SJÄLVSTÄNDIGA ARBETEN I MATEMATIK

MATEMATISKA INSTITUTIONEN, STOCKHOLMS UNIVERSITET

Kontrollteoretiska ri-funktioner

av

Ernst Cederholm

2020 - No K16

Kontrollteoretiska ri-funktioner

Ernst Cederholm

Självständigt arbete i matematik 15 högskolepoäng, grundnivå

Handledare: Yishao Zhou

2020

Abstract

Spline functions are an important tool when it comes to approximation of data. Magnus Egerstedt and Clyde Martin [1] demonstrated in *Control Theoretic Splines* (2010) how interpolating and smoothing splines can be approached using the theory of linear control systems. This thesis includes a mathematcical background that is sufficient to reconstruct several of their results, including spline functions on the Cartesian plane and on the unit sphere. Egerstedt and Martin demonstrated that the optimal and smooth spline functions converge as the amount of data grows [1]. In this thesis, an error in their proposed limit function was found and corrected.

Keywords

spline, linear system, optimal control, smoothing spline, Hilbert space

Tack

Jag vill rikta ett stort tack till min handledare professor Yishao Zhou som bidragit med teori, förklaringar, allmänna tips och vägledning när så behövts. Denna uppsats hade jag inte kunnat göra utan henne. Det har varit ett sant nöje att ha Yishao som handledare. Våra möten har varit givande och framförallt roliga. Jag har alltid lämnat dem med ett leende på läpparna. Sedan vill jag även tacka professor Annemarie Luger vars lusläsning har varit till stor nytta och mycket uppskattad. Till sist vill jag passa på att tacka Joel Fredin för trevligt sällskap och för att han alltid varit pigg på livliga mattediskussioner.

Innehåll

1 Introduktion

I detta arbete presenteras grunderna i matematisk kontrollteori som behövs för att förstå Magnus Egerstedt och Clyde Martins [1] idéer och resultat om att använda kontrollteori till att generalisera släta ri-funktioner. De visar, vilket kommer att ses i detta arbete, att problemet att skapa interpolerade kurvor till en datamängd, kan översättas till att minimera en norm i ett Hilbertrum. Med samma metod anpassar Egerstedt och Martin släta ri-funktioner efter datamängder som kan tänkas innehålla slumpmässiga fel, på ett sätt att sådana felaktigheter minimeras. Detta görs både i $\mathbb{R}^n$ samt på sfären och exempel på släta ri-funktioner återskapas i detta arbete. Sist visas att under vissa antaganden konvergerar släta ri-funktioner, mot den underliggande kurvan som datan härstammar från, när datamängden växer.

2 Bakgrundsmaterial

Detta kapitel ämnar lägga en tillräcklig matematisk grund som krävs för att förstå Egerstedt och Martins idéer och resultat om släta kontrollteoretiska ri-funktioner i boken *Control theoretic splines* [1].

2.1 Hilbertrum och normminimering

I detta avsnitt byggs teorin upp som behövs för att definiera ett Hilbertrum samt formulera Hilberts projektionssats. Definitioner och satser från detta delkapitel bygger på David Luenbergers bok *Optimization by Vector Space Methods* [2] om inte annat anges.

Till varje vektorrum finns en associerad mängd skalärer. Skalärerna behöver vara element i en algebraisk kropp. I denna uppsats används framför allt de reella talen som skalärer.

Definition 2.1 (vektorrum)**.** *Ett vektorrum X är en mängd element kallade vektorer tillsammans med operationerna addition och skalärmultiplikation. För två godtyckliga element $x, y \in X$ ligger $x + y$ i X och för varje skalär α är $\alpha x \in X$. För operationen addition ska kommutativa och associativa lagen gälla. Det ska existera ett neutralt element θ sådan att $x + \theta = x$ för alla $x \in X$. För skalärmultiplikationen ska associativa lagen gälla och för operationerna tillsammans ska distributiva lagarna gälla. Sist ska $0x = \theta$ och $1x = x$ för alla $x \in X$.*

Definition 2.2 (normerat vektorrum)**.** *Ett normerat vektorrum är ett vektorrum X tillsammans med en norm, en reellvärd funktion som avbildar varje vektor $x \in X$ till ett reellt tal $\|x\|$. En norm uppfyller villkoren att:*

$$\begin{aligned}
&\|x\| \geq 0 \forall x \in X,\ \|x\| = 0 \text{ om och endast om } x = 0,\\
&\|x + y\| \leq \|x\| + \|y\| \text{ för alla } x, y \in X,\\
&\|\alpha x\| = |\alpha| \cdot \|x\| \text{ för alla skalärer } \alpha \text{ och alla vektorer } x \in X.
\end{aligned}$$

Definition 2.3 (pre-Hilbertrum)**.** *Ett pre-Hilbertrum är ett linjärt vektorrum X med en inre produkt definierad på $X \times X$ och som till varje par av vektorer $x, y \in X$ associerar en skalär $\langle x, y \rangle$. Den inre produkten uppfyller, för alla $x, y, z \in X$ och alla skalärer α, villkoren att:*

$$\begin{aligned}
&\langle x, y \rangle = \overline{\langle y, x \rangle},\\
&\langle x + y, z \rangle = \langle x, y \rangle + \langle y, z \rangle,\\
&\langle \alpha x, y \rangle = \alpha \langle x, y \rangle,\\
&\langle x, x \rangle \geq 0,\ \langle x, x \rangle = 0 \text{ om och endast om } x = 0.
\end{aligned}$$

Sats 2.4. *I ett pre-Hilbertsrum utgör $\|x\| = \sqrt{\langle x, x \rangle}$ en norm.*

Definition 2.5 (Cauchyföljd)**.** *Låt n och m vara naturliga tal. En Cauchyföljd är en följd $\{x_n\}$ i ett normerat rum sådan att $\|x_n - x_m\| \to 0$ då $n, m \to \infty$.*

Definition 2.6 (fullständigt)**.** *Ett normerat linjärt vektorrum X är fullständigt om alla Cauchyföljder i X konvergerar i X.*

Definition 2.7 (Hilbertrum)**.** *Ett Hilbertrum är ett fullständigt pre-Hilbertrum.*

Exempel 2.1. *Vektorrummet $\mathbb{R}^n$ tillsammans med den inre produkten*

$$\langle x, y \rangle = \sum_{i=1}^{n} x_i y_i$$

för alla $x, y \in \mathbb{R}^n$ utgör ett Hilbertrum.

Exempel 2.2 nedan bygger på Egerstedt och Martins i [1] och följer deras notation.

Exempel 2.2. *Beteckna med $L_2^m[0, T]$ rummet av reella m-dimensionella kvadratiskt integrerbara funktioner. Enligt författarna är då $L_2^m[0, T]$ med inre produkt*

$$\langle v, w \rangle_{L_2^m} = \int_0^T v^\top(t) w(t) dt$$

ett Hilbertrum.

Med $L_2^m[0, \infty]$ menas

$$\lim_{T \to \infty} L_2^m[0, T]$$

och när det är förstått från sammanhanget vilket intervall $[0, T]$ samt dimension m som betraktas, skrivs i fortsättningen endast L_2.

Definition 2.8 (Gramian)**.** *Betrakta mängden av vektorer $v_1, \cdots, v_n$, $n \in \mathbb{N}$, tillhörande ett vektorrum med en inre produkt. Matrisen G bestående av elementen $G_{ij} = \langle v_i, v_j \rangle$, $i, j \in \{1, \cdots, n\}$ kallas för en gramian.*

Definition 2.9 (Gateaus differential)**.** *Låt $\mathcal{H}$ vara ett Hilbertrum och $F : \mathcal{H} \to \mathbb{R}$. Gateaus differentialen av F i $p \in \mathcal{H}$ längs $q \in \mathcal{H}$ definieras som*

$$\delta F(p, q) = \lim_{\epsilon \to 0} \frac{F(p + \epsilon q) - F(p)}{\epsilon}.$$

Med denna teori går det nu att formulera och bevisa Hilberts projektionssats.

2.1.1 Hilberts projektionssats

Innan Hilberts projektionssats presenteras behövs en definition av vad det innebär att två vektorer är ortogonala, samt en definition av innebörden av att en vektor är ortogonal mot ett delrum.

Definition 2.10. *Att x och y, tillhörande Hilbertrummet $\mathcal{H}$, är ortogonala menas att $\langle x, y\rangle = 0$ och skrivs $x \perp y$. Med att x är ortogonalt mot delrummet $\mathcal{V}$ till $\mathcal{H}$ menas att $\langle x, v\rangle = 0$, $\forall v \in \mathcal{V}$ och skrivs $x \perp \mathcal{V}$.*

Lemma 2.11. *Om x och y är vektorer i ett Hilbertrum och om $x \perp y$ gäller att $\|x+y\|^2 = \|x\|^2 + \|y\|^2$.*

Bevis. Eftersom $\langle x, y\rangle = 0$ och $\langle x, y\rangle = 0$ fås att

$$\|x+y\|^2 = \langle x+y, x+y\rangle = \langle x, x\rangle + \langle y, y\rangle + \langle x, y\rangle + \langle y, x\rangle = \|x\|^2 + \|y\|^2 .$$

□

Lemma 2.12 (parallellogramlagen)**.** *I ett pre-Hilbertrum gäller att $\|x+y\|^2 + \|x-y\|^2 = 2\|x\|^2 + 2\|y\|^2$.*

Bevis. Genom att skriva normerna i termer av den inre produkten fås att

$$\begin{aligned}\|x+y\|^2 + \|x-y\|^2 &= \langle x+y, x+y\rangle + \langle x-y, x-y\rangle \\ &= 2\|x\|^2 + 2\|y\|^2 + \langle y, x\rangle + \langle x, y\rangle + \langle -y, x\rangle + \langle x, -y\rangle \\ &= 2\|x\|^2 + 2\|y\|^2 .\end{aligned}$$

□

Sats 2.13 (Hilberts projektionssats)**.** *Låt p vara en godtycklig vektor i ett Hilbertrum $\mathcal{H}$. Varje slutet delrum V till $\mathcal{H}$ har en unik punkt v_0 sådan att $\|p - v_0\| \leq \|p - v\|$ för alla $v \in V$. Denna punk bestäms entydigt av att $p - v_0 \perp V$.*

Bevis. Antag att $p \notin V$ och låt $\delta = \operatorname{Inf}_{v \in V} \|p - v\|$. Skapa följden $\{v_n\}$ sådan att $\|p - v_n\| \to \delta$ då $n \to \infty$. Enligt parallellogramlagen gäller att

$$\|(v_m - p) + (p - v_n)\|^2 + \|(v_m - p) - (p - v_n)\|^2 = 2\|v_m - p\|^2 + 2\|v_n - p\|^2 ,$$

vilket kan skrivas om som

$$\|v_m - v_n\|^2 = 2\|v_m - p\|^2 + 2\|v_n - p\|^2 - 4\left\|\frac{v_m + v_n}{2} - p\right\|^2 .$$

Eftersom V är ett linjärt delrum gäller att $\frac{v_m + v_n}{2} \in V$ och

$$\|v_m - v_n\|^2 \leq 2\|v_m - p\|^2 + 2\|v_n - p\|^2 - 4\delta^2 \to 0 \text{ då } n, m \to \infty.$$

Alltså är $\{v_n\}$ en Cauchyföljd och eftersom V är ett slutet vektorrum konvergerar följden mot en punkt v_0 vars avstånd från p är minimalt.

Det återstår bara att visa att v_0 bestäms entydigt av att $p - v_0 \perp V$. Antag till en början att $v \not\perp p - v_0$ för något $v \in V$. Antag också - utan förlust av allmängiltighet - att $\|v\| = 1$ samt att $\langle p - v_0, v\rangle = \epsilon \neq 0$. Då är $v_0 + \epsilon v = v_1 \in V$ och

$$\|p - v_1\|^2 = \|p - v_0 - \epsilon v\|^2 = \|p - v_0\|^2 - \langle p - v_0, \epsilon v\rangle - \langle \epsilon v, p - v_0\rangle + \|\epsilon v\|^2 .$$

Sätts de antagna värdena in fås att

$$\|p - v_1\|^2 = \|p - v_0\| - |\epsilon|^2 < \|p - v_0\| ,$$

vilket motsäger antagandet att $\|p - v_0\| \leq \|p - v\| , \forall v \in V$. Antag nu istället att $p - v_0 \perp V$ då följer av Lemma 2.10 att

$$\|p - v\|^2 = \|p - v_0 + v_0 - v\|^2 = \|p - v_0\|^2 + \|v_0 - v\|^2$$

för alla $v \in V$. Alltså är $\|p - v_0\| \leq \|p - v\| , \forall v \in V$. □

Paul Klein [3] visar att Hilberts projektionssats kan användas för att minimera kvadratsumman av residualer i linjär regression. Sättet som Klein visar detta på återges i Exempel 2.3 nedan.

Exempel 2.3 (Minsta kvadratmetoden)**.** *Låt $y \in \mathbb{R}^m$ och låt X vara en reell $m \times n$ matris. Beteckna med $\bar{x}_i$ den i:te raden i X. Betrakta nu problemet att minimera minsta kvadratsumman med avseende på en parametervektor $\beta \in \mathbb{R}^n$, alltså:*

$$\min_{\beta \in \mathbb{R}^n} \sum_{i=1}^{n} (y_i - \bar{x}_i \beta)^2 = \min_{\beta \in \mathbb{R}^n} (y - X\beta)^\top (y - X\beta) = \min_{\beta \in \mathbb{R}^n} \|y - X\beta\|^2 .$$

Problemet går att betrakta som att hitta punkten $y^\star \in im(X)$ som minimerar avståndet mellan y och $y^\star$. Detta kan enligt Hilberts projektionssats göras genom att projicera y på delrummet

$$im(X) = \{z \in \mathbb{R}^m \mid z = X\alpha \text{ för något } \alpha \in \mathbb{R}^n\}.$$

Enligt projektionssatsen är residualerna $y - X\beta \perp im(X)$. Speciellt är $y - X\beta$ ortogonal mot varje kolumn i X, vilket ses genom att välja α_j, i definitionen av im(X) ovan, som kolonvektorn av dimension n med en etta på rad j och nollor annars. Detta ger att:

$$X^\top (y - X\beta) = 0$$

och om $X^\top X$ är inverterbart följer att optimum fås av

$$\beta^\star = (X^\top X)^{-1} X^\top y.$$

2.1.2 Normminimering

Detta avsnitt bygger på personlig kommunikation med Yishao Zhou. Låt c vara en vektor i $\mathbb{R}^n$ och A en $m \times n$-matris med full rang m (och $m < n$). Betrakta normminimeringsproblemet

$$\begin{aligned} &\text{minimera } \|c + x\|^2, \\ &\text{då } Ax = b \end{aligned}$$

i $\mathbb{R}^n$ med Euklidisk norm. Då $b = 0$ löses detta problem med Hilberts projektionssats genom att projicera $-c$ på vektorrummet $\ker(A)$. Den optimala lösningen $x^\star$ är enligt satsen entydig om och endast om skalärprodukten i $\mathbb{R}^n$ uppfyller villkoret att

$$\langle c + x^\star, x \rangle = 0, \ \forall x \in \ker(A).$$

Det innebär även att $c + x^\star$ ligger i $\text{im}(A^\top)$, eftersom $\text{im}(A^\top)$ och $\text{Ker}(A)$ tillsammans spänner upp R^n enligt dimensionssatsen. Av detta existerar en vektor β i R^m sådan att summan

$$c + x^\star = A^\top \beta,$$

vilket även ger att

$$A(c + x^\star) = AA^\top \beta.$$

Eftersom $Ax^\star = 0$ och $AA^\top$ är inverterbar fås att

$$Ac = AA^\top \beta \implies \beta = (AA^\top)^{-1} Ac$$

och då $c + x^\star = A^\top \beta$ ges $x^\star$ av

$$x^\star = -c + A^\top \beta = -(I + A^\top (AA^\top)^{-1} A)c.$$

Lösningen till optimeringsproblemet då $b \neq 0$ fås nu genom variabelbyterna $y = x - \bar{x}$ och $\bar{c} = \bar{x} + c$, där $\bar{x}$ är en vektor i $\mathbb{R}^n$ som uppfyller att $A\bar{x} = b$. Med dessa variabelbyten gäller att

$$Ax = b \iff Ax - A\bar{x} = 0$$

och då även att

$$A(x - \bar{x}) = Ay = 0.$$

Minimeringsproblemet översätts nu till att

$$\begin{aligned} &\text{minimera } \|y + \bar{c}\|^2, \\ &\text{då } Ay = 0. \end{aligned}$$

Analogt med tidigare beräkning fås då att

$$y^\star = -(I + A^\top (AA^\top)^{-1} A)\bar{c},$$

vilket implicerar att

$$\begin{aligned} x^\star &= y^\star + \bar{x} \\ &= -(I + A^\top (AA^\top)^{-1} A)\bar{c} + \bar{x} \\ &= -(I + A^\top (AA^\top)^{-1} A)(\bar{x} + c) + \bar{x}. \end{aligned}$$

Geometriskt kan ovanstående procedur tolkas som att

$$V_b = \{x \in \mathbb{R}^n \mid Ax = b\}$$

först parallellförflytas till $V_0 = \ker(A)$. Sedan beräknas $V_0^\top$ med hjälp av Hilberts projektionsats. Då den optimala lösningen identifierats flyttas den till $V_0^\perp + c$ som skär V_b i en punkt. Detta kan generaliseras till linjära avbildningar $F : \mathcal{H} \to \mathbb{R}^m$, vilket visas nedan.

Låt $F : \mathcal{H} \to \mathbb{R}^m$, där $\mathcal{H}$ är ett Hilbertrum, vara en linjär operator och definiera

$$V_r = \{w \in \mathcal{H} \mid Fw = r,\ r \in \mathbb{R}^m\},$$

samt

$$V_0^\perp = \{w \in \mathcal{H} \mid w \perp V_0\}.$$

Önskvärt är att kunna hitta en punkt $w^\star \in V_r$ sådan att $\|w - p\|^2$ minimeras för en godtycklig punkt $p \in \mathcal{H}$. Definiera också $V_0^\perp + p$ som

$$V_0^\perp + p = \{z \in \mathcal{H} \mid z = w + p,\ \text{för något } w \in V_0^\perp\}.$$

Det tidigare resultatet kan då generaliseras i följande sats.

Sats 2.14. *Låt p vara en godtycklig punkt i Hilbertrummet $\mathcal{H}$. Lösningen till problemet*

$$\begin{aligned} &\min_{w \in \mathcal{H}} \|w - p\|, \\ &\text{då } w \in V_r, \end{aligned}$$

ges av $w^\star$ där $\{w^\star\} = V_r \cap (V_0^\perp + p)$.

Bevis. Parallellförflytta V_r till vektorrummet V_0. En vektor som är ortogonal mot V_0 är ortogonal mot V_r. Hilberts projektionssats ger då att lösningen ligger i $V_0^\perp$. Parallellförflyttas $V_0^\perp$ med p till $V_0^\perp + p$ fås en entydig lösning av $V_r \cap (V_0^\perp + p)$.

□

2.2 Linjära kontrollsystem

Börjar med att definiera e^A där A är en matris som

$$e^A = \sum_{k=0}^{\infty} \frac{A^k}{k!}.$$

Att denna definition är väldefinierad visas av Kenneth Hoffman i [4]. Med detta definierat kan tillståndsformen som kommer vara central i resten av denna uppsats undersökas.

2.2.1 Tillståndsform och överföringsfunktion

Tillståndsformen i Sats 2.15 nedan är formulerad som i [1] medan beviset för satsen bygger på Derek Rowells anteckningar i [5].

Sats 2.15. *Låt A, B och C vara givna konstanta matriser av storlek $n \times n$, $n \times m$ och $p \times n$. Lösningen till tillståndsformen*

$$\begin{cases} \dot{x} = Ax(t) + Bu(t), \\ y = Cx(t), \end{cases} \qquad x(0) = x_0 \text{ och } t \in [0, T], \tag{1}$$

där $x(t) \in \mathbb{R}^n$, $y(t) \in \mathbb{R}^p$ och $u(t) \in \mathbb{R}^m$ ges av

$$y(t) = Ce^{At}x_0 + \int_0^t Ce^{A(t-s)}Bu(s)ds.$$

Bevis. Multiplicera första ekvationen med e^{-At} och skriv den på formen

$$e^{-At}\dot{x} - e^{-At}x(t) = e^{-At}Bu(t).$$

Med hjälp av kedjeregeln kan detta skrivas som

$$\frac{d}{dt}(e^{-At}x(t)) = e^{-At}Bu(t).$$

Integrering av bägge sidor ger

$$\int_0^t \frac{d}{dt}(e^{-At}x(t))dt = e^{-At}x(t) - e^{-A0}x(0) = \int_0^t e^{-As}Bu(s)ds,$$

varav

$$x(t) = e^{At}x(0) + e^{At}\int_0^t e^{-As}Bu(s)ds = e^{At}x_0 + \int_0^t e^{A(t-s)}Bu(s).$$

Insättning i $y = Cx(t)$ ger lösningen

$$y(t) = Ce^{At}x_0 + \int_0^t Ce^{A(t-s)}Bu(s)ds.$$

till kontrollsystemet. □

Egerstedt och Martin [1] noterar också att om

$$l_t(s) = \begin{cases} Ce^{A(t-s)}B, & s \leq t, \\ 0 \quad \text{annars} \end{cases}$$

och

$$L_t(u) = \int_0^T l_t(s)u(s)ds,$$

kan lösningen till kontrollsystemet skrivas som

$$y(t) = Ce^{At}x_0 + L_t(u).$$

Vissa satser och bevis för tillståndsformen kommer att formuleras endast för $y(t)$ och $u(t)$ som skalärvärda funktioner. Det indikeras då av att B och C skrivs med gemener.

Innan det ges exempel på problem som kan skrivas på tillståndsform, definieras ett begrepp relaterat till beviset ovan och som är nödvändigt för den kommande teorin om det linärkvadratiska regulatorproblemet.

Betrakta sytemet $\dot{x} = A(t)x(t)$ för $t_0 \leq t \leq T$ där A är en $n \times n$- matris och $x(t)$ är en n-vektor. För detta system kan tillståndsöverföringsmatrisen definieras.

Definition 2.16 (tillståndsöverföringsmatris)**.** *Matrisen $\Phi(t, s)$ kallas tillståndsöverföringsmatrisen till systemet ovan om*

$$\begin{cases} \frac{\partial}{\partial t}\Phi(t, s) = A(t)\Phi(t, s), \\ \Phi(s, s) = I. \end{cases}$$

Enligt föreläsningsanteckningar [6] från kursen *Dynamiska system och optimal kontrollteori* ges lösningen till systemet

$$\dot{x}(t) = \frac{\partial}{\partial t}\Phi(t, t_0)x(t_0) = A(t)\Phi(t, t_0)x(t_0) = A(t)x(t)$$

av

$$x(t) = \Phi(t, t_0)x(t_0).$$

Insikten att detta är en lösning fås av att sätta in $\Phi(t, t_0)x(t_0)$ i systemet och observera att likheten håller. Om A är konstant ges att $\Phi(t, t_0) = e^{A(t-s)}$.

Ett exempel på ett system som kan skrivas på tillståndsform ges av Torkel Glad och Lennart Ljung i [7]. Deras exempel återges i Exempel 2.4 nedan.

Exempel 2.4. *Inomhustemperaturen för en sommarstuga bestående av ett rum, ett element och endast ytterväggar påverkas av rumsluftstemperatur T_r, utomhustemperaturen T_{ute} och elementets temperatur T_e. Temperaturen för rumsluften ökar proportionellt mot $T_e - T_r$ och minskar proportionellt mot $T_r - T_{ute}$. Detta samband kan beskrivas som att*

$$\dot{T}_r = \alpha_1(T_e - T_r) + \alpha_2(T_r - T_{ute}),$$

där α_1 och α_2 är proportionalitetskonstanter. Temperaturskillnaden för elementet kan beskrivas som

$$\dot{T}_e = -\alpha_3(T_e - T_r) + \alpha_4 w,$$

där α_3 och α_4 också är proportionalitetskonstanter och w är tillagd elektrisk effekt till elementet. Om vektorerna

$$x = \begin{pmatrix} T_r \\ T_e \end{pmatrix},\ y = T_r \text{ och } u = \begin{pmatrix} w \\ T_{ute} \end{pmatrix}$$

införs kan detta system skrivas på tillståndsformen

$$\begin{aligned} \dot{x} &= \begin{pmatrix} -\alpha_2 - \alpha_1 & \alpha_1 \\ \alpha_3 & \alpha_4 \end{pmatrix} x + \begin{pmatrix} 0 & \alpha_2 \\ \alpha_4 & 0 \end{pmatrix} u \\ y &= \begin{pmatrix} 1 & 0 \end{pmatrix} x. \end{aligned}$$

Ett annat exempel ges av Julius Orion Smith i [8] och återges i Exempel 2.5.

Exempel 2.5. *Låt f beteckna kraft, x position, m massa, a acceleration och v hastighet. Från fysiken gäller de kända lagarna $f = ma$, $v = \dot{x}$ och $a = \dot{v}$. Tillståndsformen*

$$\begin{aligned} \begin{pmatrix} \dot{x}(t) \\ \dot{v}(t) \end{pmatrix} &= \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x(t) \\ v(t) \end{pmatrix} + \begin{pmatrix} 0 \\ \frac{1}{m} \end{pmatrix} f(t) \\ y &= \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} x(t) \\ v(t) \end{pmatrix} \end{aligned}$$

beskriver då sambandet mellan styrfunktionen f och positionen y för ett föremål med massa m.

Definitionen av Laplacetransformen genom Definition 2.17 och Definition 2.18 bygger på teori av Eduardo D. Sontag i [9].

Definition 2.17 (Exponentiell tillväxt)**.** *En funktion $f \in L_2[0, \infty]$ vars norm är lokalt integrerbar och uppfyller villkoret att $\|f(t)\| e^{-\sigma t} \to 0$ då $t \to +\infty$ för något $\sigma \in R$ sägs vara av exponentiell tillväxt.*

Definition 2.18 (Laplacetransform)**.** *För en funktion f av exponentiell tillväxt definiera Laplacetransformen $\mathcal{L}[f](s)$ som*

$$\mathcal{L}[f](s) = \int_0^\infty e^{-st} f(t)dt.$$

Denna integral är väldefinierad och analytisk för alla t och för alla $s \in \mathbb{C}$ sådan att $Re(s) \geq \sigma$.

Sats 2.19. *Laplacetransformen $\mathcal{L}$ är en linjär operator.*

Bevis. låt f och g vara funktioner i $L_2[0,\infty]$ samt α och β vara skalärer då gäller att

$$\mathcal{L}[\alpha f+\beta g](s) = \int_0^\infty e^{-st}(\alpha f(t)+\beta g(t))dt = \alpha \int_0^\infty e^{-st} f(t)+\beta \int_0^\infty e^{-st} g(t).$$

□

Sats 2.20. *För Laplacetransformen $\mathcal{L}$ gäller att $\mathcal{L}[\frac{f}{dt}](s) = s\mathcal{L}[f](s)$ om $f(0) = 0$.*

Bevis.

$$\mathcal{L}[\frac{df}{dt}](s) = \int_0^\infty e^{-st}\frac{df}{dt}dt = \left[e^{-st} f\right]_0^\infty + s\int_0^\infty e^{-st} f dt = 0 - f(0) + s\mathcal{L}[f](s)$$

□

Betrakta igen kontrollsystemet som utgörs av ekvation (1). Om det systemet Laplacetransformeras fås följande resultat.

Sats 2.21. *Om u är av exponentiell tillväxt kan Laplacetransformen för utdatan $y(t)$ till tillståndsbeskrivningen, där $x(0) = 0$, skrivas som*

$$Y(s) = H(s)U(s),$$

där överföringsfunktionen $H(s) = C(sI-A)^{-1}B, Y(s) = \mathcal{L}[y](s)$ och $U(s) = \mathcal{L}[u](s)$.

Bevis. Laplacetransformering av tillståndsbeskrivningen, ekvation (1), med $x(0) = 0$ ger att

$$\mathcal{L}[\dot{x}] = A\mathcal{L}[x] + B\mathcal{L}[u],$$

vilket enligt Sats 3.4 kan skrivas som

$$s\mathcal{L}[x] = A\mathcal{L}[x] + B\mathcal{L}[u].$$

Alltså är

$$\mathcal{L}[x] = (sI - A)^{-1}B\mathcal{L}[u]$$

och Laplacetransformen för utdatan ges då av

$$\mathcal{L}[y] = C(sI - A)^{-1}B\mathcal{L}[u].$$

□

Betrakta igen överföringsfunktionen $H(s) = C(sI - A)^{-1}B$. Enligt Cramers regel kan $(sI - A)^{-1}$ skrivas som

$$\frac{\text{adj}(sI - A)}{\det(sI - A)},$$

där adj$(\cdot)$ betecknar transponatet av kofaktormatrisen. Alltså kan överföringsfunktionen skrivas som

$$H(s) = \frac{C\text{adj}(sI - A)B}{\det(sI - A)}.$$

Notera att $\det(sI - A)$ utgör det karaktäristiska polynomet $q(s)$ för A.

2.2.2 Styrbarhet

Detta avsnitt samt de två kommande bygger på teori från Geir Dullerud och Fernando Paginis bok *A Course in Robust Control Theory* [10] om inte annat anges.

Betrakta första ekvationen i tillståndsbeskrivningen

$$\dot{x} = Ax(t) + Bu(t)$$

och låt $x(0) = 0$. Kom ihåg att lösningen till differentialekvationen ges av

$$x(t) = \int_0^t {}^{A(t-s)}Bu(s)ds.$$

En naturlig fråga är vilka värden som $x(t)$ kan anta genom olika val av $u(s)$.

För att svara på frågan betraktas först mängden av alla värden som $x(t)$ kan anta vid tidpunkten t:

$$\mathcal{R}_t = \{\xi \in \mathbb{R}^n \mid \exists u \text{ sådan att } x(t) = \xi\}.$$

Definition 2.22 (styrbarhetsmatris)**.** *Givet tillståndsbeskrivningen*

$$\dot{x} = Ax(t) + Bu(t),$$

kallas matrisen

$$\begin{pmatrix} B & AB & A^2B & \cdots & A^{n-1}B \end{pmatrix}$$

för styrbarhetsmatrisen till systemet.

Definition 2.23 (styrbarhetsdelrum)**.** *Bildrummet till styrbarhetmatrisen*

$$\mathcal{C}_{AB} = \text{im}\begin{pmatrix} B & AB & A^2B & \cdots & A^{n-1}B \end{pmatrix}$$

kallas för styrbarhetsdelrummet.

Definition 2.24 (styrbarhetsgramian)**.** *För varje $t > 0$ definiera $n \times n$ matrisen styrbarhetsgramianen som*

$$\Gamma_t = \int_0^t e^{At} B B^T e^{A^\top t} dt.$$

Sats 2.25 (Cayley-Hamiliton)**.** *Givet en $n \times n$-matris A gäller det att*

$$A^n + a_{n-1}A^{n-1} + a_{n-2}A^{n-2} + \cdots + a_0 I = 0,$$

där a_i, $0 \leq i \leq n-1$, är koefficienterna för det karaktäristiska polynomet av A.

Beviset som följer är hämtat från föreläsningsanteckningar [11] från kursen *Dynamiska system och optimal kontrollteori*.

Bevis. Låt det karaktäristiska polynomet $p(s)$ vara

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

och låt $B(s) = \{b_{ij}(s)\}$ vara transponatet av kofaktor matrisen av $(A - sI)$, där A är en $n \times n$-matris. Kofaktorerna $b_{ij}(s)$ är polynom av gradtal högst $n-1$. Alltså kan kofaktorerna skrivas som

$$b_{ij}(s) = b_{i,j_0} + b_{i,j_1}s + + b_{i,j_{n-1}}s^{n-1}.$$

Låt $B_k = \{b_{i,j_k}\}$ för $k = 0, 1, \cdots, n-1$. Då kan $B(s)$ skrivas som

$$B(s) = B_0 + B_1 s + + B_{n-1}s^{n-1}.$$

Av likheterna

$$(A - sI)[\text{adj}(A - sI)] = [\text{adj}(A - sI)](A - sI) = \det(A - sI)I$$

gäller det att $(A - sI)B(s) = [\text{adj}(A - sI)]I$. Det följer att

$$(A - sI) = B_0 + B_1 s + + B_{n-1}s^{n-1} = (a_0 + a_1 s + \cdots + a_n s^n).$$

Jämförs koefficienterna framför de olika potenserna av s mellan höger och vänsterled fås att

$$-A^n B_{n-1} = a_n A^n,\ A^n B_{n-1} - A^{n-1}B_{n-2} = a_{n-1}A^{n-1},\ \cdots,\ AB_0 = a_0 I.$$

Adderas ekvationerna ovan fås att $p(A) = 0$. □

Sats 2.26. *Låt A vara en $n \times n$ matris. Då existerar skalära funktioner $\phi_0(t), \cdots, \phi_{n-1}(t)$ sådan att $e^{At} = \phi_0(t)I + \cdots + \phi_{n-1}(t)A^{n-1}$.*

Bevis. Enligt definitionen av e^{At} är

$$e^{At} = I + At + \frac{(At)^2}{2!} + \frac{(At)^3}{3!} + \cdots$$

och skrivs A^k om för $k \geq n$ med hjälp av Cayley-Hamilitons sats följer teoremet. □

Sats 2.27. *För varje $t > 0$ gäller att*

$$\mathcal{R}_t = \mathcal{C}_{AB} = \text{im}\Gamma_t.$$

Bevis. Detta bevisas genom att visa att $\mathcal{R}_t$ är ett delrum till $\mathcal{C}_{AB}$ som är ett delrum till $\text{im}\Gamma_t$. Sist visas att $\text{im}\Gamma_t$ är ett delrum till $\mathcal{R}_t$.

Först visas att $\mathcal{R}_t$ är ett delrum till $\mathcal{C}_{AB}$. Fixera $t > 0$ och välj ett nåbart stadie ξ. Då existerar en styrfunktion u sådan att

$$\xi = \int_0^t e^{A(t-s)}u(s)ds.$$

Används Sats 2.26 fås att

$$\xi = \int_0^t \phi_0(t-s)Bu(s)ds + \cdots + A^{n-1}\int_0^t \phi_{n-1}(t-s)Bu(s)ds.$$

Vilket kan skriva som

$$\xi = \begin{pmatrix} B & AB & \cdots & A^{n-1}B \end{pmatrix} \begin{pmatrix} \int_0^t \phi_0(t-s)u(s)ds \\ \vdots \\ \int_0^t \phi_{n-1}(t-s)u(s)ds \end{pmatrix},$$

varav ξ ligger i bildrummet av kontrollbarhetsmatrisen.

Härnäst visas att $\mathcal{C}_{AB} \subset \text{im}\Gamma_t$, genom att istället visa att $(\text{im}\Gamma_t)^\perp \subset \mathcal{C}_{AB}^\perp$. Från den linjära algebran gäller att

$$(\text{im}\Gamma_t)^\perp = \ker(\Gamma_t^\top) = \ker(\Gamma_t),$$

där den sista likheten gäller för att styrbarhetsgramianen är symmetrisk. Det räcker alltså att visa att om $\xi \in \ker(\Gamma_t)$ är $\xi \in \mathcal{C}_{\mathcal{AB}}{}^\perp$. Låt $\xi \in \ker(\Gamma_t)$ då gäller det att

$$\xi^\top \Gamma_t \xi = 0.$$

Med styrbarhetsgramianen utskriven fås att

$$\xi^\top \Big(\int_0^t e^{As}BB^\top e^{A^\top s}ds \Big)\xi = \int_0^t (\xi^\top e^{As}B)(B^\top e^{A^\top s}\xi)ds.$$

Låt $y(s) = B^\top e^{A^\top s}\xi$ då är

$$\int_0^t y^\top(s)y(s)dt = 0$$

och eftersom integranden är icke negativ måste $y^\top(s) = \xi^\top e^{As}B = 0$ för alla $0 \leq s \leq t$. Detta leder till att även alla derivator

$$\frac{d^k y^\top}{ds^k}$$

evaluerade i punkten noll är lika med noll för alla $k \geq 0$ och därmed är

$$\frac{d^k y^\top}{ds^k}\Big|_{s=0} = \xi^\top A^k B.$$

Detta ger att

$$\xi^\top \begin{pmatrix} B & AB & \cdots & A^{n-1}B \end{pmatrix} = 0,$$

varav $\xi \in \mathcal{C}_{\mathcal{AB}}^\perp$ vilket skulle visas.

Sist visas att $\text{im}\Gamma_t$ är ett delrum till $\mathcal{R}_t$. Välj en godtycklig tid $t > 0$ och $\xi \in \text{im}\Gamma_t$. Per definition existerar det v i $\mathbb{R}^m$ sådan att

$$\xi = \Gamma_t v.$$

Definiera nu

$$u(s) = B^\top e^{A^\top(t-s)}v, \text{ för } 0 \leq s \leq t.$$

Då är lösningen till $\dot{x} = Ax + Bu$, då $x(0) = 0$, vid tidpunkten t

$$\begin{aligned} x(t) &= \int_0^T e^{A(t-s)}Bu(s)ds = \int_0^T e^{A^\top(t-s)}BB^\top e^{A^\top(t-s)}vds \\ &= \int_0^T e^{A^\top(s)}BB^\top e^{A^\top(s)}dsv = \Gamma_t v = \xi. \end{aligned}$$

och per definition av $\mathcal{R}_t$ är $\xi \in \mathcal{R}_t$. □

Om styrbarhetsmatrisen har full rang n kallas paret (A, B) för styrbart. Senare visas att om styrbarhetsgramianen har full rang går det hitta en optimal styrfunktion till punkt till punkproblemet i kapitel 3.

2.2.3 Observerbarhet

Betrakta systemet av ekvationer

$$\begin{cases} \dot{x}(t) = Ax(t), \\ y(t) = Cx(t), \end{cases}$$

med villkoret att $x(0) = x_0$. Lösningen till systemet ges som bekant av $y(t) = Ce^{At}x_0$. Frågan är nu om det går att bestämma x_0 genom att känna till $y(t)$ för ett visst tidsintervall $[0, T]$.

För att få svar på frågan betraktas funktionen ψ, från $\mathbb{R}^n$ till vektorrummet av $\mathbb{R}^p$-värda funktioner, som definieras av

$$x_0 \xrightarrow{\psi} Ce^{At}x_0.$$

Då gäller ekvationen

$$y(t) = \psi x_0$$

som har en unik lösning då $\ker\psi = 0$.

Sats 2.28. *Kärnan av ψ ges av*

$$\ker\psi = \ker C \cap \cdots \cap \ker CA^{n-1} = \ker \begin{pmatrix} C \\ \vdots \\ CA^{n-1} \end{pmatrix}.$$

Bevis. Börjar med att visa att $\ker\psi \subset \ker C \cap \cdots \cap \ker CA^{n-1}$. Låt $x_0 \in \ker\psi$. Per definition är då $Ce^{At}x_0 = 0$ för alla $t \geq 0$. Eftersom

$$0 = \frac{d^k}{dt^k} Ce^{At}x_0\Big|_{t=0} = CA^k x_0,$$

ligger x_0 i nollrummet av CA^k för alla ickenegativa k.

Det återstår endast att visa att $\ker C \cap \cdots \cap \ker CA^{n-1} \subset \ker\psi$. Enligt Sats 2.26 existerar skalära funktioner $\phi_k(t)$ sådan att

$$e^{At} = \phi_0 I + +\phi_{n-1}A^{n-1}$$

för $t \geq 0$. Multipliceras bägge sidor med C från vänster och x_0 från höger ser man att om x_0 ligger i $\ker C \cap \cdots \cap \ker CA^{n-1}$ är också $Ce^{At}x_0 = 0$. □

Matrisen till höger i teoremet ovan kallas för *observerbarhetsmatrisen*. Om $\ker\psi = 0$ kallas paret (C, A) för observerbart.

2.2.4 Minimal realisering

Återgå till tillståndsbeskrivningen

$$\begin{cases} \dot{x} = Ax(t) + Bu(t), \\ y = Cx(t), \end{cases} \quad x(0) = 0 \text{ och } t \in [0, T]. \tag{2}$$

Betrakta nu istället lösningen

$$y(t) = \int_0^T Ce^{A(t-s)}Bu(s)ds$$

till problemet och betraktaekvation (2) som en *realisering* (A, B, C) av funktionen $y(t)$. Två stycken realiseringar (A, B, C) och (A_1, B_1, C_1) sägs vara ekvivalenta ifall likheten

$$\int_0^t Ce^{A(t-s)}Bu(s)ds = \int_0^t C_1e^{A_1(t-s)}B_1u(s)ds \tag{3}$$

håller för alla u och alla $t \geq 0$. Med *ordning* av en realisering menas dimensionen av matrisen A.

Definition 2.29 (Minimal realisering)**.** *En realisering (A, B, C) är minimal om det inte existerar en annan realisering av lägre ordning.*

Sats 2.30. *Två realiseringar (A, B, C) och (A_1, B_1, C_1) är ekvivalenta om och endast om $H(s) = C(sI - A)^{-1}B = C_1(sI - A_1)^{-1}B_1 = H_1(s)$ för alla s där överföringsfunktionen är väldefinierad.*

Bevis. För att visa att två ekvivalenta realiseringar (A, B, C) och (A_1, B_1, C_1) har samma överföringsfunktion räcker det med att ta Laplacetransformen av ekvation (3) som gäller för alla u och alla $t \geq 0$.
För att visa att två realiseringar med samma överföringsfunktioner $H_1(s) = H_2(s)$ är ekvivalenta tas inversa Laplacetransformen av

$$H_1(s)U(s) = H_2(s)U(s).$$

□

Sats 2.31. *Två systemrealiseringar (A, B, C) och (A_1, B_1, C_1) är ekvivalenta om och endast om $Ce^{At}B = C_1e^{A_1t}B_1$ för alla $t \geq 0$.*

Bevis. Om $Ce^{At}B = C_1e^{A_1t}B_1$ håller för alla $t \geq 0$ fås att realiseringarna är ekvivalenta av att

$$\int_0^t Ce^{A(t-s)}Bu(s)ds = \int_0^t C_1e^{A_1(t-s)}B_1u(s)ds$$

håller för alla u och alla $t \geq 0$.

Det återstår att visa om realiseringarna är ekvivalenta måste $Ce^{At}B = C_1e^{A_1t}B_1$ för alla $t \geq 0$. För att göra det skrivs ekvivalens villkoret om som

$$\int_0^t (Ce^{A(t-s)}B - C_1e^{A_1(t-s)}B_1)u(s)ds = 0$$

för alla u och alla $t \geq 0$. Det återstår att visa att faktorn som multipliceras med $u(s)$ är noll. Antag att för något $t \geq 0$ gäller inte $Ce^{At}B = C_1 e^{A_1 t}B_1$. Definiera

$$u(t) = |Ce^{A(t_0+1-t)} - C_1 e^{A_1(t_0+1-t)}B_1|.$$

Om $u(t)$ väljs som styrfunktion fås motsägelsen

$$\int_0^{t_0+1} (Ce^{A(t_0+1-s)} - C_1 e^{A_1(t_0+1-s)}B_1)u(s)ds = \int_0^{t_0+1} |u(s)|^2 ds > 0,$$

eftersom $u(1) > 0$. Alltså måste $Ce^{At}B = C_1 e^{A_1 t}B_1$ för alla $t \geq 0$. □

Sats 2.32. *Två systemrealiseringar (A, B, C) och (A_1, B_1, C_1) är ekvivalenta om och endast om $CA^kB = C_1 A_1^k B_1$ för alla $k \geq 0$.*

Bevis. Att $CA^kB = C_1 A_1^k B_1$ är ett tillräckligt villkor följer av att $Ce^{At}B$ kan skrivas som

$$Ce^{At}B = CB + CABt + CA^2B\frac{t^2}{2} + \cdots,$$

enligt definitionen av e upphöjt i en matris. För att inse att villkoret är nödvändigt noteras att

$$\frac{d^k}{dt^k}Ce^{At}B = CA^k e^{At}B.$$

Eftersom derivatorna till $C_1 e^{A_1 t}B_1$ kan skrivas på samma form och måste vara lika med derivatorna till $Ce^{At}B$ för alla $t \geq 0$, fås att $CA^kB = C_1 A_1^k B_1$ om derivatorna evalueras i $t = 0$. □

Nästa sats är formulerad endast för u och y som skalära funktioner och därför betecknas B samt C med gemener.

Sats 2.33. *Realiseringen (A, b, c) är minimal om och endast om det karaktäristiska polynomet $q(s) = \det(sI - A)$ saknar gemensam delare med polynomet $p(s) = c^\top \mathrm{adj}(sI - A)b$ där*

$$H(s) = \frac{p(s)}{q(s)} = \frac{c^\top \mathrm{adj}(sI - A)b}{\det(sI - A)} = c^\top (sI - A)^{-1} b.$$

Följande bevis bygger på Jitkomut Songsiri anteckningar i [12].

Bevis. Antag att (A, b, c) är minimal men att $p(s)$ och $q(s)$ har gemensam delare. Då existerar två polynom $p_1(s)$ och $q_1(s)$ sådan att $H(s) = \frac{p_1(s)}{q_1(s)}$, där $q_1(s)$ är det karakteristiska polynomet för en matris A_1 av lägre gradtal än A, varav det går att hitta en realisering (A_1, b_1, c_1) av lägre grad än (A, b, c), vilket är en motsägelse.

Antag nu istället att $H(s) = \frac{p(s)}{q(s)}$ är irreducibelt men att (A, b, c) inte är minimal. Då finns per definition av minimal realisering en annan realisering (A_1, b_1, c_1) av lägre gradtal med överföringsfunktion $H_1(s) = \frac{p_1(s)}{q_1(s)}$. Men enligt Sats 2.30 måste $H(s) = \frac{p(s)}{q(s)} = \frac{p_1(s)}{q_1(s)} = H_1(s)$. Alltså har $p(s)$ och $q(s)$ gemensam delare, vilket leder till en motsägelse. □

Sats 2.34. *Varje överföringsfunktion $H(s)$ till en tillståndsform kan skrivas som*

$$H(s) = \begin{pmatrix} H_{11}(s) & \cdots & H_{m1}(s) \\ \vdots & \ddots & \vdots \\ H_{1p}(s) & \cdots & H_{mp}(s) \end{pmatrix}$$

där varje element H_{ij}, för $i \in \{1, \cdots, p\}$ och $j \in \{1, \cdots, m\}$, är en rationell funktion där täljare har lägre gradtal än nämnaren.

Bevis. Enligt Cramers regel gäller att $C(sI - A)^{-1}B = \frac{C\text{adj}(sI-A)B}{\det(sI-A))}$ där $\text{adj}(sI-A)$ är en matris med element bestående av polynom av gradtal lägre än n och eftersom nämnaren är ett polynom av grad n följer satsen. □

Sats 2.35. *Antag att*

$$H(s) = \frac{b_{n-1}s^{n-1} + b_{n-2}s^{n-2} + \cdots + b_0}{s^n + a_{n-1}s^{n-1} + \cdots + a_0}$$

är en överföringsfunktion där täljaren och nämnaren utgör reella polynom. Då finns det en tillståndsformsrealisering (A, B, C) där A är en $n \times n$ matris.

Bevis. Låt

$$Y(s) = \frac{b_{n-1}s^{n-1} + b_{n-2}s^{n-2} + \cdots + b_0}{s^n + a_{n-1}s^{n-1} + \cdots + a_0}U(s)$$

och

$$X(s) = \frac{U(s)}{s^n + a_{n-1}s^{n-1} + \cdots + a_0}. \tag{4}$$

Då kan $Y(s)$ skrivas som

$$Y(s) = (b_{n-1}s^{n-1} + b_{n-2}s^{n-2} + \cdots + b_0)X(s). \tag{5}$$

Om ekvation (4) multipliceras med nämnaren i högerledet varpå inversa Laplacetransformen används fås att

$$x^{(n)} + a_{n-1}x^{(n-1)} + \cdots + a_0 x = u(s). \tag{6}$$

Inför variabelbytet

$$\begin{aligned} x_1 &= x \\ \dot{x}_1 = x_2 &= \dot{x} \\ &\vdots \\ \dot{x}_n = x_{n+1} &= x^{(n)} \end{aligned}$$

Ekvation (6) kan då skrivas som

$$\dot{x}_n = x^{(n)} = u(s) - a_{n-1}x^{(n-1)} - \cdots - a_0 x$$

och på matrisform,

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \vdots \\ \dot{x}_n \end{pmatrix} = \underbrace{\begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -a_0 & -a_1 & -a_2 & \cdots & a_{n-1} \end{pmatrix}}_{A} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix} u(s).$$

Inversa Laplacetranformen av ekvation (5) ger nu att

$$y = b_{n-1}x^{(n-1)} + b_{n-2}x^{n-2} + \cdots + b_0 x = b_{n-1}x_{n-1} + b_{n-2}x_{n-2} + \cdots + b_0 x_1$$

och i matrisform kan y skrivas som

$$y = \begin{pmatrix} b_0 & b_1 & \cdots & b_{n-1} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_3 \end{pmatrix}.$$

□

Tillståndsformen i beviset ovan kallas för styrbar kanonisk form eftersom systemet är styrbart. Vilket visas genom att beräkna styrbarhetsmatrisen, som då blir en triangulärmatris med ettor på diagonalen. Av samma överföringsfunktion som i Sats 2.35 går det även konstruera en ekvivalent tillståndsrealisering (A_1, B_1, C_1) som är skriven på observerbar kanonisk form och vars oberserverbarhetsmatris är en triangulärmatris med ettor på diagonalen. I sådant fall är

$$A_1 = \begin{pmatrix} -a_{n-1} & 1 & 0 & \cdots & 0 \\ -a_{n-2} & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -a_0 & 0 & 0 & \cdots & 1 \end{pmatrix}, \; B_1 = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_{n-1} \end{pmatrix} \text{ och } C_1 = \begin{pmatrix} 1 & 0 & \cdots & 0 \end{pmatrix}.$$

Sats 2.36. *En realisering av (A, b, c) av en överföringsfunktion $H(s)$ är minimal om och endast om (A, b) är styrbar och (c, A) är observerbar.*

Bevis. Antag att realiseringen (A, b, c) är minimal men inte styrbar och har överföringsfunktion

$$H(s) = \frac{c^\top \operatorname{adj}(sI - A)b}{\det(sI - A)}.$$

Då är täljaren och nämnaren relativt prima och eftersom det enligt Sats 2.30 och Sats 2.35 går att konstruera en ekvivalent realisering som är styrbar är det en motsägelse. På analogt sätt kan det visas att en minimal realiseringen även måste vara observerbar.

Det återstår att visa att om realiseringen är observerar och styrbar existerar ingen annan ekvivalent realisering (A_1, b_1, c_1) av lägre grad. Antag att (A_1, b_1, c_1) är en ekvivalent realisering. Då är enligt Sats 3.32

$$c^\top A^k b = c_1^\top A_1^k b_1$$

för alla $k > 0$ vilket implicerar att

$$\begin{pmatrix} c^\top \\ c^\top A \\ \vdots \\ c^\top A^{n-1} \end{pmatrix} \begin{pmatrix} b & Ab & \cdots & A^{n-1}b \end{pmatrix} = \begin{pmatrix} c_1^\top \\ c_1^\top A_1 \\ \vdots \\ c_1^\top A_1^{n-1} \end{pmatrix} \begin{pmatrix} b_1 & A_1 b_1 & \cdots & A_1^{n-1} b_1 \end{pmatrix}.$$

Vänsterledet utgörs av en matrismultiplikation av observerbarhetsmatrisen och styrbarbarhetsmatrisen som båda är av full rang varav deras produkt bildar en $n \times n$-matris. Alltså måste även högerledet ha rang n varav matrisen A_1 måste vara minst en $n \times n$-matris. □

Lemma 2.37. *Det karaktäristiska polynomet till matrisen*

$$A = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ a_0 & a_1 & a_2 & \cdots & a_{n-1} \end{pmatrix}$$

är

$$q(s) = s^n - s^{n-1}a_{n-1} - \cdots - sa_1 - a_0.$$

Bevis. Börjar med att beräkna

$$q(s) = \det(sI - A) = \begin{vmatrix} s & -1 & 0 & \cdots & 0 \\ 0 & s & -1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & s & -1 \\ -a_0 & -a_1 & -a_2 & \cdots & -a_{n-1} \end{vmatrix}_{(n\times n)}.$$

Utveckling utifrån första kolumnen ger

$$q(s) = s\underbrace{\begin{vmatrix} s & -1 & 0 & \cdots & 0 \\ 0 & s & -1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & s & -1 \\ -a_1 & -a_2 & -a_3 & \cdots & s-a_{n-1} \end{vmatrix}}_{((n-1)\times(n-1))}$$

$$+ (-1)^{n+2}a_0\underbrace{\begin{vmatrix} -1 & 0 & 0 & \cdots & 0 \\ s & -1 & 0 & \cdots & 0 \\ 0 & s & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & -1 \end{vmatrix}}_{((n-1)\times(n-1))},$$

varav den sista undertriangulära determinanten blir $(-1)^{(n-1)}$. Alltså är

$$q(s) = s\underbrace{\begin{vmatrix} s & -1 & 0 & \cdots & 0 \\ 0 & s & -1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & s & -1 \\ -a_1 & -a_2 & -a_3 & \cdots & -a_{n-1} \end{vmatrix}}_{((n-1)\times(n-1))} - a_0$$

och återupprepas samma procedur $n-3$ antal gånger på den återstående determinanten blir resultatet att

$$q(s) = s^{n-2}\begin{vmatrix} s & -1 \\ -a_{n-2} & s-a_{n-1} \end{vmatrix} - s^{n-3}a_{n-3} - \cdots - sa_1 - a_0$$

och alltså är

$$q(s) = s^n - s^{n-1}a_{n-1} - \cdots - sa_1 - a_0.$$

□

Enligt Theodore Gamelin i [13] sägs en funktion $f(s)$ vara analytisk i $s = \infty$ om funktionen $g(t) = f(\frac{1}{t})$ är analytisk i $t = 0$. Antag att $g(t)$ är analytisk för $|t| < \rho$. Genom att göra variabelbytet $s = \frac{1}{t}$ och $t = \frac{1}{s}$ kan $f(s)$ beteende i $s = \infty$ studeras genom att studera $g(t)$ i punkten $t = 0$ [13]. Om $f(s)$ är analytisk i $s = \infty$, då har $g(t) = f(\frac{1}{t})$ en potensserie i $t = 0$ som ges av

$$g(t) = \sum_{k=0}^{\infty} b_k t^k, \quad |t| < \rho,$$

och därför kan $f(s)$ representeras av potensserien

$$f(s) = \sum_{k=0}^{\infty} \frac{b_k}{s^k}, \quad |s| > \frac{1}{\rho},$$

enligt [13].

Sats 2.38. *Givet en rationell skalärvärd överföringsfunktion $H(s)$ där täljaren och nämnaren utgör reella polynom och graden av polynomet i nämnaren har högre gradtal än det i täljaren, existerar konstanta matriser A, b och c sådan att*

$$c^\top (sI - A)^{-1} b = H(s).$$

Bevis. Sätt $q(s) = s^n + q_{n-1}s^{n-1} + ... + q_0$. Utvecklas $H(s)$ i ∞ fås att

$$H(s) = b_0 s^{-1} + b_1 s^{-2} + b_2 s^{-3} + ...,$$

där b_i är konstanta skalärer för alla heltal $i \geq 0$. Låt

$$A = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -q_0 & -q_1 & -q_2 & \cdots & -q_{n-1}, \end{pmatrix} \quad b = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_{n-1} \end{pmatrix} \quad \text{och } c = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Att dessa matriser utgör en realisering inses genom att utveckla

$$c^\top (sI - A)^{-1} b$$

i $s = \infty$ och sedan jämföra koefficienterna med utvecklingen av $H(s)$. Det följer då att

$$c^\top (sI - A)^{-1} b = c^\top b s^{-1} + c^\top A b s^{-2} + c^\top A^2 b s^{-3} + \cdots,$$

vilket ger att $c^\top = e_1, c^\top A = e_2, \cdots$ och $c^\top A^{n-1} = e_n$. Detta resulterar i att $c^\top (sI - A)^{-1} b$ kan skrivas som

$$c^\top (sI - A)^{-1} b = b_0 s^{-1} + b_1 s^{-2} + \cdots + b_{n-1} s^{-n} + \cdots,$$

där de n första koefficienterna stämmer överens med de i utvecklingen av $H(s)$. Eftersom $q(s)H(s)$ är ett polynom måste koefficienterna framför termerna i

$$q(s)H(s) = (s^n + q_{n-1}s^{n-1} + \cdots + q_0)(b_0 s^{-1} + b_1 s^{-2} + \cdots)$$

som innehåller negativa exponenter av s vara noll. Detta inkluderar koefficienterna framför s^{-1} som måste vara noll:

$$q_0 b_0 + q_1 b_1 + \cdots + q_{n-1} b_{n-1} + b_n = 0.$$

Av Cayley–Hamiltons sats gäller att $q(A) = 0$ och därför är

$$q_0 c^\top b + q_1 c^\top A b + \cdots + q_{n-1} c^\top A^{n-1} b + c^\top A^n b = c^\top q(A) b = 0.$$

Sätts de två övre vänsterleden lika med varandra fås att $b_n = c^\top A^n b$.

Proceduren ovan resulterar i att om de n första termerna i utvecklingarna är lika följer att nästa term i utvecklingarna kommer att vara lika. Matchas koefficienterna framför s^{-2} termerna istället fås att

$$q_0 b_1 + q_1 b_2 + \cdots + q_{n-1} b_n + b_{n+1} = 0$$

och det gäller även att

$$q_0 c^\top A b + q_1 c^\top A^2 b + \cdots + q_{n-1} c^\top A^n b + c^\top A^{n+1} b = c^\top q(A) A b = 0.$$

Detta ger i sin tur att $b_{n+1} = c^\top A^{n+1} b$. Analogt visas att $b_i = c^\top A^i b$ för alla heltal $i \geq 0$. □

Följdsats 2.39. *Låt A och c vara som i satsen ovan och låt $b = \begin{pmatrix} 0 & 0 & \cdots, c^\top A^{n-2} b \end{pmatrix}^\top$. Antag att $c^\top b = c^\top A b = \cdots = c^\top A^{n-2} b = 0$ då är*

$$H(s) = c^\top (sI - A)^{-1} b = \frac{c^\top A^{-1} b}{q(s)}.$$

Bevis. Från det beviset av Sats 2.38 är $b_i = c^\top A^i b$ och med det nya antagandet är $b_i = 0$ för $i \in \{0, \cdots, n-2\}$. □

Följdsats 2.40. *Realiseringen i Följdsats 2.39 är minimal.*

Bevis. Eftersom $c^\top A^{n-1} b$ och $q(s)$ är relativt prima är realiseringen minimal av Sats 2.33. □

Följdsats 2.41. *Utsignalen y och styrningsfunktionen u som fås från överföringsfunktionen i Följdsats 2.39 uppfyller differentialekvationen:*

$$y^{(n)}(t) + q_{n-1} y^{(n-1)}(t) + \cdots + q_0 y(t) = c^\top A^{n-1} b u(s).$$

Bevis. Om $Y(s)$ och $U(s)$ betecknar Laplacetranformerna av y respektive u följer att

$$Y(s) = H(s)U(s) \Longleftrightarrow q(s)Y(s) = c^\top A^{n-1} b U(s),$$

varav följdsatsen följer av inversa Laplacetransformen. □

2.2.5 Linjärkvadratiska regulatorproblemet

Detta avsnitt behandlar linjärkvadratiska regleringingsproblem som en typ av minsta kvadratproblem i L_2. Upplägget av materialet bygger på personlig kommunikation med Yishao Zhou.

I optimal kontrollteori är det vanligt att ibland behöva lösa det linjärkvadratisk regulatorproblemet (LQ problemet). Ett sådant problem som

dyker upp i detta arbete är att hitta styrningsfunktionen $u(t)$, definierad på intervallet $[0, T]$, som minimerar

$$\eta = \int_0^T \begin{pmatrix} x^\top(t) & u^\top(t) \end{pmatrix} \begin{pmatrix} L & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} x(t) \\ u(t) \end{pmatrix} dt + x^\top(t) Q x(t),$$

då

$$\dot{x} = Ax + Bu, x(0) = x_0$$

är givet och L samt Q är positivt semidefinita matriser.

För enkelhetens skull och i överensstämmelse med uppsatsens mål, betraktas nu endast fall där alla ovannämnda matriser är konstanta. Observera att detta är ett normminimeringsproblem i $L_2[0, T]$, där projektionen på det linjära rummet bestående av lösningar till $\dot{x} = Ax + Bu$, $x(0) = x_0$, sökes. Hur en sådan lösning kan hittas kommer nu att visas.

Lemma 2.42. *Låt A, B och $K(t)$ vara givna matriser. Antag att $\dot{K} = \frac{dK}{dt}$ existerar på intervallet $0 \leq t \leq T$. För x och u som uppfyller ekvationen $\dot{x} = Ax + Bu$ gäller att*

$$\begin{aligned} 0 = &\int_0^T \begin{pmatrix} u^\top(t) & x^\top(t) \end{pmatrix} \begin{pmatrix} 0 & B^\top K(t) \\ K(t)B(t) & \dot{K}(t) + A^\top K(t) + K(t)A \end{pmatrix} \begin{pmatrix} u(t) \\ x(t) \end{pmatrix} dt \\ &- x^\top(t) K(t) x(t) \Big|_0^T. \end{aligned}$$

Bevis. Om x är ett godtyckligt differentierarbart tillstånd och om K är en godtycklig differentierbar matris, då är

$$\begin{aligned} &\int_0^T \big(x^\top(t) \dot{K}(t) x(t) + \dot{x}^\top(t) K(t) x(t) + x^\top K(t) \dot{x}(t) \big) dt - x^\top(t) K(t) x(t) \Big|_0^T \\ &= \int_0^T \frac{d}{dt} (x^\top(t) K(t) x(t)) dt - x^\top(t) K(t) x(t) \Big|_0^T \\ &= x^\top(t) K(t) x(t) \Big|_0^T - x^\top(t) K(t) x(t) \Big|_0^T = 0. \end{aligned}$$

Men eftersom $\dot{x} = Ax + Bu$ gäller även att

$$\begin{aligned} &\int_0^T \big(x^\top(t) \dot{K}(t) x(t) + \dot{x}^\top(t) K(t) x(t) + x^\top K(t) \dot{x}(t) \big) dt \\ &= \int_0^T \big(x^\top(t) \dot{K}(t) x(t) + (Ax + Bu)^\top K(t) x(t) + x^\top K(t) (Ax + Bu) \big) dt \\ &= \int_0^T \begin{pmatrix} u^\top(t) & x^\top(t) \end{pmatrix} \begin{pmatrix} 0 & B^\top K(t) \\ K(t)B(t) & \dot{K}(t) + A^\top K(t) + K(t)A \end{pmatrix} \begin{pmatrix} u(t) \\ x(t) \end{pmatrix} dt, \end{aligned}$$

varav lemmat följer. □

Sats 2.43. *Låt $A, B, L = L^\top$ och $Q = Q^\top$ vara givna matriser. Antag att $\dot{K} = \frac{dK}{dt}$ existerar på intervallet $0 \le t \le T$. Låt även $K(t) = K(t)^T$ uppfylla den så kallade Riccati ekvationen*

$$\dot{K}(t) = -A^\top K(t) - K(t)A + K(t)BB^\top K(t) - L,\ K(T) = Q.$$

Då existerar en styrfunktion u som minimerar

$$\eta = \int_0^T \big(x^\top(t)Lx(t) + u^\top(t)u(t)\big)dt + x^\top(T)Qx(T)$$

för systemet

$$\dot{x}(t) = Ax(t) + Bu(t), x(0) = x_0.$$

Den minimerande styrningen i reglerad form (styrfunktionen beror på x) är

$$u(t) = -B^\top K(t)x(t),$$

medan i oreglerad form (styrfunktionen beror ej på x) är

$$u(t) = -B^\top \lambda(t),$$

där $\lambda(t)$ uppfyller att

$$\dot{\lambda} = -A^\top \lambda(t) - Lx(t)$$

och $\lambda(T) = Qx(T)$.

Bevis. Genom identiteten i Lemma 2.42 och eftersom $\dot{x} = Ax + Bu$ fås att

$$\begin{aligned}\eta = &\int_0^T \begin{pmatrix} u^\top(t) & x^\top(t)\end{pmatrix}\begin{pmatrix} I & 0 \\ 0 & L\end{pmatrix}\begin{pmatrix} u(t) \\ x(t)\end{pmatrix} dt + x^\top(T)Qx(T) \\ &+ \int_0^T \begin{pmatrix} u^\top(t) & x^\top(t)\end{pmatrix}\begin{pmatrix} 0 & B^\top K(t) \\ K(t)B(t) & \dot{K}(t) + A^\top K(t) + K(t)A\end{pmatrix}\begin{pmatrix} u(t) \\ x(t)\end{pmatrix} dt \\ &- x^\top(t)K(t)x(t)\Big|_0^T.\end{aligned}$$

Om integralerna läggs ihop och då

$$\dot{K}(t) = -A^\top K(t) - K(t)A + K(t)BB^\top K(t) - L$$

samt $K(T) = Q$ fås att

$$\begin{aligned}\eta &= \int_0^T \begin{pmatrix} u^\top(t) & x^\top(t)\end{pmatrix}\begin{pmatrix} I & B^\top K(t) \\ K(t)B(t) & K(t)BB^\top K(t)\end{pmatrix}\begin{pmatrix} u(t) \\ x(t)\end{pmatrix} dt \\ &\quad + x^\top(0)K(0)x(0) \\ &= \int_0^T \Big\|u + B^\top K(t)x(t)\Big\|^2 dt + x^\top(0)K(0)x(0).\end{aligned}$$

Det går nu att se att minsta värdet på η fås av $u^\star = -B^\top K(t)x(t)$.

Insättning av $u^\star$ i systemet ger att

$$\dot{x} = Ax + Bu = (A - BB^\top K(t))x.$$

Låt $\lambda = K(t)x(t)$ och $\dot{\lambda} = \dot{K}(t)x(t) + K(t)\dot{x}(t)$. Det råder då att

$$\begin{aligned}\dot{\lambda} &= (-A^\top K(t) - K(t)A + K(t)BB^\top K(t) - L)x(t) + K(t)(A - BB^\top K(t))x \\ &= -A^\top K(t)x(t) - Lx(t) \\ &= -A^\top \lambda - Lx(t)\end{aligned}$$

och $\lambda(T) = K(T)x(T) = Qx(T)$. □

Anmärkning 2.44. *Från Sats 2.43 gäller att den optimala styrningen i oreglerad form utgör ett system av $2n$ stycken differentialekvationer med två punkters randvillkor:*

$$\begin{pmatrix} \frac{dx}{dt} \\ \frac{d\lambda}{dt} \end{pmatrix} = \underbrace{\begin{pmatrix} A & -BB^\top \\ -L & -A^\top \end{pmatrix}}_{:=D} \begin{pmatrix} x \\ \lambda \end{pmatrix}, \; x(0) = x_0 \text{ och } \lambda(T) = Qx(T).$$

Anmärkning 2.45. *Randvärdesproblemet kan under vissa antaganden istället lösas som begynnelsevärdesproblemet, vilket visas nedan. Integreras systemet fås att*

$$\begin{pmatrix} x(t) \\ \lambda(t) \end{pmatrix} = e^{Dt} \begin{pmatrix} x \\ \lambda \end{pmatrix}, \text{ där } \begin{pmatrix} x(T) \\ \lambda(T) \end{pmatrix} = \begin{pmatrix} I \\ Q \end{pmatrix} x(T).$$

Låt

$$e^{D(t-s)} = \underbrace{\begin{pmatrix} \Phi_{11}(t,s) & \Phi_{12}(t,s) \\ \Phi_{21}(t,s) & \Phi_{22}(t,s) \end{pmatrix}}_{:=\Phi(t,s)}.$$

Då är

$$\begin{pmatrix} x(t) \\ \lambda(t) \end{pmatrix} = \Phi(t,s) \begin{pmatrix} x(s) \\ \lambda(s) \end{pmatrix}$$

och

$$\begin{pmatrix} x(0) \\ \lambda(0) \end{pmatrix} = \begin{pmatrix} \Phi_{11}(0,T) & \Phi_{12}(0,T) \\ \Phi_{21}(0,T) & \Phi_{22}(0,T) \end{pmatrix} \begin{pmatrix} x(T) \\ \lambda(T) \end{pmatrix}.$$

Kom ihåg att $\lambda(T) = QX(T)$. Ovanstående ger då ekvationssystemet:

$$\begin{cases} x(0) = (\Phi_{11}(0,T) + \Phi_{12}(0,T)Q)x(T), \\ \lambda(0) = (\Phi_{21}(0,T) + \Phi_{22}(0,T)Q)x(T). \end{cases}$$

Om $\Phi_{11}(0,T) + \Phi_{12}(0,T)Q$ är inverterbar gäller att

$$x(T) = (\Phi_{11}(0,T) + \Phi_{12}(0,T)Q)^{-1}x(0)$$

och därmed är

$$\lambda(0) = (\Phi_{21}(0,T) + \Phi_{22}(0,T)Q)(\Phi_{11}(0,T) + \Phi_{12}(0,T)Q)^{-1}x(0).$$

Eftersom $\lambda(t) = K(t)x(t)$ *gäller även att* $\lambda(0) = K(0)x(0)$ *vilket medför att*

$$K(0) = (\Phi_{21}(0,T) + \Phi_{22}(0,T)Q)(\Phi_{11}(0,T) + \Phi_{12}(0,T)Q)^{-1}.$$

Alltså kan randvärdesproblemet reduceras till initialvärdesproblemet om Riccati ekvationen går att lösa.

Sats 2.46. *Lösningen av Riccati ekvationen i diskussionen ovan är på formen*

$$K(t) = (\Phi_{21}(t,T) + \Phi_{22}(t,T)Q)(\Phi_{11}(t,T) + \Phi_{12}(t,T)Q)^{-1}$$

för alla $0 \leq t \leq T$ *givet att* $(\Phi_{11}(t,T) + \Phi_{12}(t,T)Q)$ *är inverterbar.*

Bevis. Detta visas genom att utveckla

$$\begin{pmatrix} x(t) \\ \lambda(t) \end{pmatrix} = \Phi(t,T) \begin{pmatrix} x(T) \\ \lambda(T) \end{pmatrix}.$$

Då fås ekvationssystemet

$$\begin{cases} x(t) = (\Phi_{11}(t,T) + \Phi_{12}(t,T)Q)x(T), \\ \lambda(t) = (\Phi_{21}(t,T) + \Phi_{22}(t,T)Q)x(T) \end{cases}$$

och $\lambda(t)$ kan skrivas som

$$\lambda(t) = (\Phi_{21}(t,T)\Phi_{22}(t,T)Q)(\Phi_{11}(t,T) + \Phi_{12}(t,T)Q)^{-1}x(t) = K(t)x(t),$$

varav $K(t)$ är på den önskade formen. Eftersom lösningen till LQ problemet är entydig är $K(t)$ lösningen till Ricatiekvationen.

Proceduren som beskrivs ovan är de facto en lösningsmetod till differentialekvationen

$$\begin{pmatrix} \frac{dX}{dt} \\ \frac{dY}{dt} \end{pmatrix} = \begin{pmatrix} A & -BB^\top \\ -L & -A^\top \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix}, \begin{pmatrix} X(T) \\ Y(T) \end{pmatrix} = \begin{pmatrix} I \\ Q \end{pmatrix}$$

och $K(t) = Y(t)X(t)^{-1}$. □

Anmärkning 2.47. *Hittills har det antagits att* $\Phi_{11}(t,T) + \Phi_{12}(t,T)Q$ *är inverterbar. Detta visar sig vara sant om* (A,B) *är styrbart. Notera att* $\Phi_{11}(t,T) + \Phi_{12}(t,T)Q$ *är en tillståndsöverföringsmatris till*

$$\dot{x} = (A - BB^\top K(t))x$$

och att tillståndsöverföringsmatrisen är inverterbar.

Anmärkning 2.48. *Kom ihåg att styrfunktionen för den oreglerade formen är*

$$u(t) = -B^\top \lambda(t),$$

vilket ekvivalent kan skrivas som

$$u(t) = -B^\top K(t)x(t) = -B^\top K(t)\Phi(t,0)x_0$$

där Φ *är tillståndsöverföringsmatrisen för*

$$\dot{x} = (A - BB^\top K(t))x.$$

3 Ri-funktioner

Hela kapitel tre bygger på Magnus Egerstedt och Clyde Martins bok *Control Theoretic Splines* [1] om inte annat anges.

Låt $D = \{(t_i, \alpha_i) \mid i = 1, \cdots, N\}$ vara en datamängd där t_i:na betecknar tidpunkter. Problemet att skapa så kallade ri-funktioner för ett Hilbertrum beskrevs av Grace Whaba på 1970 talet och formulerades då som

$$\min_{f \in L_2[0,T]} \int_0^T f''(t)^2 dt + \lambda (f(t_i) - \alpha_i)^2.$$

Egerstedt och Martin anpassade dessa ri-funktioner till tillståndsformen beskriven i ekvation (1) och problemet kom då att formuleras som

$$\min_{u \in L_2[0,T]} \int_0^T u(t)^2 dt + \lambda (y(t_i) - \alpha_i)^2.$$

Den senare kontrollteoretiska beskrivningen av ri-funktioner är den som kommer studeras i detta kapitel.

3.1 Punkt till punktproblemet

Punkt till punktproblemet innebär huruvida det givet $x(0) = x_0$ och tillståndet x_T, går att få tillståndsformen

$$\dot{x} = Ax(t) + Bu(t)$$

att uppfylla villkoret $x(T) = x_T$, genom att välja styrfunktion $u(t) \in L_2^m[0, T]$.

Kom ihåg från Exempel 2.2 att $L_2^m[0, T]$ är ett Hilbertrum. Med en stor förenkling av definitionen går det att tänka på $L_2^m[0, T]$ som mängden

$$\{w : [0, T] \longrightarrow \mathbb{R}^m \mid \int_0^T w^\top(t) w(t) dt < \infty\}$$

tillsammans med den inre produkt

$$\langle v, w \rangle_{L_2} = \int_0^T v^\top(t) w(t) dt.$$

Lösningen till differentialekvationen i punkt till punktproblemet hittades i beviset för Sats 2.15 men skrivs nu om på formen

$$x(T) = e^{AT} x_0 + \int_0^T e^{A(T-t)} B u(t) dt = e^{AT} x_0 + \Lambda u,$$

där $\Lambda u = \int_0^T e^{A(T-t)}Bu(t)dt$ och $\Lambda : L_2 \longrightarrow \mathbb{R}^n$.

Punkt till punktproblemet innebär alltså att hitta en lösning till

$$x(T) - e^{AT}x_0 = \Lambda u \Longleftrightarrow x(T) - e^{AT}x_0 \in \text{im}(\Lambda)$$

där $\text{im}(\Lambda) = \{z \in \mathbb{R}^n \mid \exists u \in L_2 \text{ sådan att } z = \Lambda u\}$. Definiera operatorn $\Lambda^\star$ från $\mathbb{R}^n$ till L_2 som

$$\langle z, \Lambda v\rangle_{\mathbb{R}^n} = \langle \Lambda^\star z, v\rangle_{L_2}.$$

För att hitta $\text{im}(\Lambda)$ är det praktiskt att först beräkna $\text{im}(\Lambda\Lambda^\star)$. Det gäller att

$$\langle z, \Lambda u\rangle_{\mathbb{R}^n} = z^\top \int_0^T e^{A(T-t)}Bu(t)dt,$$

där R^n-normen definieras som $\langle z, w\rangle_{\mathbb{R}^n} = z^\top w$. Detta kan även skrivas som

$$\int_0^T (B^\top e^{A^\top(T-t)}z)^\top u(t)dt = \langle \Lambda^\star z, u\rangle_{L_2},$$

vilket resulterar i att

$$\Lambda^\star z = B^\top e^{A^\top(T-t)}z.$$

Nu när $\Lambda^\star$ är identifierad kan produkten av den adjungerande operatorn och Λ skrivas som den så kallade styrbarhetsgramianen från Definition 2.24

$$\Gamma = \Lambda\Lambda^\star = \int_0^T e^{A(T-t)}BB^\top e^{A^\top(T-t)}dt. \tag{7}$$

Satsen som följer och tillhörande bevis bygger på teori av Eduardo D. Sontag i [9].

Sats 3.1. *Låt Λ vara en begränsad linjär operator mellan två Hilbertsrum $\mathcal{H}$ till $\mathcal{X}$, då gäller att $im(\Lambda) = im(\Lambda\Lambda^\star)$*

Bevis. Per definition av bildrummet gäller att

$$\text{im}\Lambda\Lambda^\star \subseteq \text{im}\Lambda.$$

Det räcker att visa att

$$(\text{im}\Lambda\Lambda^\star)^\perp \subset \text{ker}\Lambda^\star \subseteq (\text{im}\Lambda)^\perp.$$

För z $\in (\text{im}\Lambda\Lambda^\star)^\perp$ gäller att $\langle \Lambda\Lambda^\star x, z\rangle = 0$ för alla $x \in \mathcal{X}$. Alltså är även

$$0 = \langle \Lambda\Lambda^\star z, z\rangle = \langle \Lambda^\star z, \Lambda^\star z\rangle = \|\Lambda^\star z\|^2,$$

vilket betyder att $\Lambda^\star z = 0$ och därmed är $z \in \text{ker}\Lambda^\star$ samt $z \in (\text{im}\Lambda)^\perp$, ty

$$\langle \Lambda w, z\rangle = \langle w, \Lambda^\star z\rangle = 0$$

för alla $w \in \mathcal{H}$. □

Alltså har punkt till punktproblemet en lösning om och endast om

$$x(T) - e^{AT}x_0 \in \text{im}(\Lambda\Lambda^\star) = \text{im}(\Lambda).$$

Av de styrfunktioner som gör problemet lösbart, ska den med minst norm $u^\star$ hittas i nästa avsnitt.

3.1.1 Minimering av $\|u\|_{L_2}^2$

Med hjälp av Hilberts projektionssats, Sats 2.13, går det att hitta funktionen $u^\star$ som minimerar $\|u\|_{L_2}^2$ sådant att punkt till punktproblemet har en lösning. Låt $\rho = x(T) - e^{AT}x_0$ och kom ihåg att punkt till punktproblemet har en lösning om och endast om $\rho \in \text{im}(\Lambda)$. Bilda mängden

$$V_\rho = \{u \in L_2 \mid \rho = \Lambda u\},$$

alltså de funktioner u_2 som ger upphov till en lösning av punkt till punktproblemet. Den unika funktionen $\{u^\star\}$ fås nu av $V_0^\perp \cap V_\rho$, där

$$V_0^\perp = (\ker\Lambda)^\perp = ((\text{im}\Lambda^\star)^\perp)^\perp = \text{im}\Lambda^\star.$$

Det gäller att $v \in V_0^\perp$ är ekvivalent med att $\Lambda^\star z = v$ för något $z \in \mathbb{R}^n$. Om $\Lambda^\star z = v$ multipliceras med Λ från vänster fås att $\Lambda\Lambda^\star z = \Lambda v$ för något $z \in \mathbb{R}^n$. I fall att v ligger i V_ρ och då även i $V_0^\perp \cap V_\rho$ är

$$\Lambda\Lambda^\star z = \Lambda v = \rho.$$

Om $\Gamma = \Lambda\Lambda^\star$ har maximal rang ges z av $(\Lambda\Lambda^\star)^{-1}\rho$ och då är

$$u^\star = \Lambda^\star z = \Lambda^\star(\Lambda\Lambda^\star)^{-1}\rho.$$

3.1.2 Interpolation

Betrakta kontrollsystemet

$$\begin{cases} \dot{x} = Ax(t) + bu(t), \\ y = c^\top x(t), \end{cases} \qquad x(0) = 0$$

och data

$$D = \{(t_i, \alpha_i) \mid i = 1, \cdots, N\}.$$

En ri-funktion och tillhörande styrfunktion $u \in L_2[0, T]$ som minimerar

$$J(u) = \int_0^T u^2(s)ds$$

kan konstrueras, sådan att ri-funktionen skär punkterna i datamängden ovan. Låt

$$l_{t_i}(s) = \begin{cases} c^\top e^{A(t_i - s)}b, & s \le t_i, \\ 0 \quad \text{annars.} \end{cases}$$

Eftersom $x(0) = 0$ gäller att $y(t_i) = L_{t_i}(u)$. Det interpolerande villkoret är därför att $\alpha_i = L_{t_i}(u)$ för $i = 1, \cdots, N$. För att konstruera den interpolerande funktionen $y(s)$, visas först att funktionerna $\{l_{t_i}(s) \mid i = 1, \cdots, N\}$ är linjärt oberoende.

Sats 3.2. *Låt c och b vara skilda från nollvektorn. Då är funktionerna $\{l_{t_i}(s) \mid i = 1, \cdots, N\}$ linjärt oberoende.*

Bevis. Antag att det existerar skalärer $\beta_1, \cdots, \beta_N$ sådana att för alla s är

$$\beta_1 l_{t_i}(s) + \cdots + \beta_N l_{t_N}(s) = 0.$$

Det behöver visas att alla koefficienterna $\beta_1, \cdots, \beta_N$ då måste vara noll. Om $t_{N-1} < s \leq t_N$ är

$$\beta_N l_{t_N}(s) = 0,$$

vilket implicerar att $\beta_N = 0$. Om istället $t_{N-2} < s \leq t_{N-1}$ är

$$\beta_{N-1} l_{t_{N-1}}(s) + \beta_N l_{t_N}(s) = 0,$$

men eftersom $\beta_N = 0$ fås även att $\beta_{N-1} = 0$. Återupprepas samma procedur tills att $0 \leq s \leq t_1$ fås till sist att även $\beta_1 = 0$ vilket skulle visas. □

Mängden funktioner u som gör kontrollsystemet lösbart är

$$V_\alpha = \{u \in L_2 \mid \alpha_i = L_{t_i}(u),\ i = 1, \cdots, N\}, \tag{8}$$

där $L_{t_i}(u) = \int_0^T l_{t_i}(s)u(s)ds$. Konstruera även mängden

$$V_0 = \{u \in L_2 \mid 0 = L_{t_i}(u),\ i = 1, \cdots, N\}.$$

Eftersom den optimala styrfunktionen $u^\star$ ges av snittet mellan $V_0^\perp$ och V_α konstrueras

$$\begin{aligned} V_0^\perp &= \{u \in L_2 \mid L_{t_i}(u) = 0,\ i = 1, \cdots, N\}^\perp \\ &= \{v \in L_2 \mid \langle v, u \rangle_{L_2} = 0, \forall u \in V_0\}. \end{aligned}$$

Eftersom det finns N stycken oberoende funktioner $l_{t_i}(s)$ för $i = 1, \cdots N$ i V_0, kan den optimala styrningen $u^\star$ kan skrivas som

$$u^\star(s) = \sum_1^N \tau_i l_{t_i}(s), \tag{9}$$

för några skalärer $\tau_1, \cdots, \tau_N$. Av (8) och (9) tillsammans fås ekvationerna:

$$\begin{aligned} y(t_1) &= \sum_1^N \tau_i L_{t_1}(l_{t_i}) \\ &\vdots \\ y(t_N) &= \sum_1^N \tau_i L_{t_N}(l_{t_i}). \end{aligned}$$

De kan på matrisform skrivas som

$$\hat{y} = G\tau = \alpha \Longleftrightarrow \tau = G^{-1}\alpha,$$

där

$$G = \begin{pmatrix} L_{t_1}(l_{t_1}) & \cdots & L_{t_1}(l_{t_N}) \\ \vdots & & \vdots \\ L_{t_N}(l_{t_1}) & \cdots & L_{t_N}(l_{t_N}) \end{pmatrix} = \int_0^T l(s)l^\top(s)ds,$$

$\hat{y} = (y(t_1), \cdots, y(t_N))^\top$, $\alpha = (\alpha_1, \cdots, \alpha_N)^\top$ och $l(s) = (l_{t_1}(s), \cdots, l_{t_N}(s))^\top$. På matrisform kan nu den optimala styrfunktionen skrivas som

$$u^\star(t) = \tau^\top l(t) = \alpha^\top G^{-1} l(t).$$

Med hjälp av programmet RStudio och metoden ovan kan Exempel 3.2.2 på sida 31 i [1] rekonstrueras. Låt

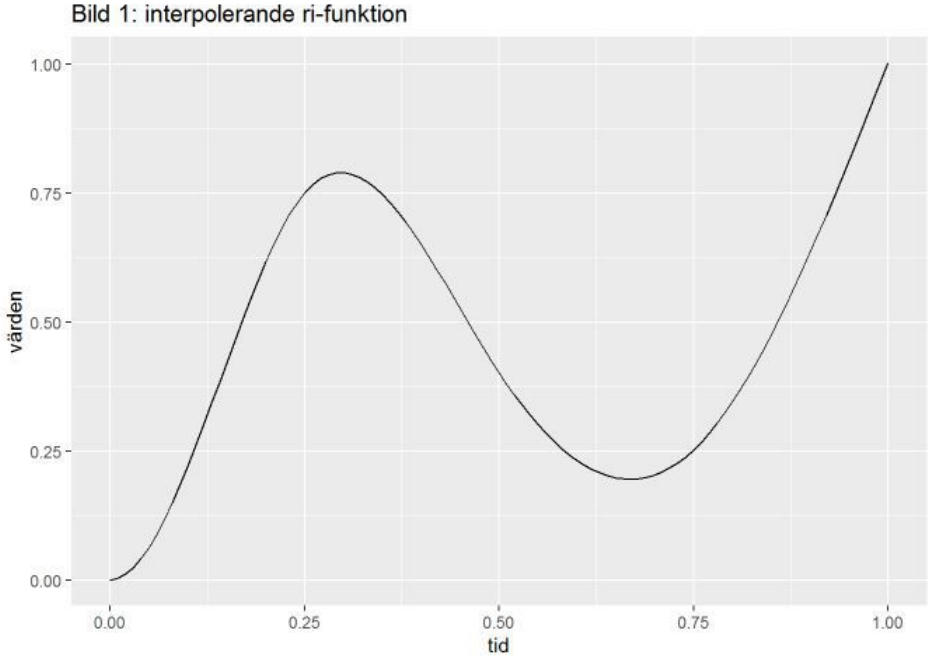
$$T = 1,\ N = 4\ A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix},\ b = \begin{pmatrix} 0 \\ 1 \end{pmatrix},\ \ c = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

och

$$D = \{(t_1, \alpha_1) = (\frac{1}{4}, \frac{3}{4}), (t_2, \alpha_2) = (\frac{1}{2}, \frac{2}{5}), (t_3, \alpha_3) = (\frac{3}{4}, \frac{1}{4}), (t_4, \alpha_4) = (1, 1)\}.$$

Den interpolerande ri-funktionen till exemplet ovan ges av R-koden i bilaga A och visas i Bild 1 nedan.

Bild 1: interpolerande ri-funktion



3.2 Släta ri-funktioner

Betrakta det interpolerande exemplet i föregående avsnitt, men istället för att tvinga den anpassade funktionen genom datapunkterna straffas nu kurvan beroende på avståndet till punkterna.

Betrakta ett observerbart och styrbart kontrollsystem

$$\begin{cases} \dot{x} = Ax(t) + bu(t), \\ y = c^\top x(t), \end{cases} \qquad x(0) = x_0,\ t \in [0, T]$$

och tillhörande data

$$D = \{(t_i, \alpha_i) \mid i = 1, \cdots, N\}.$$

Det går att hitta den funktion $u^\star$ som minimerar

$$J(u, x(0)) = \int_0^T u^2(t)dt + (\hat{y} - \hat{\alpha})^\top Q(\hat{y} - \hat{\alpha}) + x(0)^\top Rx(0),$$

där

$$\hat{y} = (y_1, \cdots, y_n)^\top,\ y_i = ce^{At_i}x_0 + \int_0^{t_i} ce^{A(t_i - s)}bu(s),$$

$\hat{\alpha} = (\alpha_1, \cdots, \alpha_n)^\top$ och Q samt R är positivt definita matriser. Låt

$$l_i(s) = \begin{cases} c^\top e^{A(t_i - s)}b, & t_i \geq s, \\ 0 \quad \text{annars}, \end{cases}$$

och $\beta_i = R^{-1}e^{A^\top t_i}c$. Nu kan y_i skrivas som

$$y_i = ce^{At_i}x_0 + \int_0^T l_i(s)u(s)ds = \langle \beta_i, x_0 \rangle_R + \langle l_i, u \rangle_{L_2}.$$

och det går att konstruerar Hilbertsrummet $\mathcal{H} = L_2[0, T] \times \mathbb{R}^n \times \mathbb{R}^N$ med norm

$$\|(u; x; d)\|_{\mathcal{H}} = \int_0^T u^2(t)dt + x^\top Rx + d^\top Qd$$

och inre produkt

$$\langle (u; x; d), (v; z; f) \rangle_{\mathcal{H}} = \langle u, v \rangle_{L_2} + \langle x, z \rangle_R + \langle d, f \rangle_Q.$$

Den optimala styrfunktionen $u^\star$ fås då av att lösa minimeringsproblemet

$$\min_{(u;x;d) \in \mathcal{H}} \|(u; x; d) - (0; 0; \hat{\alpha})\|$$

sådan att

$$(u; x; d) \in V_0 = \{(u; x; d) \mid d_i = \langle \beta_i, x \rangle_R + \langle l_i, u \rangle_{L_2}\}. \tag{10}$$

Om V_0 är ett slutet delrum av $\mathcal{H}$ följer att $u^\star = V_0 \cap (V_0^\perp + p)$, där $p = (0; 0; \hat{\alpha})$, av Hilberts projektionssats. Att V_0 är slutet visas med följande sats.

Sats 3.3. *V_0 är ett slutet delrum av $\mathcal{H}$.*

Bevis. Betrakta funktionen

$$K_i((u;x)) = \langle \beta_i, x\rangle_R + \langle l_i, u\rangle_{L_2}$$

från $L_2[0,T]\times\mathbb{R}^n$ till $\mathbb{R}^N$. Eftersom $K_i((u;x))$ är kontinuerlig och V_0 utgör grafen av K följer det att V_0 är sluten enligt satsen om den slutna grafen. □

För att hitta $V_0 \cap (V_0^\perp + p)$, börja med att betrakta

$$V_0^\perp = \{(v;w;z) \mid \langle v,u\rangle_{L_2} + \langle w,x\rangle_R + \langle z,d\rangle_Q = 0, \forall (u;x;d)\in V_0\}.$$

Med hjälp av omskrivningen

$$\begin{aligned}\langle z,d\rangle_Q &= \sum_{i=1}^N \langle z,e_i\rangle_Q d_i = \sum_{i=1}^N \langle z,e_i\rangle_Q(\langle\beta_i,x\rangle_R + \langle l_i,u\rangle_{L_2})\\ &= \langle\sum_{i=1}^N \langle z,e_i\rangle_Q\beta_i, x\rangle_R + \langle\sum_{i=1}^N\langle z,e_i\rangle_Q l_i, u\rangle_{L_2}\end{aligned}$$

fås att

$$\begin{aligned}0 &= \langle v,u\rangle_{L_2} + \langle w,x\rangle_R + \langle z,d\rangle_Q\\ &= \langle v,u\rangle_{L_2} + \langle w,x\rangle_R + \langle\sum_{i=1}^N\langle z,e_i\rangle_Q\beta_i,x\rangle_R + \langle\sum_{i=1}^N\langle z,e_i\rangle_Q l_i,u\rangle_{L_2}\\ &= \langle w + \sum_{i=1}^N\langle z,e_i\rangle_Q\beta_i, x\rangle_R + \langle v + \sum_{i=1}^N\langle z,e_i\rangle_Q l_i, u\rangle_{L_2}.\end{aligned}$$

För att ovanstående likhet ska gälla för alla $u\in L_2[0,T]$ och för alla $x\in\mathbb{R}^n$ måste både $w+\sum_{i=1}^N\langle z,e_i\rangle_Q\beta_i$ och $v+\sum_{i=1}^N\langle z,e_i\rangle_Q l_i$ vara lika med noll. Varav $V_0^\perp$ kan skrivas som

$$V_0^\perp = \{(v;w;z) \mid w + \sum_{i=1}^N\langle z,e_i\rangle_Q\beta_i = 0, v+\sum_{i=1}^N\langle z,e_i\rangle_Q l_i = 0\}. \tag{11}$$

Det återstår bara att hitta interceptet $u^\star = V_0\cap(V_0^\perp + p)$. Börja med att betrakta definitionen av V_0 i ekvation (10). Det gäller att

$$d_i = \langle\beta_i,x\rangle_R + \langle l_i,u\rangle_{L_2} = -\sum_{i=1}^N\langle z,e_i\rangle_Q\langle\beta_j,\beta_i\rangle_R - \sum_{i=1}^N\langle z,e_i\rangle_Q\langle l_j,l_i\rangle_{L_2}.$$

Med d som $\hat{y}$ och z som $\hat{y}-\hat{\alpha}$ fås att

$$\begin{aligned}y_i &= -\sum_{i=1}^N\langle\hat{y}-\hat{\alpha},e_i\rangle_Q\langle\beta_j,\beta_i\rangle_R - \sum_{i=1}^N\langle\hat{y}-\hat{\alpha},e_i\rangle_Q\langle l_j,l_i\rangle_{L_2}\\ &= e_j^\top GQ(\hat{y}-\hat{\alpha}) - e_j^\top FQ(\hat{y}-\hat{\alpha}),\end{aligned}$$

där G och F är gramianerna av $\beta_i^\top$ respektive l_i. I och med att F är inverterbar, då l_i:na är oberoende med varandra för $i \in \{1, \cdots, N\}$, och eftersom både F och Q är positivt definita samt att G är positivt semidefinit fås den optimalt anpassade datan av

$$\hat{y} = -(GQ + FQ)(\hat{y} - \hat{\alpha}) = (I + GQ + FQ)^{-1}(GQ + FQ)\hat{\alpha}.$$

Av ekvation (11) fås att

$$\begin{aligned} u^\star(t) &= -\sum_{i=1}^{N} \langle \hat{y} - \hat{\alpha}, e_i \rangle_Q l_i(t) \\ &= -\sum_{i=1}^{N} \langle (I + GQ + FQ)^{-1}(GQ + FQ)\hat{\alpha} - \hat{\alpha}, e_i \rangle_Q l_i(t) \\ &= \sum_{i=1}^{N} \langle (I - (I + GQ + FQ)^{-1}(GQ + FQ))\hat{\alpha}, e_i \rangle_Q l_i(t). \end{aligned}$$

Det går även att göra ri-funktionerna släta genom att införa villkoret att

$$c^\top b = c^\top Ab = c^\top A^2 b = \cdots = c^\top A^{n-2} b = 0.$$

Då fås att de $k - 2$ första derivatorna till $l_i(s)$ är kontinuerliga och kan skrivas som

$$l_i^{(k)}(s) = \begin{cases} c^\top A^k e^{A(t_i - s)} b, & t_i \geq s, \\ 0 \quad \text{annars.} \end{cases}$$

Deriveras

$$y(t) = c^\top e^{At} x_0 + \int_0^t c^\top e^{A(t-s)} b u(s) ds,$$

fås att

$$\dot{y}(t) = c^\top A e^{At} x_0 + c^\top b u(t) + \int_0^t c^\top A e^{A(t-s)} b u(s) ds.$$

Nästa derivata blir

$$\ddot{y}(t) = c^\top A^2 e^{At} x_0 + c^\top b \dot{u}(t) + c^\top A b u(t) + \int_0^t c^\top A^2 e^{A(t-s)} b u(s) ds$$

och den k:te derivatan för $1 \leq k \leq n - 1$ ges av

$$y^{(k)}(t) = c^\top A^k e^{At} x_0 + \sum_{n=1}^{k} c^\top A^{n-1} b u^{(k-n)} + \int_0^t c^\top A^k e^{A(t-s)} b u(s) ds.$$

Nu, med $u = l_i(t)$ och villkoret att $c^\top b = c^\top Ab = c^\top A^2 b = \cdots = c^\top A^{n-2} b = 0$, fås att

$$y^{(n-1)}(t) = c^\top A^k e^{At} x_0 + \int_0^t c^\top A^{n-1} e^{A(t-s)} b l_i(s) ds.$$

vilket är den sista derivatan av y som är kontinuerlig.

3.2.1 Släta ri-funktioner utan initialvillkor

Betrakta problemet från början av avsnitt 3.2 men med $x(0) = 0$ och låt kostfunktionen vara

$$J(u) = \sum_{i=1}^{N} w_i (L_{t_i}(u) - \alpha_i)^2 + \rho \int_0^T u(t)^2 dt,$$

där w_i och ρ är positiva konstanter för alla $i \in \{1, \cdots, N\}$. Om variabelbytet $\mu(t)^2 = \rho u(t)^2$ görs fås problemet att minimiera

$$J(u) = \sum_{i=1}^{N} (\psi_i (L_{t_i}(\mu) - \beta_i)^2 + \int_0^T \mu(t)^2 dt,$$

där $\psi_i = \frac{w_i}{\sqrt{\rho}}$ och $\beta_i = \sqrt{\rho}\alpha_i$. Skrivs summan om på matrisform fås att

$$J(u) = (\hat{y} - \hat{\beta})^\top Q(\hat{y} - \hat{\beta}) + \int_0^T \mu(t)^2 dt$$

där $\hat{\beta} = (\beta_1, \cdots, \beta_N)$, Q är diagonalmatrisen med elementen ψ_i på rad i kolonn i där $i \in \{1, \cdots, N\}$ och eftersom att $x(0) = 0$ är $\hat{y}_i = L_{t_i}(\mu)$ och $\hat{y} = (\hat{y}_1, \cdots, \hat{y}_N)$. Den optimala styrfunktionen fås på liknande sätt som i förra delkapitlet och ges av

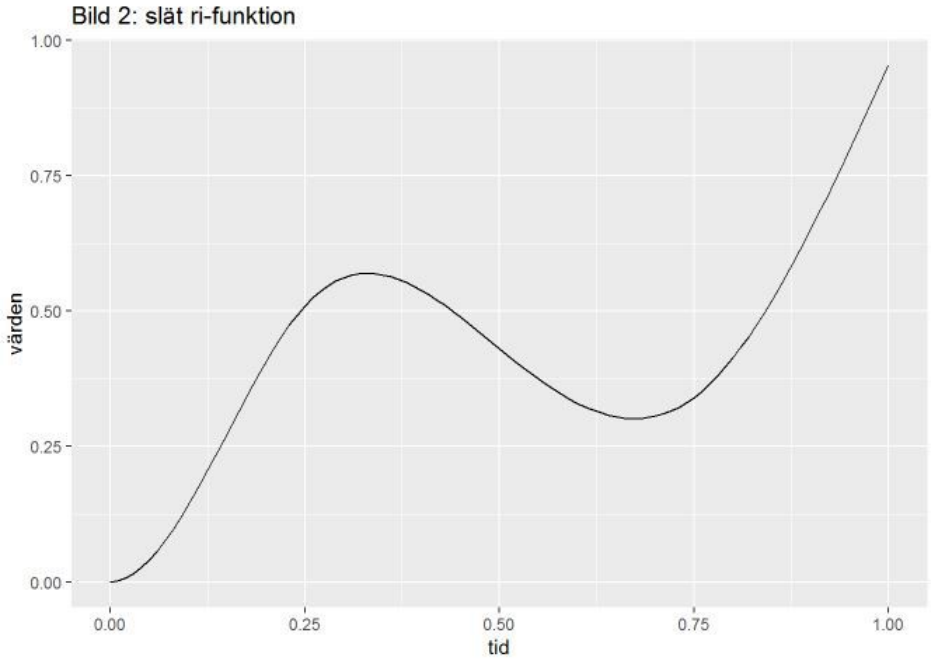
$$\mu^\star(t) = \sum_{i=1}^{N} \langle (I - (I + FQ)^{-1}(FQ))\hat{\alpha}, e_i \rangle_Q l_i(t),$$

där F är gramianen för l_i. Om variablerna byts tillbaka gäller det att

$$y(t) = L_t(u) = \sqrt{\rho} \int_0^t c^\top e^{A(t-s)} u(s).$$

Med samma data som exemplet i avsnitt 3.1.2 och med $\rho = 2 \cdot 10^6$ samt $w_i = 1$ för $i \in \{1, \cdots, N\}$, fås den anpassade ri-funktionen av koden i Bilaga A. Den släta ri-funktionen visas i Bild 2 nedan och liknar resultatet

på sida 6 i [1].



3.2.2 Val av släthetsparameter genom korsvalidering

Detta delkapitel bygger på teori av Grace Wahba i [14]. Betrakta igen kostfunktionen

$$J(u) = \sum_{i=1}^{N}(w_i(L_{t_i}(u) - \alpha_i)^2 + \rho \int_0^T u(t)^2 dt.$$

Beroende på valet av ρ kan man bestämma hur mycket den anpassade funktionen ska följa trender i datan. Ett sätt att välja släthetsparametern är genom så kallad korsvalidering. Givet data $D = \{(t_i, \alpha_i) \mid i = 1, \cdots, N\}$ och villkoret att $x(0) = 0$, låt $u_\rho^{[k]}$ vara funktionen som minimerar

$$\sum_{i=1,\, i\neq k}^{N} w_i(L_{t_i}(u) - \alpha_i)^2 + \rho \int_0^T u(t)^2 dt.$$

Då ges den ordinära korsvalideringsfunktionen av $K_0(\rho)$ av

$$K_0(\rho) = \frac{1}{n}\sum_{k=1}^{n}(\alpha_i - L_{t_i}(u_\rho^{[k]}))^2,$$

varav den ordinära korsvalideringsskattningen OCV för ρ ges av ρ_{ocv} som minimerar $K_0(\rho)$.

3.3 Ri-funktioner på sfärer

Ett sätt att konstruera släta ri-funktioner på enhetssfären är att först avbilda sfären i $\mathbb{R}^2$ och där konstruera ri-funktionerna. Sedan kan de släta ri-funktionerna avbildas tillbaka på sfären.

Låt

$$S(z_1, z_2, z_3) = (y_1, y_2) = (\frac{z_1}{1-z_3}, \frac{z_2}{1-z_3}) \tag{12}$$

vara projektionen från enhetssfären exklusive punkten $(0, 0, 1)$ till $\mathbb{R}^2$. Inversen till denna projektion ges då av

$$S^{-1}(y_1, y_2) = (z_1, z_2, z_3) = \Big(\frac{2y_1}{y_1^2+y_2^2+1}, \frac{2y_2}{y_1^2+y_2^2+1}, \frac{y_1^2+y_2^2-1}{y_1^2+y_2^2+1}\Big).$$

Deriveras punkterna på sfären med avseende på tiden, fås för första koordinaten att

$$\begin{aligned}\dot{z}_1 &= \frac{d}{dt}(\frac{2y_1}{y_1^2+y_2^2+1}) = \frac{2\dot{y}_1}{y_1^2+y_2^2+1} - \frac{2y_1(2y_1\dot{y}_1+2y_2\dot{y}_2)}{(y_1^2+y_2^2+1)^2} \\ &= (1-z_3)\dot{y}_1 - z_1(z_1\dot{y}_1+z_2\dot{y}_2).\end{aligned}$$

Av symmetriskäl fås att

$$\dot{z}_2 = (1-z_3)\dot{y}_2 - z_2(z_1\dot{y}_1+z_2\dot{y}_2).$$

För sista derivatan gäller att

$$\begin{aligned}\dot{z}_3 &= \frac{d}{dt}(\frac{y_1^2+y_2^2-1}{y_1^2+y_2^2+1}) = \frac{2y_1\dot{y}_1+2y_2\dot{y}_2}{y_1^2+y_2^2+1} - \frac{(y_1^2+y_2^2-1)(2y_1\dot{y}_1+2y_2\dot{y}_2)}{(y_1^2+y_2^2+1)^2} \\ &= \frac{2}{y_1^2+y_2^2+1}(y_1\dot{y}_1+y_2\dot{y}_2 - \frac{(y_1^2+y_2^2-1)(y_1\dot{y}_1+y_2\dot{y}_2)}{(y_1^2+y_2^2+1)}) = \\ &= (1-z_3)\frac{2y_1\dot{y}_1+2y_2\dot{y}_2}{y_1^2+y_2^2+1} = (1-z_3)(z_1\dot{y}_1+z_2\dot{y}_2).\end{aligned}$$

I matrisform är alltså

$$\dot{z} = \begin{pmatrix}\dot{z}_1\\ \dot{z}_2\\ \dot{z}_3\end{pmatrix} = \begin{pmatrix}1-z_3-z_1^2 & -z_1z_2\\ -z_1z_2 & 1-z_3-z_2^2\\ (1-z_3)z_1 & (1-z_3)z_2\end{pmatrix}\begin{pmatrix}\dot{y}_1\\ \dot{y}_2\end{pmatrix} = A(z)\dot{y},$$

där $z = (z_1, z_2, z_3)^\top$, $y = (y_1, y_2)^\top$. Andra derivatan fås av:

$$\ddot{z} = \frac{d}{dt}(A(z))\dot{y} + A(z)\ddot{y} \tag{13}$$

och skrivs $\dot{y}$ genom att derivera (12) fås att

$$\dot{y} = \frac{d}{dt}\begin{pmatrix}\frac{z_1}{1-z_3}\\ \frac{z_2}{1-z_3}\end{pmatrix} = \begin{pmatrix}\frac{\dot{z}_1}{1-z_3} - \frac{z_1\dot{z}_3}{(1-z_3)^2}\\ \frac{\dot{z}_2}{1-z_3} - \frac{z_2\dot{z}_3}{(1-z_3)^2}\end{pmatrix} = \begin{pmatrix}\frac{1}{1-z_3} & 0 & \frac{z_1}{(1-z_3)^2}\\ 0 & \frac{1}{1-z_3} & \frac{z_2}{(1-z_3)^2}\end{pmatrix}\begin{pmatrix}\dot{z}_1\\ \dot{z}_2\\ \dot{z}_3\end{pmatrix},$$

varav $\ddot{z}$ kan skrivas som

$$\ddot{z} = \frac{d}{dt}(A(z))B(z)\dot{z} + A(z)\ddot{y},$$

där

$$B(z) = \begin{pmatrix} \frac{1}{1-z_3} & 0 & \frac{z_1}{(1-z_3)^2} \\ 0 & \frac{1}{1-z_3} & \frac{z_2}{(1-z_3)^2} \end{pmatrix}.$$

Om $\ddot{y}_1 = u_1$ och $\ddot{y}_2 = u_2$ är styckvis kontinuerliga styrfunktioner kan (13) skrivas som:

$$\begin{aligned}
\ddot{z}_1 &= -z_1\frac{\dot{z}_1^2 + \dot{z}_2^2 + \dot{z}_3^2}{1-z_3} - 2\frac{\dot{z}_1\dot{z}_3}{1-z_3} + (1 - z_3 - z_1^2)u_1 - z_1z_2u_2 \\
\ddot{z}_2 &= -z_2\frac{\dot{z}_1^2 + \dot{z}_2^2 + \dot{z}_3^2}{1-z_3} - 2\frac{\dot{z}_2\dot{z}_3}{1-z_3} + (1 - z_3 - z_2^2)u_1 - z_1z_2u_2 \\
\ddot{z}_3 &= \dot{z}_1^2 + \dot{z}_2^2 - \frac{1+z_3}{1-z_3}\dot{z}_3^2 + z_1(1-z_3)u_1 + z_2(1-z_3)u_2.
\end{aligned}$$

Med variabelbytet $(w_1, w_2, w_3, w_4, w_5, w_6) = (z_1, z_2, z_3, \dot{z}_1, \dot{z}_2, \dot{z}_3)$ fås att:

$$\begin{aligned}
\frac{d}{dt}\begin{pmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \\ w_6 \end{pmatrix} &= \begin{pmatrix} w_4 \\ w_5 \\ w_6 \\ \frac{w_1}{1-w_3}(w_4^2 + w_5^2 + w_6^2) - \frac{2}{1-w_3}w_4w_6 \\ \frac{w_2}{1-w_3}(w_4^2 + w_5^2 + w_6^2) - \frac{2}{1-w_3}w_5w_6 \\ w_4^2 + w_5^2 - \frac{1+w_3}{1-w_3}w_5w_6 \end{pmatrix} \\
&\quad + \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 - w_3 - w_1^2 \\ -w_1w_2 \\ w_1(1-w_3) \end{pmatrix} u_1 + \begin{pmatrix} 0 \\ 0 \\ 0 \\ -w_1w_2 \\ 1 - w_3 - w_2^2 \\ w_2(1-w_3) \end{pmatrix} u_2.
\end{aligned}$$

Enligt Egerstedt och Martin [1] är systemet ovan styrbart på enhetssfären exklusive punkten $(0, 0, 1)$.

Låt t_i vara tidpunkten för respektive datapunkt $\xi_i = (z_{1i}, z_{2i}, z_{3i})$ och $i \in \{1, \cdots, m\}$. Ett sätt att konstruera släta ri-funktioner på sfären är genom att hitta styrfunktionerna u_1 och u_2 som löser

$$\inf_{u_1,u_2 \in L_2[0,T]} \{\frac{1}{2}\rho \int_0^T (u_1(t)^2 + u_2(t)^2)dt + \frac{1}{2}\sum_{i=1}^m \gamma(z(t_i), \xi)\},$$

där $\rho > 0$ är en släthetsparameter och

$$\gamma(z(t_i), \xi_i) = \frac{\|S(\xi_i)\|}{\sum_{j=1}^m \|S(\xi_j)\|} \|S(z(t_i)) - S(\xi_i)\|^2$$

är en funktion från $S^2 \times S^2$ till $\mathbb{R}_+ \cup 0$ som beräknar kostnaden för $z(t_i)$:s avvikelse från datapunkterna ξ_i. Eftersom små avstånd från punkten $(0,0,1)$ på enhetssfären vid avbildningen till $\mathbb{R}^2$ blir större, borde större vikter

$$\tau_i = \frac{\|S(\xi_i)\|}{\sum_{j=1}^{m} \|S(\xi_j)\|}$$

ges åt de datapunkterna. På samma sätt ges lägre vikt åt datapunkterna nära $(0,0,-1)$. Till sist fås trevligare egenskaper om vikterna normaliseras sådan att $\sum_{i=1}^{m} \tau_i = 1$. Optimeringsproblemet blir att hitta styrfunktionerna $u_1, u_2 \in L_2[0,T]$ som minimerar

$$\frac{1}{2}\rho \int_0^T u_1(t)^2 + u_2(t)^2 dt + \frac{1}{2}\sum_{i=1}^{m} \tau_i((y_1(t_i) - S(\xi_i)_1)^2 + (y_2(t_i) - S(\xi_i)_2)^2, \quad (14)$$

med avseende på

$$\dot{\bar{x}}_i = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \bar{x}_i + \begin{pmatrix} 0 \\ 1 \end{pmatrix} u_i, \; y_i = \begin{pmatrix} 1 & 0 \end{pmatrix} \bar{x}_i,$$

där $\bar{x}_i = (y_i, \dot{y}_i)$, $i \in \{1,2\}$, och $S(\xi_i)_j$, för $j \in \{1,2\}$, är j:te koordinaten i planet av datapunkten i.

Beteckna kostfunktionen med

$$\begin{aligned} J(u_1,u_2) =& \frac{1}{2}\rho \int_0^T u_1(t)^2 + u_2(t)^2 dt \\ &+ \frac{1}{2}\sum_{i=1}^{m} \tau_i((y_1(t_i) - S(\xi_i)_1)^2 + (y_2(t_i) - S(\xi_i)_2)^2) \end{aligned}$$

och låt

$$J_1(u_1) = \frac{1}{2}\rho \int_0^T u_1(t)^2 dt + \frac{1}{2}\sum_{i=1}^{m} \tau_i((y_1(t_i) - S(\xi_i)_1)^2,$$

samt

$$J_2(u_2) = \frac{1}{2}\rho \int_0^T u_2(t)^2 dt + \frac{1}{2}\sum_{i=1}^{m} \tau_i((y_2(t_i) - S(\xi_i)_2)^2$$

Problemet i ekvation (14) kan då översättas till

$$\min_{u_1,u_2 \in L_2[0,T]} J(u_1,u_2) = \min_{u_1 \in L_2[0,T]} J_1(u_1) + \min_{u_2 \in L_2[0,T]} J_2(u_2),$$

det vill säga att lösa två separata minimeringsproblem vilka stötts på tidigare. Med denna metod återskapas, i programmet R (se bilaga A), ri-funktionerna i bilderna på sida 175 i [1]. En ri-funktion anpassas i $\mathbb{R}^2$ vilket

visas i Bild 3. Samma ri-funktion avbildad på sfären ses i Bild 4.

Bild 3: slät ri-funktion på reella planet

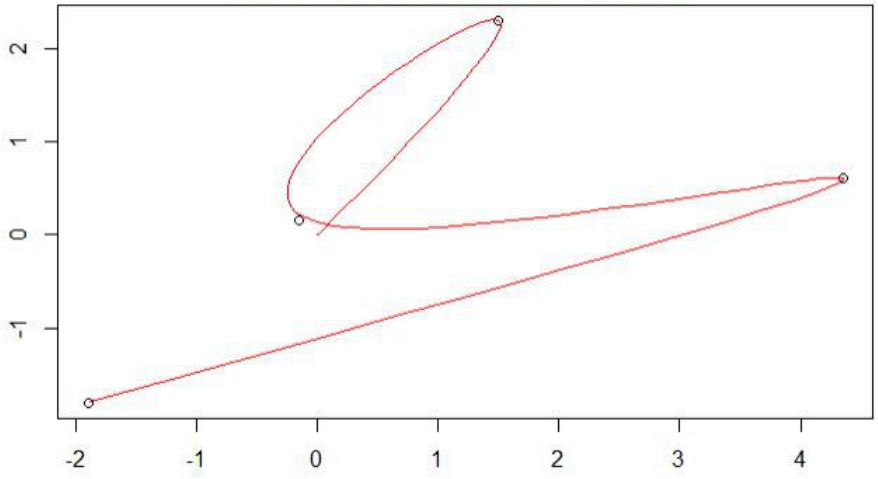
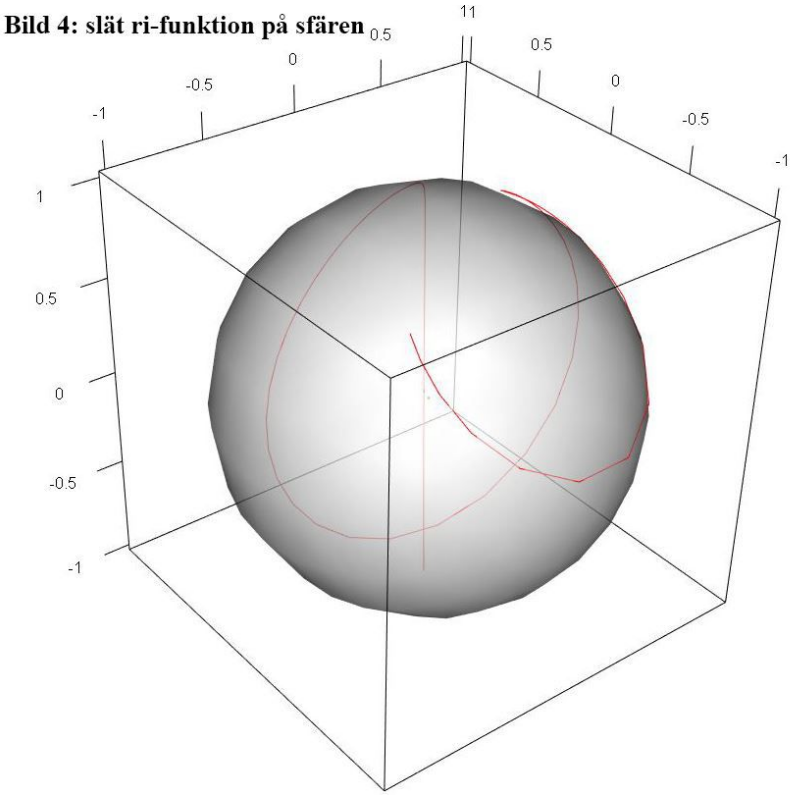


Bild 4: slät ri-funktion på sfären



4 Konvergens av släta ri-funktioner

Detta kapitel bygger på resultat av Egerstedt och Martins i [1], förutom Lemma 4.2 som korrigerats efter att ett fel upptäcktes i personlig kommunikation med Yishao Zhou.

I detta avsnitt approximeras en tillräckligt slät kurva $f(t)$ på intervallet $[0, T]$. Antag att det finns ändligt många datamängder genom återupprepade stickprov från funktionen $f(t)$ på intervallet $[0, T]$. Låt

$$D_N = \{(t_{iN}, \alpha_{iN}) \mid i = 1, \cdots, N\}$$

vara den N:te datamängden och låt unionen av tidsmängderna vara tät i intervallet $[0, T]$. Låt u_N beteckna den optimala styrningen som minimerar kostfunktionen

$$J_N(u) = \frac{1}{2N} \sum_{i=1}^{N} w_{iN}(L_{t_{iN}}(u) - f(t_{iN}))^2 + \frac{\rho}{2} \int_0^T u^2(t)dt,$$

sådan att

$$\begin{cases} \dot{x} = Ax(t) + bu(t), \\ y = c^\top x(t), \end{cases} \qquad x(0) = 0,$$

där w_{iN} är vikterna för den N:te datamängden, ρ är släthetsparametern och kontrollsystemet är styrbart och observerbart. Kom ihåg att lösningen till kontrollsystemet ges av

$$y(t) = L_t(u) = \int_0^t c^\top e^{A(t-s)} bu(s)ds.$$

Att u_N existerar och är unik visades i avsnitt 3.2.1. Innan huvudresultaten i detta kapitel presenteras behövs följande antaganden göras:

1. $c^\top b = c^\top Ab = c^\top A^2 b = \cdots = c^\top A^{n-2} b = 0$,
2. matrisen A har endast reella egenvärden,
3. att den underliggande funktionen $f(t)$ är tillräckligt många gånger deriverbar på intervallet $[0, T]$,
4. gränsvärdet

$$\lim_{N\to\infty} \frac{1}{2N} \sum_{i=1}^{N} w_{iN} h(t_{iN}) = \frac{1}{2} \int_0^T h(t)dt,$$

existerar för alla tal w_{iN}, t_{iN}, $i \in \{1, \cdots N\}$, och alla kontinuerliga funktioner $h(t)$ definierade på $[0, T]$.

Det kan nu presenteras en sats som säger att följden $\{u_N(t)\}_{N=1}^{\infty}$ konvergerar mot $u^\star$ som minimerar

$$J(u) = \frac{1}{2}\int_0^T (L_t(u) - f(t))^2 + \frac{\rho}{2}\int_0^T u^2(t)dt.$$

Sats 4.1. *Under de fyra antagandena ovan konvergerar*

$$\{u_N(t)\}_{N=1}^{\infty}$$

mot funktionen $u^\star$ i L_2-norm och följden av släta ri-funktioner

$$\{L_t(u_N)\}_{N=1}^{\infty}$$

konvergerar mot $L_t(u^\star)$ i L_2-norm.

Satsen kommer bevisas med hjälp av Hilberts projektionssats. Proceduren går som tidigare till genom att konstruera ett affint rum $V_{(\cdot;\cdot)}$ som sedan translateras till vektorrummet $V_{(0;0)}$. Det optimala tillståndet fås sedan av $(V_{(0;0)}^{\perp} + p) \cap V_{(0;0)}$.

Låt $w(t) = L_t(u) - f(t)$. Eftersom $c^\top b = c^\top Ab = c^\top A^2 b = \cdots = c^\top A^{n-2}b = 0$ ges de n första derivatorna av $w(t)$ av

$$\begin{aligned}
w^{(0)}(t) &= \int_0^t c^\top e^{A(t-s)}bu(s)ds - f(t)\\
w^{(1)}(t) &= \int_0^t c^\top Ae^{A(t-s)}bu(s)ds - f^{(1)}(t)\\
&\vdots\\
w^{(n-1)}(t) &= \int_0^t c^\top A^{n-1}e^{A(t-s)}bu(s)ds - f^{(n-1)}(t)\\
w^{(n)}(t) &= c^\top A^{n-1}bu(t) + \int_0^t c^\top A^n e^{A(t-s)}bu(s)ds - f^{(n)}(t).
\end{aligned}$$

Följdsats 2.41 tillsammans med att $y(t) = L_t(u) = w(t) + f(T)$ ger att

$$w^{(n)} - \zeta_{n-1}w^{(n-1)} - \cdots - \zeta_0 w = c^\top A^{n-1}bu(t) - f^{(n)} + \zeta_{n-1}f^{(n-1)} + \cdots + \zeta_0 f,$$

där ζ_i, för $i \in \{1, \cdots n-1\}$ betecknar koefficienterna i det karaktäristiska polynomet av A. Införs $\hat{w} = (w^0, \cdots, w^{n-1})$ gäller att tidsderivatan av $\hat{w}(t)$ kan skrivas som

$$\begin{pmatrix}
w^{(1)}\\
\vdots\\
w^{(n-1)}\\
c^\top A^{n-1}bu + \zeta_{n-1}w^{(n-1)} + \cdots + \zeta_0 - (f^{(n)} - \zeta_{n-1}f^{(n-1)} - \cdots - \zeta_0 f)
\end{pmatrix}$$

och med

$$A = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ \zeta_0 & \zeta_1 & \zeta_2 & \cdots & \zeta_{n-1} \end{pmatrix}$$

fås att

$$\frac{d}{dt}\hat{w}(t) = A\hat{w}(t) + \begin{pmatrix} 0 \\ \vdots \\ 0 \\ c^\top A^{n-1}bu - (f^{(n)} - \zeta_{n-1}f^{(n-1)} - \cdots - \zeta_0 f) \end{pmatrix},$$

vilket i tillståndsform skrivs

$$\frac{d}{dt}\hat{w}(t) = A\hat{w}(t) + (c^\top A^{n-1}b)e_n u(t) + (f^{(n)} - \zeta_{n-1}f^{(n-1)} - \cdots - \zeta_0 f)e_n.$$

Låt nu

$$F(s) = (f^{(n)} - \zeta_{n-1}f^{(n-1)} - \cdots - \zeta_0 f)$$

och konstruera affinrummet

$$V_{(\hat{w}_0;F(s))} = \left\{(\hat{w};u)|\ \hat{w} = \int_0^t e^{\hat{A}(t-s)}(c^\top A^{n-1}b)e_n u(s)ds\right\} + \left(e^{\hat{A}t}\hat{w}_0 + \int_0^T e^{\hat{A}(t-s)}F(s)e_n ds, 0\right)$$

av de par av tillstånd $\hat{w}$ och styrningsfunktioner u som uppfyller differentialekvationen (1). Konstruera vektorrummet

$$V_{(0;0)} = \left\{(\hat{w};u)|\ \hat{w} = \int_0^t e^{\hat{A}(t-s)}(c^\top A^{n-1}b)e_n u(s)ds\right\}$$

och definiera även kostfunktionalen

$$\eta = \int_0^T (L_t(u) - f(t))^2 + u^2(t)dt = \int_0^T (\hat{w}(t)^\top e_1 e_1^\top \hat{w}(t) + u^2(t))dt.$$

Att nu hitta den styrfunktion $u(t)$ som minimerar η längs

$$\frac{d\hat{w}}{dt} = A\hat{w}(t) + c^\top A^{n-1}be_n u \text{ med } \hat{w}(0) = \hat{w}_0 = 0$$

motsvarar en projicering av $\hat{w}$ på $V_{(0;0)}$ - eftersom lösningen till differentialekvationen ges av

$$w = \int_0^T e^{A(t-s)}c^\top A^{n-1}be_n u(s)ds.$$

Detta betyder att det optimala tillståndet samt dess tillhörande optimala styrningsfunktion $(\hat{w}^\star; u^\star)$ är en punkt i $V_{(0;0)}^{\perp}$ enligt Hilberts projektionssats. Den optimala styrfunktionen $u^\star$ beräknas i beviset nedan. Lemma 4.2 och dess bevis bygger på personlig kommunikation med Yishao Zhou.

Lemma 4.2. *Låt*

$$D = \begin{pmatrix} A & -(c^\top A^{n-1}b)^2 e_n e_n^\top) \\ -e_1 e_1^\top & -A^\top \end{pmatrix}.$$

Den optimala kontrollfunktionen $L_t(u^\star)$ ges av

$$L_t(u^\star) = \begin{pmatrix} e_1^\top & 0 \end{pmatrix} e^{Dt} \begin{pmatrix} \hat{w}(0) \\ \lambda(0) \end{pmatrix} - \int_0^t \begin{pmatrix} e_1^\top & 0 \end{pmatrix} e^{H(t-s)} \begin{pmatrix} e_n^\top \\ 0 \end{pmatrix} F(s) ds$$

och den optimala styrningen är

$$u(t) = -c^\top A^{n-1} b e_n^\top \lambda(t),$$

där $\lambda(t)$ är definierat som i Sats 2.42.

Bevis. Betrakta problemet att

$$\begin{aligned} \text{minimera } \eta &= \int_0^T (\hat{w}(t)^\top e_1 e_1^\top \hat{w}(t) + u^2(t)) dt \\ \text{då } \frac{d\hat{w}}{dt} &= A\hat{w}(t) + c^\top A^{n-1} b e_n u \text{ och } \hat{w}(0) = \hat{w}_0 = 0. \end{aligned}$$

Med hjälp av Sats 2.42 i teorin om minsta kvadrater, med $L = e_1 e_1^\top$ och $Q = 0$, ges den optimala styrningen av $u(t) = -c^\top A^{n-1} b e_n^\top \lambda(t)$ där $\lambda(t)$ uppfyller villkoren

$$\dot{\lambda} = -A^\top \lambda - e_1 e_1^\top \hat{w}(t)$$

och $\lambda(T) = 0$. Detta ger att

$$\frac{d}{dt} \begin{pmatrix} \hat{w}(t) \\ \lambda(t) \end{pmatrix} = \begin{pmatrix} A & -(c^\top A^{n-1} b)^2 e_n e_n^\top) \\ -e_1 e_1^\top & -A^\top \end{pmatrix} \begin{pmatrix} \hat{w}(t) \\ \lambda(t) \end{pmatrix},$$

där $\hat{w}(0) = 0$ och $\lambda(T) = 0$. Parallellförflyttas detta system tillbaka sådant att det skär $V_{(\hat{w}_0, F(t))}$ fås differentialekvationerna

$$\frac{d}{dt} \begin{pmatrix} \hat{w}(t) \\ \lambda(t) \end{pmatrix} = \underbrace{\begin{pmatrix} A & -(c^\top A^{n-1} b)^2 e_n e_n^\top) \\ -e_1 e_1^\top & -A^\top \end{pmatrix}}_{:=D} \begin{pmatrix} \hat{w}(t) \\ \lambda(t) \end{pmatrix} - \begin{pmatrix} e_n \\ 0 \end{pmatrix} F(t).$$

Enligt Sats 2.13 är den unika lösningen till detta system

$$\begin{pmatrix} \hat{w} \\ \lambda(t) \end{pmatrix} = e^{Dt} \begin{pmatrix} \hat{w}(0) \\ \lambda(0) \end{pmatrix} - \int_0^t e^{D(t-s)} \begin{pmatrix} e_n \\ 0 \end{pmatrix} F(s) ds.$$

Av detta följer att $L_t(u^\star)$ kan skrivas som

$$\begin{aligned} L_t(u^\star) &= \begin{pmatrix} e_1^\top & 0 \end{pmatrix} \begin{pmatrix} \hat{w}(t) \\ \lambda(t) \end{pmatrix} \\ &= \begin{pmatrix} e_1^\top & 0 \end{pmatrix} e^{Dt} \begin{pmatrix} \hat{w}(0) \\ \lambda(0) \end{pmatrix} - \int_0^t \begin{pmatrix} e_1^\top & 0 \end{pmatrix} e^{D(t-s)} \begin{pmatrix} e_n^\top \\ 0 \end{pmatrix} F(s) ds. \end{aligned}$$

och den optimala styrningen är

$$u^\star(t) = -(c^\top A^{n-1} b) e_n^\top \lambda(t),$$

där

$$\frac{d}{dt} \begin{pmatrix} \hat{w}(t) \\ \lambda(t) \end{pmatrix} = \begin{pmatrix} A & -(c^\top A^{n-1} b e_n)^2 e_n e_n^\top) \\ -e_1 e_1^t & -A^\top \end{pmatrix} \begin{pmatrix} \hat{w}(t) \\ \lambda(t) \end{pmatrix} - \begin{pmatrix} e_n \\ 0 \end{pmatrix} F(t),$$

$\lambda(0) = K(0)x(0)$ och $K(t)$ uppfyller Riccati ekvationen

$$\begin{cases} \dot{K}(t) = -A^\top K(t) - K(t)A + K(t)(c^\top A^{n-1} b)^2 e_n e_n^\top K(t) - e_1 e_1^\top, \\ K(T) = 0. \end{cases}$$

□

Det är nu möjligt att bevisa satsen om konvergensen.

Bevis av Sats 4.1. Från antagandet att f är tillräckligt deriverbar följer att även $u^\star$ är tillräckligt deriverbar. Det finns en unik styrfunktion som minimerar $J(u)$ och styrfunktionen som minimerar $J_N(u)$ är unik. Funktionen som minimera det kvadratiska problemet ges av att Gateaux derivatorna ska vara noll. Beräkning av Gateaux derivatorna ger följande linjära funktioner:

$$DJ(u; w) = \int_0^T (L_t(u) - f(t)) L_t(w) dt + \int_0^T u(t) w(t) dt,$$

$$DJ_N(u; w) = \sum_{i=1}^N w_{iN} (L_{iN}(u) - f(t_{iN})) L_t(w) dt + \int_0^T u(t) w(t) dt.$$

Det är klart att för varje u och w konvergerar $DJ_N(u; w)$ enligt antagande nummer fyra.

Gateaux derivatorna kan skrivas i termer av inre produkter genom att byta ordning på integrationerna

$$DJ(u; w) = \int_0^T \Big(\int_0^T l_t(s)(L_t(u) - f(t)) + u(s) \Big) w(s) dt,$$

$$DJ_N(u;w) = \int_0^T \Big(\sum_{i=1}^{N} w_{iN} l_{t_iN}(s)(L_{iN}(u) - f(t_{iN})) + u(s) \Big) w(s) dt.$$

Från ovanstående två ekvationer fås att konvergensen är oberoende av w, eftersom

$$\sum_{i=1}^{N} l_{t_{iN}}(s) w_{iN}(L_{iN}(u) - f(t_{iN})) + u(s)$$

konvergerar mot

$$\int_0^T l_t(s)(L_t(u) - f(t))dt + u(s)$$

för varje $s \in [0, T]$. Frågan handlar nu snarare om konvergens för linjära operationer snarare än linjära funktioner.

Låt nu

$$B(s)(u) = \int_0^T l_t(s) L_t(u) dt + u(s)$$

och definiera $B_N(s)$ som

$$B_N(s)(u) = \sum_{i=1}^{N} l_{t_{iN}}(s) w_{iN} L_{iN}(u) + u(s).$$

Låt dessutom

$$b(s) = \int_0^T l_t(s) f(t) dt$$

och

$$b_N(s) = \sum_{i=1}^{N} l_{t_{iN}} w_{iN} f(t_{iN}).$$

Det är nu klart att $b_N(s)$ konvergerar mot $b(s)$ punktvis. Därför, givet ϵ, gäller för tillräckligt stort N att

$$|B_N(s)(u_N - u^\star)| < \epsilon.$$

Nu när $B_N(s)x = b_N(s)$ har en unik lösning och då $B_N(s)$ därför inte är singulär, konvergerar

$$u_N(s) - u^\star(s)$$

mot 0 punktvis, eftersom båda termerna är släta och definierade på ett kompakt intervall sker konvergensen även i L_2-norm. □

5 Diskussion och slutsats

I denna uppsats har exempel på optimala släta ri-funktioner från Egerstedt och Martins bok *Control Theoretic Splines* [1] kunnat återskapas genom programmering i programmet R. Med hjälp av min handledare Yishao Zhou har ett fel i deras bok gällande konvergensen av släta ri-funktioner upptäckts och korrigerats. Egerstedt och Martin [1] presenterar den optimala styrfunktionen

$$u^\star = e_1\lambda(t)$$

och den optimala ri-funktionen

$$\begin{aligned} L_t(u^\star) = &\begin{pmatrix} e_1^\top & 0 \end{pmatrix} \exp\begin{pmatrix} A & -e_n e_n^\top \\ -e_1 e_1^\top & -A^\top \end{pmatrix} t \begin{pmatrix} \hat{w}(0) \\ \lambda(0) \end{pmatrix} \\ &- \int_0^t \begin{pmatrix} e_1^\top & 0 \end{pmatrix} \exp\begin{pmatrix} A & -e_n e_n^\top \\ -e_1 e_1^\top & -A^\top \end{pmatrix} (t-s) \begin{pmatrix} e_n^\top \\ 0 \end{pmatrix} F(s)ds, \end{aligned}$$

medan denna uppsats kommer fram till den optimala styrfunktionen

$$u^\star = -c^\top A^{n-1} b e_n^\top \lambda(t).$$

Detta resulterar i att den optimala ri-funktionen som konvergeras mot när datamängden växer är

$$\begin{aligned} L_t(u^\star) = &\begin{pmatrix} e_1^\top & 0 \end{pmatrix} \exp\begin{pmatrix} A & -(c^\top A^{n-1} b)^2 e_n e_n^\top \\ -e_1 e_1^\top & -A^\top \end{pmatrix} t \begin{pmatrix} \hat{w}(0) \\ \lambda(0) \end{pmatrix} \\ &- \int_0^t \begin{pmatrix} e_1^\top & 0 \end{pmatrix} \exp\begin{pmatrix} A & -(c^\top A^{n-1} b)^2 e_n e_n^\top \\ -e_1 e_1^\top & -A^\top \end{pmatrix} (t-s) \begin{pmatrix} e_n^\top \\ 0 \end{pmatrix} F(s)ds. \end{aligned}$$

Släta ri-funktioner har i denna uppsats konstruerats i $\mathbb{R}^2$ samt på sfären. Det kan tänkas att dessa funktioner även går att konstruera på andra geometriska kroppar. Till exempel konstruerar Freja Egebrand, Magnus Egerstedt och Clyde Martin ri-funktioner även på torusen i [15].

Referenser

[1] Egerstedt M, Martin C. Control Theoretic Splines: Optimal Control, Statistics, and Path Planning. Princeton: Princeton University Press; 2010.

[2] Luenberger DG. Optimization by Vector Space Methods. New York: Wiley; 1969.

[3] Klein P. Hilbert spaces and the projection theorem [Internet]. [citerad 2020-05-22]. Hämtad från http://paulklein.ca/newsite/teaching/projections.pdf.

[4] Hoffman K. Analysis in Euclidean Space. Englewood Cliffs, N.J.: Prentice-Hall; 1975.

[5] Rowell D. Time-Domain Solution of LTI State Equations [Internet]. MIT; 2002 [citerad 2020-05-22]. Hämtad från http://web.mit.edu/2.14/www/Handouts/StateSpaceResponse.pdf.

[6] Zhou Y. Linear differential equations [Internet]. Stockholm: Stockholms universitet, Matematiska institutionen; 2019. [citerad 2020-05-29]. Hämtad från https://kurser.math.su.se/pluginfile.php/74698/mod_resource/content/6/LinSysHT14HT19.pdf.

[7] Glad T, Ljung L. Reglerteknik: grundläggande teori. 4., [omarb.] uppl. Lund: Studentlitteratur; 2006.

[8] Smith J. MUS420 Introduction to Linear State Space Models [Internet]. Stanford University; 2019-02-05 [citerad 2020-05-29]. Hämtad från https://ccrma.stanford.edu/ jos/StateSpace/StateSpace.pdf.

[9] Sontag ED. Mathematical Control Theory: Deterministic Finite-Dimensional Systems. 2. ed. New York: Springer; 1998.

[10] Dullerud GE, Paganini FG. A Course in Robust Control Theory: A Convex Approach. New York: Springer; 1999.

[11] Zhou Y. Cayley-Hamilton Theorem. Stockholm: Stockholms universitet, Matematiska institutionen; 2019. [citerad 2020-05-29]. Hämtad från https://kurser.math.su.se/mod/page/view.php?id=46380.

[12] Songsiri J. 4. Minimal realization [Internet]. [citerad 2020-05-22]. Hämtad från http://jitkomut.eng.chula.ac.th/ee635/minreal.pdf.

[13] Gamelin TW. Complex Analysis. New York: Springer; 2001.

[14] Wahba G. Spline Models for Observational Data. Philadelphia: Society for Industrial and Applied Mathematics; 1990.

[15] Egebrand F, Egerstedt M, Martic C. Smoothing Splines on the Torus [Internet]. 2010. [citerad 2020-05-29]. Hämtad från https://magnus.ece.gatech.edu/Papers/mtns10-torus.pdf.

A R-kod

```
library(tidyverse)
library(expm)
library(cubature)
library(rgl)

# Interpolation
A <- matrix(c(0,0,1,0),2,2)
b <- matrix(c(0,1),2,1)
c <- matrix(c(1,0), nrow = 1, ncol = 2)
Time <- 1
N <- 4
times <- c(1/4,1/2,3/4,1)
points <- c(3/4,2/5,1/4,1)

lt <- function(t,s,c,A,b){
  value <- 0
  if (t>s){
    value=c %*% expm(A*(t-s)) %*% b
  }
  return(value)
}

l <- function(times,s,c,A,b){
  v <- c()
  for (i in times){
    v <- c(v,lt(i,s,c,A,b))
  }
  return(v)
}

G <- function(times,c,A,b,N){
  sum <- matrix(numeric(N*N),N,N)
  for (i in 1:10000){
    s <- i/10000
    sum <- sum + 0.0001*l(times,s,c,A,b) %*%
      t(l(times,s,c,A,b))
  }
  return(sum)
}

gramian <- G(times,c,A,b,N)
```

```
Ginv <- solve(gramian)

integrand <- function(s,t,times,points,Ginv,c,A,b){
  c %*% expm(A*(t-s)) %*% b * points %*% Ginv %*%
    l(times,s,c,A,b)
}

y <- function(t,times,points,Ginv,c,A,b){
  return(adaptIntegrate(integrand, t=t,times = times,
                        points = points, Ginv = Ginv,
                        c = c, A = A, b = b,
                        lower = 0,
                        upper = t)$integral)
}

tid <- seq(0,Time,0.01)

values <- unlist(lapply(tid,y, times = times,
                        points = points, Ginv = Ginv,
                        c = c, A = A, b = b))

df <- data.frame(y = values,tid)

ggplot(df,aes(x = tid, y = values))+geom_line()

# Smooth splines
rho <- 2*10^(6)
Q <- diag(1/sqrt(rho),4,4)
points <- c(3/4,2/5,1/4,1)*sqrt(rho)
I <- diag(1,4,4)

F1 <- G(times,c,A,b,N)*rho

IplusFQinv <- solve(I + F1 %*% Q)

constant <- t((I-IplusFQinv %*% F1 %*% Q)
              %*% points) %*% Q

integrand <- function(s,t,times,points,constant,
                      Q,c,A,b){
  c %*% expm(A*(t-s)) %*% b *
    (c(constant %*% c(1,0,0,0),constant %*%
         c(0,1,0,0),constant %*% c(0,0,1,0),
       constant %*% c(0,0,0,1)) %*%
```

```
        l(times,s,c,A,b))
}

y <- function(t,times,points,constant,Q,c,A,b){
  return(adaptIntegrate(integrand, t=t,times = times,
                        points = points,
                        constant = constant,
                        Q = Q, c = c, A = A, b = b,
                        lower = 0,
                        upper = t)$integral)
}

values <- unlist(lapply(tid,y, times = times,
                        points = points,
                        constant = constant,
                        Q = Q, c = c, A = A, b = b))

df <- data.frame(y = values,tid)

ggplot(df,aes(x = tid, y = values*sqrt(rho)))+
  labs(y="values")+geom_line()

# Sphere
A <- matrix(c(0,0,1,0),2,2)
b <- matrix(c(0,1),2,1)
c <- matrix(c(1,0), nrow = 1, ncol= 2)
rho <- 2*10^13
Q <- diag(1/sqrt(rho),4,4)
Time <- 1
N <- 4
times <- c(1/4,2/4,3/4,1)
I <- diag(1,4,4)

z <- matrix(numeric(12),4,3)
z[,1] <- c(0.3512881,-0.2870813,0.4289412,-0.4840764)
z[,2] <- c(0.5386417,0.2870813,0.0591643,-0.4585987)
z[,3] <- c(0.7658080,-0.9138756,0.9013928,0.7452229)

points1 <- (z[,1]/(1-z[,3]))*sqrt(rho)
points2 <- (z[,2]/(1-z[,3]))*sqrt(rho)

weight <- sqrt(points1^2+points2^2)/
  sum(sqrt(points1^2+points2^2))
```

```
Q <- diag(weight/sqrt(rho),4,4)

F1 <- G(times,c,A,b,N)*rho

IplusFQinv <- solve(I + F1 %*% Q)

constant1 <- t((I-IplusFQinv %*% F1 %*% Q)
                %*% points1) %*% Q
constant2 <- t((I-IplusFQinv %*% F1 %*% Q)
                %*% points2) %*% Q

values1 <- unlist(lapply(tid,y, times = times,
                         points = points1,
                         constant = constant1,
                         Q = Q, c = c, A = A, b = b))

values2 <- unlist(lapply(tid,y, times = times,
                         points = points2,
                         constant = constant2,
                         Q = Q, c = c, A = A, b = b))

plot(points1/sqrt(rho),points2/sqrt(rho), xlab = "",
     ylab = "", main = "")

lines(values1*sqrt(rho),values2*sqrt(rho),
      col = "red")

y1 <- (values1)*sqrt(rho)
y2 <- (values2)*sqrt(rho)

z2 <- matrix(c(2*y1/(y1^2+y2^2+1),2*y2/(y1^2+y2^2+1),
               (y1^2+y2^2-1)/(y1^2+y2^2+1)),
             length(y1),3)

open3d()
plot3d(0,0,0,col=3,type="p", radius=0.5, xlab = "",
       ylab = "", zlab = "")

plot3d(0,0,0,col="white",alpha=0.5, add=T,type="s",
       radius=1)

plot3d(z2[,1],z2[,2],z2[,3], col = "red", type = "l",
       add=T, radius=0.01)
```