $\widehat{v} = v_0/|v_0|^2$ obeys

$$\langle \widehat{v}\,,\, s - \widehat{x}\rangle \leq -1,$$

which gives the desired result. (It is important to think about why the vector $\widehat{v}$ is well-defined, i.e., why $v_0 \neq 0$, and to recognize that $\widehat{v}$ may well fail to be a unit vector.)

"(c)⇐" Suppose x is a point for which some vector v satisfies $\langle v\,,\, s - x\rangle \leq -1$ for all s in S. If x were actually a point in $\operatorname{cl} S$, then there would be a sequence of points s_k in S converging to x. Every point in this sequence would satisfy $\langle v\,,\, s_k - x\rangle \leq -1$, and taking the limit as $k \to \infty$ would produce $0 \leq -1$. This cannot be: thus x must lie outside $\operatorname{cl} S$.

"(d)⇒" Consider now the case where the given point $\widehat{x}$ lies outside $\operatorname{int} S$. This is more delicate, because it allows for the possibility that $\widehat{x}$ is a boundary point of S. It is here that part (b) comes into play: since S is convex, knowing $\widehat{x} \notin \operatorname{int} S$ implies that in fact $\widehat{x} \notin \operatorname{int}(\operatorname{cl} S)$. Therefore $\widehat{x}$ is realizable as a limit of a sequence of points x_k lying outside $\operatorname{cl} S$. For each k, we may apply part (c) to find a vector v_k such that

$$\langle v_k\,,\, s - x_k\rangle \leq -1 \quad \forall s \in S,\ \forall k.$$

In particular, if we set $\widehat{v}_k = v_k/|v_k|$, we obtain a sequence of unit vectors in $\mathbb{R}^n$, each satisfying

$$\langle \widehat{v}_k\,,\, s - x_k\rangle \leq 0 \quad \forall s \in S. \tag{$**$}$$

Along a suitable subsequence, we have $\widehat{v}_k \to \widehat{v}$ for some unit vector $\widehat{v}$; taking the limit in $(**)$ above gives the desired result:

$$\langle \widehat{v}\,,\, s - \widehat{x}\rangle \leq 0 \quad \forall s \in S.$$

"(d)⇐" Suppose x is a point for which some vector v satisfies $\langle v\,,\, s - x\rangle \leq 0$ for all s in S. If x were actually a point in $\operatorname{int} S$, then there would be some positive radius $\rho > 0$ such that the open ball $\mathbb{B}(x;\rho)$ is a subset of S. In particular, for every unit vector u in $\mathbb{R}^n$, the point $s = x + \frac{1}{2}\rho u$ would lie in S, and the inequality above would give $\langle v\,,\, u\rangle \leq 0$. This certainly implies $v = 0$. So if the indicated inequality holds for $v \neq 0$, then it must be true that $x \notin \operatorname{int} S$. ////

The situation in parts (c) and (d) of the previous result is so important that we define the terms formally.

C.3. Definition. *Given a convex set S, a point $s \in \operatorname{cl} S$, and a vector v, we say that v* **is an outward normal to S at s,** *and write $v \perp S$ at $\widehat{s}$, exactly when*

$$\langle \widehat{v}\,,\, x - s\rangle \leq 0 \qquad \forall x \in S.$$

The set of all outward normals at s is written $N_S(s)$.

The point of the theory just developed is that for convex sets, boundary points have an easily accessible geometric characterization: a boundary point is one where there exists at least one nonzero outward normal to the set, and the outward normal is described by a simple inequality.

D. Attainable Sets and Boundary Trajectories

Now a new ingredient enters our problem setup: explicit constraints on the control values. Suppose the values of the control input are all required to lie in some preassigned subset U of $\mathbb{R}^m$. Then the dynamics become

$$\dot{x}(t) = Ax(t) + Bu(t), \qquad u(t) \in U, \qquad \text{a.e.} \tag{$*$}$$

The attainable set will be influenced by this new constraint, so we adapt our notation to express this: for each $t > 0$ and $\xi \in \mathbb{R}^n$, we write

$$\mathcal{A}(t; \xi, U) := \left\{ e^{At}\xi + \int_0^t e^{A(t-s)}Bu(s)\, ds \,:\, u(s) \in U \ \forall s \in [0,t] \right\}.$$

To put the notions of convexity to work in our control problem, we assume that the set of allowed input values U is convex. As an instance of the general rule that the affine image of a convex set is again a convex set, we obtain the following result:

D.0. Proposition. *If the control set $U \subseteq \mathbb{R}^m$ is convex, then each attainable set $\mathcal{A}(t; \xi, U)$ ($t > 0$, $\xi \in \mathbb{R}^n$) is convex.*

Proof. Let $\eta_0, \eta_1 \in \mathcal{A}(t; \xi, U)$. This means that there exist $u_0, u_1 \in PWC([0,t]; \mathbb{R}^n)$ such that

$$\eta_0 = e^{At}\xi + \int_0^t e^{A(t-r)}Bu_0(r)\, dr,$$
$$\eta_1 = e^{At}\xi + \int_0^t e^{A(t-r)}Bu_1(r)\, dr.$$

For any $\alpha \in (0,1)$, it follows that

$$\eta_\alpha := (1-\alpha)\eta_0 + \alpha\eta_1 = e^{At}\xi + \int_0^t e^{A(t-r)}Bu_\alpha(r)\, dr,$$

where $u_\alpha(t) := (1-\alpha)u_0(t) + \alpha u_1(t)$. Since U is convex, one has $u_\alpha \in PWC([0,t]; U)$, so the last expression shows that $\eta_\alpha \in \mathcal{A}(t; \xi, U)$, as required. ////

Boundary Trajectories. When the attainable set $\mathcal{A}(T; \xi, U)$ is convex, we can get a good idea of its shape by looking at the boundary. Once we know $\operatorname{bdy} \mathcal{A}(T; \xi, U)$, we can recover at least the closure of the set $\mathcal{A}(T; \xi, U)$ simply by including all points inside the boundary ("taking the convex hull").

As a first attempt to characterize the boundary of $\mathcal{A}(T; \xi, U)$, we look at those points of the boundary that actually belong to the set $\mathcal{A}(T; \xi, U)$—i.e., those boundary points of the form $x(T)$, where x is the state response starting from ξ associated with some piecewise continuous control taking values in U.

D.1. Lemma. *Suppose U is convex, and $(\widehat{u}, x)$ is a control-state pair for $(*)$ on $[0,T]$. If $x(T) \in \operatorname{bdy} \mathcal{A}(T, x(0), U)$, then the following statements about a vector $w \in \mathbb{R}^n$ are equivalent:*

(i) $w \in N_{\mathcal{A}(T,x(0),U)}(x(T))$,

(ii) $\displaystyle\int_0^T \left\langle e^{A^T(T-r)}w\,,\, B[u(r) - \widehat{u}(r)] \right\rangle dr \le 0$ *for all* $u \in PWC([0,T],U)$,

(iii) *for almost every* $r \in [0,T]$, *one has*

$$\left\langle e^{A^T(T-r)}w\,,\, B[v - \widehat{u}(r)] \right\rangle \le 0 \qquad \text{for all } v \in U.$$

Proof. (i⇔ii) By definition of the outward normal, statement (i) is equivalent to

$$\langle w\,,\, \eta - x(T)\rangle \le 0 \qquad \forall \eta \in \mathcal{A}(T; x(0), U).$$

In detail, this means that every $u \in PWC\,([0,T]; U)$ satisfies

$$\left\langle w\,,\, \left(e^{AT}x(0) + \int_0^T e^{A(T-r)}Bu(r)\,dr\right) - \left(e^{AT}x(0) + \int_0^T e^{A(T-r)}B\widehat{u}(r)\,dr\right) \right\rangle \le 0$$

$$\iff \quad \int_0^T \left\langle w\,,\, e^{A(T-r)}B\left(u(r) - \widehat{u}(r)\right) \right\rangle dr \le 0$$

$$\iff \quad \int_0^T \left\langle e^{A^T(T-r)}w\,,\, B\left(u(r) - \widehat{u}(r)\right) \right\rangle dr \le 0.$$

Hence (ii) holds. But the argument is completely reversible.

(ii⇒iii) Pick a time $\theta \in [0,T)$ and a point $v \in U$ arbitrarily. For each fixed $h \in (0, T-\theta)$, define the control

$$u_h(t) = \begin{cases} v, & \text{if } \theta < t < \theta + h, \\ \widehat{u}(t), & \text{otherwise.} \end{cases}$$

Each $u_h(\cdot)$ so defined belongs to $PWC([0,T],U)$, so must confirm statement (ii). Hence, dividing by $h > 0$, we find

$$\frac{1}{h}\int_\theta^{\theta+h} \left\langle e^{A^T(T-r)}w\,,\, B\left(v - \widehat{u}(r)\right)\,dr \right\rangle \le 0.$$

Since $\widehat{u}(\cdot)$ is piecewise continuous by hypothesis, its one-sided limit from the right at time θ exists. So in the limit as $h \to 0^+$ above, the fundamental theorem of calculus gives

$$\left\langle e^{A^T(T-\theta)}w\,,\, B\left(v - \widehat{u}(\theta^+)\right) \right\rangle \le 0.$$

This development works for any $v \in U$ and $\theta \in [0,T)$. Since $u(\theta^+) = u(\theta)$ holds for all $\theta \in [0,T)$ with at most finitely many exceptions, statement (iii) follows. (Just write $r = \theta$.)

(iii⇒ii) Obvious. ////

Controls that drive the state vector to the boundary of the attainable set turn out to satisfy a certain system of equations shown in Definition D.2 below. The definition relies on the **pre-Hamiltonian** function $H\colon \mathbb{R}^n \times \mathbb{R}^n \times U \to \mathbb{R}$ defined by

$$H(x,p,u) := p^T(Ax+Bu) = p \bullet (Ax+Bu).$$

The "Hamiltonian" part of this function's name comes from the analogy between conditions (a)–(b) below and the Hamiltonian equations of motion in classical mechanics. The prefix "pre-" arises because the analogy with classical mechanics gets even better when one considers the "true Hamiltonian" $\mathcal{H}(x,p) := \max_{v\in U} H(x,p,v)$—a figure in more advanced courses on this subject.

D.2. Definition. A (control,state)-pair $(\widehat{u}(\cdot), x(\cdot))$ defined on the interval $[a,b]$ is called *extremal* if there exists a piecewise smooth function $p\colon [a,b] \to \mathbb{R}^n$ obeying these four conditions:

(a) the adjoint equation, $-\dot{p}(t) = \nabla_x H(x(t),p(t),\widehat{u}(t)) = A^T p(t)$ a.e.,

(b) the state equation, $\dot{x}(t) = \nabla_p H(x(t),p(t),\widehat{u}(t)) = Ax(t) + B\widehat{u}(t)$ a.e.,

(c) the maximum condition, $H(x(t),p(t),\widehat{u}(t)) \geq H(x(t),p(t),v)$ for all $v \in U$, a.e.,

(d) the nontriviality condition, $|p(t)| \neq 0\ \forall t \in [a,b]$.

(Here the abbreviation "a.e." stands for "at all but finitely many points of $[0,T]$".†)

Remarks. **0.** Line (b) simply restates the differential equation in $(*)$, so it contains no new information about an admissible (control,state) pair. However, it does combine with line (a) to produce a system of $2n$ differential equations for the $2n$ components of $x(\cdot)$ and $p(\cdot)$ that looks a lot like the "Hamiltonian systems" loved by experts in dynamics around the world.

1. If conditions (a)–(d) hold for a function p, then they also hold for the function αp for any scalar $\alpha > 0$. This is occasionally useful ... one can simplify problem-solving by specifying a numerical value of $|p(T)|$, for example.

2. Condition (d) is essential, because the function $p(t) \equiv 0$ obviously satisfies (a)–(c). Thus if we were to drop condition (d), every control $\widehat{u}$ would be an extremal: the definition would be utterly useless.

3. Condition (c) asserts that for each fixed t, the input value $\widehat{u}(t)$ must maximize the product $p(t)^T Bv = B^T p(t) \bullet v$ over all v in U. Alternative notation for this is

$$\widehat{u}(t) \in \arg\max \{H(x(t),p(t),v)\ :\ v \in U\}.$$

Geometrically, this means that the vector $B^T p(t)$ is an outward normal to U at the point $\widehat{u}(t)$—in symbols,

$$\text{(c)} \iff B^T p(t) \in N_U(\widehat{u}(t)) \qquad \text{a.e. } t \in [a,b].$$

† This abbreviation is commonly used in the theory of the Lebesgue integral, to stand for, "at all points outside some set of Lebesgue measure zero." The theory being presented here extends without difficulty to the more advanced case, and the definition given here is appropriate with that understanding of "a.e." as well.

All the trajectories of the control system that end up on the boundary of the attainable set can be found among its extremal pairs. We prove this next, along with some interesting extensions.

D.3. Theorem. *Assume the conrol set U in $\mathbb{R}^m$ is convex. For any (control,state)-pair $(\widehat{u}, x)$ for system $(*)$ on $[0,T]$, the following are equivalent:*

(i) $x(T) \in \operatorname{bdy} \mathcal{A}(T; x(0), U)$.

(ii) $(\widehat{u}, x)$ *is extremal on* $[0,T]$.

(iii) $(\widehat{u}, x)$ *is extremal on* $[0,T]$, *and the associated costate function* p *obeys*

$$p(t) \in N_{\mathcal{A}(t;x(0),U)}(x(t)) \qquad \forall t \in [0,T].$$

(iv) $x(t) \in \operatorname{bdy} \mathcal{A}(t; x(0), U)$ *for each* $t \in [0,T]$.

Proof. (i⇒ii) The set $\mathcal{A}(T;\xi,U)$ is convex by Prop. D.1. And $x(T) \notin \operatorname{int} \mathcal{A}(T;\xi,U)$ by hypothesis. So by Thm. C.2(d), the set $\mathcal{A}(T;\xi,U)$ has at least one nonzero outward normal vector at $x(T)$. Pick one—call it $w \in N_{\mathcal{A}(T;x(0),U)}(x(T))$—and use it to define $p(t) := e^{A^T(T-t)}w$. Since the matrix exponential is always invertible, the nontriviality of w guarantees that $p(t) \neq 0$ for all t. And, by direct calculation,

$$-\dot{p}(t) = A^T e^{A^T(T-t)} = A^T p(t) = \nabla_x H(x(t), p(t), \widehat{u}(t)) \qquad \text{a.e. } t \in [0,T].$$

Thus conditions (a), (b), and (d) in definition D.2 are all in force. Condition (c) is provided by Lemma D.1.

(ii⇒iii) Suppose $(\widehat{u}, \widehat{x})$ is extremal on $[0,T]$, and let p be an associated costate. Fix any $t \in (0,T]$. Note that $p(r) \neq 0$ for all $r \in [0,t]$, by D.2(d). Rearrange the maximum condition, D.2(c):

$$\begin{aligned} & p(r)^T B\widehat{u}(r) \geq p(r)^T Bv \qquad \forall v \in U, \text{ a.e. } r \in [0,t], \\ \iff & 0 \geq \langle p(r)\,,\, B[v - \widehat{u}(r)] \rangle \qquad \forall v \in U, \text{ a.e. } r \in [0,t], \\ \iff & p(t) \in N_{\mathcal{A}(t;\hat{x}(0),U)}(\widehat{x}(t)) \qquad \text{(by Lemma D.1).} \end{aligned}$$

(iii⇒iv) Apply Theorem C.2(d).

(iv⇒i) Obvious. ////

The previous result shows that *once a trajectory $x(\cdot)$ enters* $\operatorname{int} \mathcal{A}(t;\xi,U)$, *no control can steer it back onto the boundary.* Trajectories which terminate on the boundary must evolve along the boundary for all time. This is why we call them "boundary trajectories." Let's look at some.

D.4. Example: The Rocket Car. Details in class.

Time-Optimal Control: Necessary Conditions. Typically the set U is bounded: in this case, $\mathcal{A}(t;\xi,U)$ is a *bounded* subset of the affine subspace

$$\mathcal{A}(t;\xi) = e^{At}\xi + \operatorname{Im}\left[B \,|\, AB \,|\, \ldots \,|\, A^{n-1}B\right].$$

It thus makes sense to consider the **minimum time problem**

$$\min\{T \geq 0 \,:\, \eta \in \mathcal{A}(T;\xi,U)\} \qquad P(\xi,\eta)$$

Here the initial point ξ and the target point η are prescribed, and the goal is to drive the system from ξ to η as quickly as possible. We will focus on the most important target for stabilization, namely, $\eta = 0$.

D.5. Definition. For any subset S of $\mathbb{R}^n$, the *distance function* $x \mapsto \operatorname{dist}(x;S)$ is defined by

$$\operatorname{dist}(x;S) = \inf\{|x-s| \,:\, s \in S\}.$$

D.6. Theorem. *Fix $\xi \in \mathbb{R}^n$ and let $U \subseteq \mathbb{R}^m$ be bounded. Then for each fixed $T > 0$, there exists a constant $K \geq 0$ such that*

(a) for any $t_1, t_2 \in [0,T]$, one has

$$\operatorname{dist}(y, \mathcal{A}(t_1;\xi,U)) \leq K|t_2 - t_1| \quad \forall y \in \mathcal{A}(t_2;\xi,U).$$

(b) for any $\eta \in \mathbb{R}^n$, the function $d(t) := \operatorname{dist}(\eta; \mathcal{A}(t;\xi,U))$ obeys

$$|d(t_2) - d(t_1)| \leq K|t_2 - t_1| \quad \forall t_1, t_2 \in [0,T].$$

Proof. Choose $R > 0$ so large that the right side of the matrix norm inequality

$$\left\|e^{At} - e^{As}\right\| = \left\|e^{At}\left(I - e^{A(s-t)}\right)\right\| \leq e^{\|A\|T}\left(e^{\|A\|\,|s-t|} - 1\right)$$

is bounded above by $R|t-s|$ for any $s, t \in [0,T]$. Then choose $M > 0$ so large that

$$\left|e^{At}Bu\right| \leq M \qquad \forall t \in [0,T],\ \forall u \in U.$$

We will use R and M to define K below.

(a) Fix $t_2 \in [0,T]$ and let $y \in \mathcal{A}(t_2;\xi,U)$. Then y is the endpoint $x(t_2)$ for some trajectory x with corresponding control u. Two cases arise: if $t_1 \leq t_2$, then the trajectory x provides a point $x(t_1)$ in $\mathcal{A}(t_1;\xi,U)$. Thus

$$\begin{aligned}
&\operatorname{dist}(y;\mathcal{A}(t_1;\xi,U)) \\
&\quad \leq |x(t_2) - x(t_1)| \\
&\quad = \left|\left(e^{At_2} - e^{At_1}\right)\xi + \int_{t_1}^{t_2} e^{A(t_2-s)}Bu(s)\,ds + \int_0^{t_1}\left(e^{A(t_2-t_1)} - I\right)e^{A(t_1-s)}Bu(s)\,ds\right| \\
&\quad \leq R|\xi|(t_2-t_1) + M(t_2-t_1) + TMR(t_2-t_1).
\end{aligned}$$

If, on the other hand, $t_1 > t_2$, then choose any vector v in U and extend the control u generating x to the longer interval $[0, t_1]$ by setting $u(t) = v$ for all $t \in [t_2, t_1]$. In this case,

$$\begin{aligned}
&\operatorname{dist}(y; \mathcal{A}(t_1; \xi, U)) \\
&\quad \le \left| x(t_2) - \left(e^{At_1}\xi + \int_0^{t_2} e^{A(t_1-s)}Bu(s)\,ds + \int_{t_2}^{t_1} e^{A(t_1-s)}Bv\,ds \right) \right| \\
&\quad = \left| \left(e^{At_2} - e^{At_1}\right)\xi + \int_0^{t_2} \left(I - e^{A(t_1-t_2)}\right) e^{A(t_2-s)}Bu(s)\,ds + \int_{t_2}^{t_1} e^{A(t_1-s)}Bv\,ds \right| \\
&\quad \le R|\xi|(t_1 - t_2) + TMR(t_1 - t_2) + M(t_1 - t_2).
\end{aligned}$$

In both cases, we have the same estimate, and it leads to conclusion (a) with the constant $K = TMR + M + R|\xi|$.

(b) Fix $\eta \in \mathbb{R}^n$ and define the function d as indicated. Let $t_1, t_2 \in [0, T]$ be given. Then for any $\varepsilon > 0$, the set $\mathcal{A}(t_2; \xi, U)$ contains a point y satisfying $|y - \eta| \le d(t_2) + \varepsilon$. By part (a), the set $\mathcal{A}(t_1; \xi, U)$ must containt a point x with $|x - y| \le K|t_1 - t_2| + \varepsilon$. Therefore

$$d(t_1) \le |\eta - x| \le |\eta - y| + |y - x| \le d(t_2) + \varepsilon + K|t_1 - t_2| + \varepsilon.$$

Since $\varepsilon > 0$ is arbitrary,

$$d(t_1) \le d(t_2) + K|t_1 - t_2|.$$

Interchanging labels t_1 and t_2 and repeating the argument gives (b). ////

D.7. Theorem. *Fix ξ in $\mathbb{R}^n$. Suppose $U \subseteq \mathbb{R}^m$ is compact and convex.*

(a) If $\eta \in \mathbb{R}^n$ and $T > 0$ obey $\eta \in \operatorname{int} \mathcal{A}(T; \xi, U)$, then there exists $\delta > 0$ so small that $\eta \in \operatorname{int} \mathcal{A}(t; \xi, U)$ for all t in $(T - \delta, T]$.

(b) If the minimum-time problem $P(\xi, \eta)$ has a solution and T denotes the minimum time, then $\eta \in \operatorname{bdy} \mathcal{A}(T; \xi, U)$. Therefore any minimizing control function must be extremal on $[0, T]$.

Proof. **(a)** (Contraposition.) If the conclusion were false, then there would be some sequence $T_k \to T^-$ such that

$$\eta \notin \operatorname{int} \mathcal{A}(T_k; \xi, U) \qquad \forall k.$$

Now each set $\mathcal{A}(T_k; \xi, U)$ is convex by Prop. C.4, so Thm. C.2 provides, for each k, a unit vector v_k satisfying

$$\langle v_k\,,\, \eta \rangle \ge \langle v_k\,,\, s \rangle \qquad \forall s \in \mathcal{A}(T_k; \xi, U).$$

Now fix any y in $\mathcal{A}(T; \xi, U)$. By Thm. D.6(a), some point $s_k \in \mathcal{A}(T_k; \xi, U)$ satisfies

$$|s_k - y| \le K|T_k - T| + \frac{1}{k}.$$

Thus

$$\langle v_k\,,\,\eta\rangle \geq \langle v_k\,,\,s_k\rangle = \langle v_k\,,\,s_k - y\rangle + \langle v_k\,,\,y\rangle \geq \langle v_k\,,\,y\rangle - K|T_k - T| - \frac{1}{k}.$$

Now the sequence v_k has a subsequence converging to some unit vector v. Taking the limit as $k \to \infty$ along this subsequence in the previous inequality, we find

$$\langle v\,,\,\eta\rangle \geq \langle v\,,\,y\rangle\,.$$

Since $y \in \mathcal{A}(T;\xi,U)$ is arbitrary, this inequality shows that $v \perp \mathcal{A}(T;\xi,U)$ at η. In particular, Thm. C.2(c) implies that $\eta \notin \operatorname{int}\mathcal{A}(T;\xi,U)$. Contrapose.

(b) If $P(\xi,\eta)$ has a solution with minimum time T, then $\eta \in \mathcal{A}(T;\xi,U)$. If $\eta \in \operatorname{int}\mathcal{A}(T;\xi,U)$, then statement (a) would contradict the minimality of T. Thus $\eta \in \operatorname{bdy}\mathcal{A}(T;\xi,U)$. ////

Here is the main result of this section.

D.8. Theorem (Maximum Principle). *If the control set U is compact and convex, and the control function $\widehat{u}$ solves the minimum-time problem $P(\xi,\eta)$, then $\widehat{u}$ is extremal. That is, there exists a function p satisfying*

(a) $-\dot{p}(t) = H_x(x(t),p(t),\widehat{u}(t))^T = A^T p(t)$ a.e.,

(b) $\dot{x}(t) = H_p(x(t),p(t),\widehat{u}(t))^T = Ax(t) + B\widehat{u}(t)$ a.e.,

(c) $H(x(t),p(t),\widehat{u}(t)) \geq H(x(t),p(t),v)$ *for all* $v \in U$, a.e.,

(d) $p(T) \neq 0$.

Modified Problems and Transversality Conditions. The theory developed here can be applied to other problems as well. Here are two related situations. In both cases, we assume that U is compact and convex.

A. Minimum time to reach a given set S. In standard form, this problem reads

$$\begin{aligned}\min\{T>0\ :&\dot{x} = Ax + Bu,\ x(0) = \xi,\\ &u(t) \in U \text{ a.e.},\\ &x(T) \in S\}.\end{aligned} \qquad P(\xi,S)$$

It can be reduced to a problem involving the motion of the attainable sets in $\mathbb{R}^n$ as follows:

$$\min\left\{T>0\ :\ \mathcal{A}(T;\xi,U)\cap S \neq \emptyset\right\}.$$

If this minimum is attained, and has the value T, there must be some point η in $S \cap \mathcal{A}(T;\xi,U)$. Of course, $\eta = x(T)$ for some state process x associated with the optimal control $\widehat{u}$. Now since $\eta \in S$, $\widehat{u}$ most certainly solves the ordinary minimum-time problem $P(\xi,\eta)$. (If there were any faster way to steer the system from ξ to η, it

would also be a faster way to hit the target set S.) Therefore $\eta \in \operatorname{bdy} \mathcal{A}(T;\xi,U)$ by Thm. D.7, and $\widehat{u}$ must be an extremal in the sense of Def. C.5. It is also clear from the arguments underlying Thm. D.7 that η cannot be an interior point of the set S. (Why?) When S is convex, the fact that $\eta \in (\operatorname{bdy} S) \cap (\operatorname{bdy} \mathcal{A}(T;\xi,U))$ implies (via a small extension of Thm. C.2(c)) that there exists a unit vector $w \in \mathbb{R}^n$ satisfying

$$\text{both} \quad w \perp \mathcal{A}(T;\xi,U) \quad \text{and} \quad -w \perp S \text{ at } \eta = x(T).$$

This allows us to say more about the adjoint function p we choose to describe the extremality of the optimizing control $\widehat{u}$. In view of Thm. D.3(a), we can replace the nontriviality condition (d) in Def. C.5 with a more informative assertion about $p(T)$. This statement, D.9(d) below, is called a **transversality condition.**

D.9. Theorem (Maximum Principle for $P(\xi,S)$). *If the control set U is compact and convex, and the control function $\widehat{u}$ solves the minimum-time problem $P(\xi,S)$ for a convex set S, then there exists a continuous function p satisfying*

(a) $-\dot{p}(t) = H_x(x(t),p(t),\widehat{u}(t))^T = A^T p(t)$ a.e.,

(b) $\dot{x}(t) = H_p(x(t),p(t),\widehat{u}(t))^T = Ax(t) + B\widehat{u}(t)$ a.e.,

(c) $H(x(t),p(t),\widehat{u}(t)) \geq H(x(t),p(t),v)$ *for all* $v \in U$, a.e.,

(d) $p(T) \neq 0$ *and* $-p(T) \perp S$ *at* $x(T)$.

(In simple problems, the vectors perpendicular to S are easy to find.)

B. A fixed-time problem. Here the terminal time $T > 0$ is given, along with a smooth function $\ell\colon \mathbb{R}^n \to \mathbb{R}$. The problem is

$$\min\{\ell(x(T)) \,:\dot{x} = Ax + Bu,\ x(0) = \xi,\\ u(t) \in U \text{ a.e.}\}.$$

It can be reduced to a problem involving attainable sets as follows:

$$\min\left\{\ell(\eta) \,:\, \eta \in \mathcal{A}(T;\xi,U)\right\}.$$

D.10. Lemma. *Let $\mathcal{A}$ be any convex subset of $\mathbb{R}^n$. If $\ell\colon \mathbb{R}^n \to \mathbb{R}$ is smooth and*

$$\widehat{\eta} \in \arg\min\left\{\ell(\eta) \,:\, \eta \in \mathcal{A}\right\},$$

then either $\nabla\ell(\widehat{\eta}) = 0$ or else $-\nabla\ell(\widehat{\eta}) \perp \mathcal{A}$ at $\widehat{\eta}$.

Proof. Suppose $\widehat{\eta} \in \mathcal{A}$ is a point where $\nabla\ell(\widehat{\eta})$ is not zero, and is not perpendicular to $\mathcal{A}$ at $\widehat{\eta}$. This means that there is some point $a \in \mathcal{A}$ for which

$$\begin{aligned}
&\langle -\nabla\ell(\widehat{\eta})\,,\, a\rangle > \langle -\nabla\ell(\widehat{\eta})\,,\,\widehat{\eta}\rangle\\
&\quad\iff \langle \nabla\ell(\widehat{\eta})\,,\, a - \widehat{\eta}\rangle < 0\\
&\quad\iff \lim_{t\to 0^+} \frac{\ell(\widehat{\eta} + t(a-\widehat{\eta})) - \ell(\widehat{\eta})}{T}\\
&\quad\implies \ell(\widehat{\eta} + t(a-\widehat{\eta})) < \ell(\widehat{\eta}) \text{ for all } t > 0 \text{ sufficiently small}\\
&\quad\implies \widehat{\eta} \notin \arg\min\left\{\ell(\eta) \,:\, \eta \in \mathcal{A}\right\}.
\end{aligned}$$

(The last implication relies upon the convexity of $\mathcal{A}$ to ensure that $\widehat{\eta} + t(a - \widehat{\eta}) \in \mathcal{A}$ for all $t \in (0,1)$.) ////

This observation lets us write down a version of the maximum principle for the fixed-time problem.

D.11. Theorem (Another Maximum Principle). *Let U be a compact convex set, and let $\widehat{u}$ solve the fixed-time problem above. Then there exists a function p such that*

(a) $-\dot{p}(t) = H_x(x(t), p(t), \widehat{u}(t))^T = A^T p(t)$ a.e.,

(b) $\dot{x}(t) = H_p(x(t), p(t), \widehat{u}(t))^T = Ax(t) + B\widehat{u}(t)$ a.e.,

(c) $H(x(t), p(t), \widehat{u}(t)) \geq H(x(t), p(t), v)$ *for all* $v \in U$, a.e.,

(d) $-p(T) = \nabla \ell(x(T))$.

Proof. Notice that the choice $p(\cdot) \equiv 0$ satisfies Def C.5(a)–(c) and gives the first possible conclusion of the lemma above. If the second conclusion of the lemma above holds, then the choice $p(t) = e^{A^T(T-t)} \nabla \ell(x(T))$ satisfies (d) by construction and obeys C.5(a)–(c) by virtue of Thm. D.3(a). ////

Time-Optimal Control: Sufficient Conditions. Although Theorem D.3 gives a complete characterization of boundary trajectories, its counterpart in the theory of necessary conditions (Thm. D.8) does not completely characterize the time-optimal arcs. We know that every optimal control must satisfy conditions D.8(a)–(d), but we do not know if these conditions are also satisfied by some non-optimal controls (spurious "solutions"). This is because neither theorem address the possibility that some fixed point of $\mathbb{R}^n$ might remain on the boundary of $\mathcal{A}(t; \xi, U)$ throughout an interval of time. We account for this possibility in Cor. D.13 below. That result depends upon the following "controllability result", which provides a qualitative analogue of Thm. A.1(a).

D.12. Proposition. *For any control set $U \subseteq \mathbb{R}^m$, if $0 \in \operatorname{int} U$ and (A, B) is controllable, then*

$$0 \in \operatorname{int} \mathcal{A}(t; 0, U) \text{ for all } t > 0.$$

Proof. By assumption, U contains the closed ball $\mathbb{B}[0; \varepsilon]$ for some $\varepsilon > 0$. Let $\widetilde{U} = \mathbb{B}[0; \varepsilon]$: then $\widetilde{U} \subseteq U$, so $\mathcal{A}(t; 0, \widetilde{U}) \subseteq \mathcal{A}(t; 0, U)$. It suffices to prove that $0 \in \operatorname{int} \mathcal{A}(t; 0, \widetilde{U})$ for all $t > 0$.

Suppose not. If $0 \notin \operatorname{int} \mathcal{A}(T; 0, \widetilde{U})$ for some $T > 0$, then by Thm. C.2(c) there

must be some $w \neq 0$ such that

$$\langle w\,,\,0\rangle \geq \langle w\,,\,y(T)\rangle \quad \forall y(T) \in \mathcal{A}(T;0,U)$$
$$\iff 0 \geq w^T e^{AT} \int_0^T e^{-At} Bu(t)\,dt \quad \forall u \in PWC([0,T];\widetilde{U}).$$

Define $v^T = w^T e^{AT}$. Note that the set $\widetilde{U}$ is symmetric about the origin, so the previous inequality actually implies

$$0 = v^T \int_0^T e^{-At} Bu(t)\,dt \quad \forall u \in PWC([0,T];\widetilde{U}).$$

Since the right side is linear in $u(\cdot)$, we can scale it to deduce that

$$0 = v^T \int_0^T e^{-At} Bu(t)\,dt \quad \forall u \in PWC([0,T];\mathbb{R}^m).$$

Just as in the proof of Thm. 1.1, this implies

$$v^T\,[\,B \quad AB \quad A^2B \quad \cdots \quad A^{n-1}B\,] = 0^T,$$

which contradicts the controllability of (A,B). This is impossible, so the theorem must hold. ////

A sufficient condition for the minimum-time problem $P(\xi,0)$ follows directly. (The case of $P(\xi,\eta)$ with $\eta \neq 0$ is somewhat harder to treat, and we omit it.)

D.13. Corollary. *Let $U \subseteq \mathbb{R}^m$ be compact and convex. Suppose also that $0 \in \operatorname{int} U$ and that (A,B) is controllable. Then for any $T > 0$ and $\widehat{u} \in PWC([0,T];U)$ for which the response satisfies $x(T) = 0$, the following assertions are equivalent:*

(a) $\widehat{u}$ is extremal on $[0,T]$;

(b) $\widehat{u}$ solves problem $P(\xi,0)$.

Proof. ($a \Rightarrow b$) We prove that "not (b)" implies "not (a)". Thus, we assume that $x(T) = 0$, but that $\widehat{u}$ is not optimal. This means that there is some time $T_0 < T$ for which $0 \in \mathcal{A}(T_0;\xi,U)$. This implies that for any time $t > T_0$, $\mathcal{A}(t;\xi,U) \supseteq \mathcal{A}(t-T_0;0,U)$. (Why?) We apply this observation with $t = T$ to get

$$0 \in \operatorname{int} \mathcal{A}(T-T_0;0,U) \subseteq \operatorname{int} \mathcal{A}(T;\xi,U).$$

(The first inclusion follows from Prop. D.12.) This implies that 0 is not a boundary point of $\mathcal{A}(T;\xi,U)$, so we know that the control $\widehat{u}$ cannot be extremal on [0,T] by Thm. D.3.

($b \Rightarrow a$) Proved above as Cor. D.5. ////

A Survey of Advanced Topics

Existence Theory. We would like to have a theorem something like this:

D.14. Desirable Statement. *Let $U \subseteq \mathbb{R}^m$ be compact convex. Consider the minimum time problem $P(\xi, \eta)$. If $\eta \in \mathcal{A}(T; \xi, U)$ for some $T > 0$, then $P(\xi, \eta)$ has a solution.*

If we could show that $\mathcal{A}(t; \xi, U)$ is closed for each $t > 0$, then D.14 would be easy to prove. We would simply consider the function

$$d(t) := \operatorname{dist}(\eta; \mathcal{A}(t; \xi, U))$$

first mentioned in Thm. C.5(b). That result showed that d is continuous on $[0, +\infty)$. Hence the set $\{t \geq 0 : d(t) = 0\}$ is a closed set, which contains some $T > 0$ by assumption. Hence this set contains a smallest element, say $\widehat{T}$. Then for $t < \widehat{T}$ we have $d(t) > 0$, so $\eta \notin \mathcal{A}(t; \xi, U)$, while for $\widehat{T}$ we have $0 = d(\widehat{T}) = \operatorname{dist}(\eta; \mathcal{A}(\widehat{T}; \xi, U))$. *If $\mathcal{A}(\widehat{T}; \xi, U)$ is closed,* then this implies $\eta \in \mathcal{A}(\widehat{T}; \xi, U)$ and D.14 is proved.

Unfortunately, our assumptions do not imply that $\mathcal{A}(t; \xi, U)$ is always a closed set. An example given by D. B. Silin, Soviet Math. Doklady **23**(1981), 309–311 has $m = n = 3$, $B = I$, and $U \subseteq \mathbb{R}^3$ compact convex, but $\mathcal{A}(T; \xi, U)$ is not closed for certain times $T > 0$. Silin modifies his example to formulate a version of $P(\xi, 0)$ which satisfies the hypotheses of D.14, but has no solution. Some additional hypotheses are clearly required. Here are two workable alternatives.

Approach 1 (Specialize the control set). Silin (1981, op. cit.) states that if U is compact, convex, *and polyhedral,* then the sets $\mathcal{A}(t; \xi, U)$ are closed for $t \geq 0$. ("Polyhedral" means that U is the intersection of finitely many affine half-spaces, i.e., $U = \{v \in \mathbb{R}^m : \langle g_i\,,\, v\rangle \leq c_i\}$ for a finite collection of vectors $g_1, \ldots, g_N \in \mathbb{R}^m$ and constants $c_1, \ldots, c_N$.) This covers most applications, and turns D.14 into a theorem.

Approach 2 (Generalize the control functions). If we consider Lebesgue measurable control functions $u(\cdot)$ instead of merely the piecewise continuous control functions, we obtain a larger attainable set:

$$\mathcal{A}_m(t; \xi, U) = \left\{ e^{At}\xi + \int_0^t e^{A(t-s)} B u(s)\, ds \,:\, u \in L^\infty([0, t]; U) \right\}.$$

It can be shown that for a compact convex set U,

$$\mathcal{A}_m(t; \xi, U) = \operatorname{cl} \mathcal{A}(t; \xi, U).$$

Hence the sets $\mathcal{A}_m(t; \xi, U)$ are closed, and not too much larger than the sets we have used. Our necessary and sufficient conditions for optimality all go through in this setting (except that the meaning of "a.e." changes from "from all but finitely many points" to "almost everywhere in the Lebesgue sense"), and statement D.14 again becomes a valid theorem.

Appendix: Bang-Bang Controls. In all of our examples, the optimal controls we found took on values on the boundary of the control set U. Is it possible to use a smaller control set and obtain the same attainable sets? A positive answer would be useful to anyone who has to build an optimal control law, since it would be necessary to implement only a subset of the possible control values.

D.15. Definition. Given a set $S \subseteq \mathbb{R}^m$, the *convex hull* of S is the smallest convex set containing S, denoted $\operatorname{co} S$. Analytically,

$$\operatorname{co} S = \left\{ \sum_{i=1}^{r} \lambda_i s_i \,:\, r \in \mathbf{N},\ \lambda_i \geq 0,\ \sum_{i=1}^{r} \lambda_i = 1,\ s_i \in S \right\}.$$

A *multifunction* (short for "multi-valued function") is a mapping whose values are not points, but sets. One example is the mapping $t \rightrightarrows \mathcal{A}(t; \xi, U)$ studied earlier. Another is the multifunction

$$\Gamma(t) = e^{-At} BU := \left\{ e^{-At} Bu \,:\, u \in U \right\}.$$

The *integral of a multifunction* $\Gamma\colon \mathbb{R} \rightrightarrows \mathbb{R}^n$ is defined as follows:

$$\int_a^b \Gamma(t)\, dt = \{ \int_a^b \gamma(t)\, dt \,:\, \gamma\colon [a,b] \to \mathbb{R}^m \text{ is integrable, and } \gamma(t) \in \Gamma(t) \text{ a.e. } t \in [a,b] \}.$$

A multifunction $\Gamma\colon \mathbb{R} \rightrightarrows \mathbb{R}^n$ is *measurable* if the set below is measurable for any open set $\Omega \subseteq \mathbb{R}^n$:

$$\{ t \in \mathbb{R} \,:\, \Gamma(t) \cap \Omega \neq \emptyset \}.$$

D.16. Theorem (Aumann). *If $\Gamma\colon [a,b] \rightrightarrows \mathbb{R}^n$ is measurable, and its values are nonempty compact sets, and there exists $R > 0$ so big that*

$$\Gamma(t) \subseteq \mathbb{B}[0; R] \text{ a.e. } t \in [a,b],$$

then one has

$$\int_a^b \Gamma(t)\, dt = \int_a^b \operatorname{co} \Gamma(t)\, dt,$$

and both these sets are nonempty, compact, and convex.

Our interest in Aumann's theorem derives from the following observation. Suppose that U is compact and convex, while V is a compact set for which $U = \operatorname{co} V$.

Then for any $T > 0$,

$$\begin{aligned}
\mathcal{A}_m(T;\xi,V) &= e^{AT}\xi + e^{AT}\left\{\int_0^T e^{-At}Bu(t)\,dt \,:\, u \in L^\infty([0,T];V)\right\} \\
&= e^{AT}\xi + e^{AT}\int_0^T \Gamma(t)\,dt, \text{ where } \Gamma(t) := e^{-At}BV, \\
&= e^{AT}\xi + e^{AT}\int_0^T \operatorname{co}\Gamma(t)\,dt, \text{ by Aumann's theorem,} \\
&= e^{AT}\xi + e^{AT}\left\{\int_0^T e^{-At}Bu(t)\,dt \,:\, u \in L^\infty([0,T];U)\right\} \\
&= \mathcal{A}_m(T;\xi,U).
\end{aligned}$$

Thus any point reachable using measurable controls in U can be reached in the same time using measurable controls in the smaller set V.

When V is the smallest closed subset of U such that $U = \operatorname{co} V$, then the functions in $L^\infty([0,T];V)$ are called *bang-bang controls.*

To identify the set of bang-bang controls, note first that any set V such that $U = \operatorname{co} V$ must contain all the "extreme points" of U: these are the vectors $\widehat{u}$ which have the following property:

$$\text{If } \widehat{u} = \frac{u_1 + u_2}{2} \text{ for some } u_1, u_2 \in U, \text{ then } u_1 = u_2 = \widehat{u}.$$

(In other words, $\widehat{u}$ is an extreme point of U iff it cannot be realized as the midpoint of a nontrivial line segment contained in U.) The extreme points of U are denoted $\operatorname{ext} U$. The famous Krein-Milman theorem asserts that if U is compact and convex, then $U = \operatorname{cl}\operatorname{co}\operatorname{ext} U$. It follows immediately that the smallest closed set V with the property that $\operatorname{co} V = U$ is the set $V = \operatorname{cl}\operatorname{ext} U$. Thus we have the following result.

D.17. Theorem (The Bang-Bang Theorem). *If U is compact and convex, then for every $T > 0$, one has*

$$\mathcal{A}_m(T;\xi,U) = \mathcal{A}_m(T;\xi,\operatorname{cl}\operatorname{ext} U).$$

In particular, if the min-time problem $P(\xi,\eta)$ with measurable controls has a solution, then it has a bang-bang solution—i.e., a solution $\widehat{u} \in L^\infty([0,T];\operatorname{cl}\operatorname{ext} U)$.

Notice that Theorem D.17 applies only to individual optimal control functions $\widehat{u}(t)$, and not to nonlinear feedback laws of the type we seek in typical problems. Example 17.4 in Hermes and LaSalle, *Functional Analysis and Time-Optimal Control,* illustrates this point by presenting a problem with a solution that is not bang-bang. We note, however, that the solution in this problem is not unique, and that there is still a feedback solution of bang-bang type. This raises the interesting research question: Can the bang-bang theorem be extended to assert that there exists a function $\widehat{u}\colon \mathbb{R}^n \to \operatorname{cl}\operatorname{ext} U$ such that the discontinuous feedback control dynamics $\dot{x} = Ax + B\widehat{u}(x)$ have a solution for every initial point ξ in $\mathbb{R}^n$, and the solution is always precisely the time-optimal trajectory from ξ to 0?