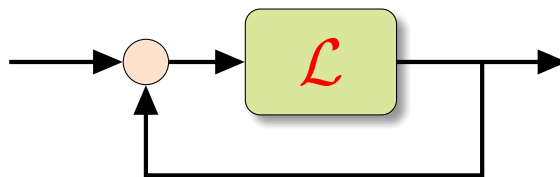


Forelesningsnotater for IELET2101 – Regulerings-teknikk 2
Versjon: v1.5 (16.04.2024)

Dynamiske systemer: Modellering, Simulering og Regulering

Christian Fredrik Sætre

`christian.f.satre@ntnu.no`



Institutt for teknisk kybernetikk, NTNU



Innhold

Forord	vii
I Introduksjon og praktisk informasjon	1
1 Introduksjon	2
1.1 Kort om faget	2
1.1.1 Arbeidskrav	3
1.1.2 Pensum og og bok	3
1.1.3 Annen litteratur og relevante kilder:	4
1.2 Motivasjon for fagets innhold	5
1.2.1 Hva er et dynamisk system?	5
1.2.2 Hvorfor trenger vi tilbakekobling?	6
1.2.3 Praktisk reguleringsteknikk og PID-regulatoren	7
1.2.4 Innstilling/tuning (og etterjustering) av PID-regulatorer	9
II Dynamiske systemer	12
2 Dynamiske systemer	13
2.1 Matematisk notasjon	13
2.2 Størrelser, symboler og variabler i et reguleringsystem	14
2.2.1 Symboler og variabler	14
2.2.2 Aktuatorer og pådragsorganer	14
2.3 Blokk- og instrumenteringsdiagrammer	15
2.3.1 Blokkdiagrammer	15
2.3.2 Instrumenteringsdiagrammer	16
2.4 Dynamiske systemer som differensialligninger	17
2.4.1 Differensialligninger	17
2.4.2 Differensialligninger på tilstandsromform	19
2.4.3 Normal struktur for en tilbakekoblingsløyfe	20
2.4.4 Lineære, tids-invariante (LTI) systemer	21
2.5 Linearisering av ulineære systemer om et arbeidsspunkt	23
2.5.1 Ulineære vs Lineære systemer	23
2.5.2 De grunnleggende ideene bak linearisering	23
2.5.3 Linearisering for multivariable systemer	25
2.6 Overføringsfunksjoner og Laplace-domenet	27
2.6.1 Reguleringsløyfe i Laplace-domenet	27
2.6.2 Laplacetransformasjonen	28
2.6.3 Fra differensialligninger til overføringsfunksjoner	29
2.6.4 Propre og strengt propre overføringfunksjoner	30
2.6.5 Fra overføringsfunksjoner til differensialligninger	31

2.6.6	(Ikke-) Minimum-fase systemer	33
2.6.7	Kausalitet og Realiserbare overføringsfunksjoner	34
2.7	Stabilitet	35
2.7.1	Inngang-utgang-stabilitet	35
2.7.2	Stabilitet i tilstandsrommet*	36
 III Modellering og simulering		38
3	Modellering	39
3.1	Masse- og energi-balanse	40
3.1.1	Massebalanse	40
3.1.2	Energibalanse	44
3.2	Elektro-mekaniske systemer	48
3.2.1	Newtons andre lov (kraftbalanse)	48
3.2.2	Rotasjonsdynamikk (momentbalanse)	51
3.2.3	Gir (ideelle)	56
3.2.4	Krichhoffs lover for elektriske kretser	58
3.2.5	*Euler–Lagrange-ligningene for stive legemer*	60
3.3	*Unlineariteter og fysiske fenomener*	60
3.3.1	Metning og rate-begrensninger	61
3.3.2	Friksjon	61
3.3.3	Hysterese	62
3.3.4	Dødbånd og backlash	62
3.3.5	Grensesvigninger	62
3.4	*Fluidmekanikk*	63
3.4.1	Bernoullis ligning:	63
3.4.2	Systemer med pumper og reguleringsventiler	65
3.4.3	Head loss og den modifiserte Bernoulli-ligningen	66
3.4.4	Reguleringsventiler	67
3.4.5	*Motorer*	69
3.4.6	Pumper	69
3.5	*Aktuatorer og sensorikk*	69
4	Modellforenkling og -tilpasning	70
4.1	Første- og andreordens lineære reguleringsystemer	70
4.1.1	Første-ordens systemer	70
4.1.2	Andre-ordens systemer	71
4.2	Modellforenkling	74
4.2.1	Systemer med tidsforsinkelser	74
4.2.2	Tidsforsinkelser representert i Laplace-domenet	75
4.2.3	Forenkling/approksimasjoner av tidsforsinkelser	75
4.2.4	Første-ordens-pluss-tidsforsinkelse (FOPTF) systemer	76
4.2.5	Andre-ordens-pluss-tidsforsinkelse (AOPTF) systemer	77
4.2.6	Modellforenkling ved sammenslåtte tidskonstanter	78
4.2.7	Modellforenkling ved Skogestads halv-regel	79
4.3	Modellering fra empiriske data	80

4.3.1	Kort om systemidentifikasjon	80
4.3.2	Tilpasning av FOPTF-modell fra sprangrespons	80
4.3.3	Visuell tilpassing av underdempet AOPTF-modell	81
4.3.4	Smiths metode for tilpassning av AOPTF-modeller	82
4.3.5	*Skogestads lukkede-sløyfe-metode*	84
5	Simulering	86
5.1	Numerisk integrasjon og Runge–Kutta-metoder	87
5.1.1	Hva er numerisk integrasjon?	87
5.1.2	Numerisk integrasjon uten tilstander	88
5.1.3	Numerisk integrasjon med tilstander	89
5.1.4	*Høyere ordens Runge–Kutta-metoder*	95
5.1.5	Aspekter ved numerisk integrasjon og ting som bør vurderes	96
5.2	Simulering vha. MATLAB og Simulink	98
5.2.1	MATLABs numeriske ODE-løsere (ODE45 og ODE23)	98
5.2.2	Innstillinger: Relativ- og absolutt toleranse	99
IV	Regulering av mono-variable systemer	101
6	PID-regulatoren	102
6.1	En PID-regulator i et nøtteskall	103
6.1.1	P-leddet	103
6.1.2	I-leddet	103
6.1.3	D-leddet	104
6.1.4	Nominelt pådrag	104
6.1.5	Direkte- vs reversvirkning	105
6.2	Parallell-, serie-form og andre former	105
6.2.1	Parallel- og serie-form	106
6.2.2	PI-D , I-PD og beta-gamma-PID	107
6.3	Derivat-filter	108
6.4	P(I) eller P(I)D?	110
6.4.1	PID for andre-ordens-dominante prosesser	111
6.5	PID-regulatoren i Simulink	111
6.6	Modellfri tuning av PID-regulatorer	111
6.6.1	Etterjustering og manuell tuning	113
6.6.2	Ytelses-karakteristikker og -metrikker	114
6.6.3	Ziegler-Nichols' metoder	114
6.6.4	Auto-tuning og Åstrøms relé-metode	118
7	Modellbasert regulator-design	120
7.1	Direktesyntese	120
7.1.1	Modifisert direktesyntese for ikke-minimum-fase systemer	122
7.1.2	Direktesyntese for systemer med tidsforsinkelser	123
7.2	Intern-modell-kontroll	124
7.3	SIMC-metoden	127
7.3.1	Utleddning av SIMC-reglene	130

7.4	Smith-prediktoren for dødtidkompensering	131
8	Foroverkobling og referansefølging	133
8.1	Foroverkobling og nominelle pådrag	133
8.1.1	Hva er foroverkobling?	133
8.1.2	Tilbakekobling vs foroverkobling	135
8.1.3	Nominelt pådrag	136
8.1.4	Ideelle foroverkoblinger for lineære systemer	138
8.1.5	Utlede realiserbare tilnærmede foroverkoblinger	139
8.1.6	Analytisk foroverkobling fra referansen	142
8.1.7	Foroverkobling basert på lead-lag-element	143
8.1.8	Regulator med to frihetsgrader	144
8.2	Referanse-glatting (myk-starting) og -følging	146
9	Anti-windup og rykkfri overføring	149
9.1	Anti-windup for PID-regulatorer	149
9.1.1	Hva er windup?	149
9.1.2	Anti-windup-metoder	151
9.2	Tracking	154
9.3	Rykkfri overføring for PID-regulatorer	156
10	Alternative reguleringsstrukturer	158
10.1	Kaskade-regulering	158
10.2	Forholdsregulering og synkronisering	161
10.3	Parameterstyring	163
11	Frekvensanalyse	170
11.1	Fase og amplitude	172
11.1.1	Kontur-plott	172
11.1.2	Amplituderatio og fasevinkel	172
11.2	Grafisk analyse: Nyquist-, Bode- og Nichols-diagram	175
11.2.1	Nyquist-diagram	175
11.2.2	Bode-diagram	176
11.2.3	Nichols-diagram	177
11.3	Lukket-sløyfe stabilitet fra frekvensrespons	179
11.3.1	Det kritiske punktet	179
11.3.2	Bodes stabilitetskriterium	180
11.3.3	Nyquists stabilitetskriterium for åpent ustabil system	181
11.3.4	Lukket-sløyfe-stabilitet fra grafiske betraktninger	182
11.4	Stabilitetsmarginer og sensitivitetsanalyse	184
11.4.1	Fase- og forsterknings-marginer	185
11.4.2	Sensivitets-analyse og robusthet mtp. forstyrrelser	187

V	Filtere, signaltilpassing og sampling	191
12	Filtre og digitale reguleringsystemer	192
12.1	Digitale reguleringsystemer	192
12.1.1	Effektiv tidsforsinkelse ved digital regulering	193
12.1.2	Tasting/sampling og Nyquist frekvensen	194
12.1.3	Fenomenet frekvens-folding/aliasing og folding-filtre	194
12.2	Lavpass- og Butterworth-filtre	196
12.2.1	Lavpassfiltre: Ideelle vs Realiserbare	196
12.2.2	Lavpassfiltre som elektriske kretser	197
12.2.3	Butterworth-filtre	198
12.2.4	Alternative lavpass-filtre: Chebyshev, Bessel og elliptiske*	202
12.3	*Høypass-, båndpass og båndstopp-filtre*	202
12.3.1	Komplementærfilter	204
12.3.2	Kalmanfilter*	204
VI	Regulering av multi-variable, koblede systemer	205
13	Multivariable, koblede systemer	206
13.1	Hva er multivariable, koblede systemer?	206
13.1.1	Motiverende eksempel: To koblede traller	207
13.1.2	Hvordan regulere koblede, multivariable systemer?	208
13.1.3	Generell form til et koblet, multivariabelt system	209
13.1.4	Vårt fokus: desentralisert regulering av koblede 2x2 systemer	209
13.2	Dekobling	212
13.3	Stasjonær analyse: RFM og kondisjontall	213
13.3.1	Relativ forsterkningsmatrise (RFM-analyse)	214
13.3.2	Singulærverdi-analyse og Kondisjonstall	220
13.4	Innstilling av PID-regulatorer via frekvensanalyse*	229
Vedlegg		235
A	Begreper og konsepter	235
A.1	Det greske alfabetet	235
A.2	Mekaniske systemer, frihetsgrader og aktueringsgrad	236
A.3	Transienter, statisk respons og likevekt	236
B	Matematiske verktøy	237
B.1	Kalkulus	237
B.1.1	Differensiering	237
B.2	Overføringsfunksjoner, Laplace-transformasjonen og s-omenet	237
B.2.1	Omskriving av strengt propre overføringsfunksjoner	237
B.2.2	Sluttverdi-teoremet	238
B.2.3	Båndbredde	238
B.3	Lineær algebra	238

B.3.1	Matriser og vektorer	238
B.3.2	Eigenverdier og egenvektorer	239
C	Ekstra ting	241
C.1	Tilstandsestimering	241
C.2	Systemidentifikasjon	241
C.3	Tilpassing av F-/A-OTF-modell fra sprangresponser	242

Forord

Dette er forelesningsnotater for emnet IELET2101 ([lenke til emnesiden til IELET2101](#)). Noe av materialet er basert på tidligere faglærerere Fredrik Dessen og Henrik Dobbe Flemmen sine forelesningsnotater og øvinger. Noen figurer og annet materiale relatert til multivariable systemer er også basert på forelesningene til Torleif Anstensrud i det utgåtte emnet TELE3003, som igjen er basert på et kompendium av Pål A. Gisvold. En takk rettet også til flere av studentene som tok emne våren 2023, for flere gode tilbakemeldinger og forslag!

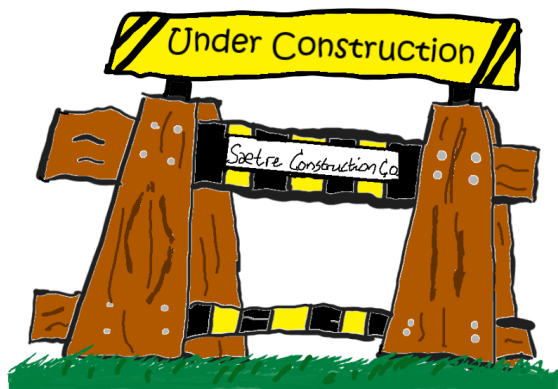
Selv om det er litt kaos nå, så har jeg har som mål at disse notatene skal følge “less is more” prinsippet; altså få frem idéene, teoriene og metodene med minst mulig blablaba. I tillegg skal motivasjonen for *hvorfor du* har lyst til å lære et tema (i motsetning til at det føles ut som at det bare er *jeg* som mener du bør det) komme frem så fort som mulig, og ikke minst bør dette faktisk motivere deg.

Help! Å få til punktene over er ikke så lett uten tilbakemeldinger fra dere, så ikke være redd for å gi meg beskjed (in-person eller via christian.f.satre@ntnu.no) hvis du synes jeg har mislyktes med noen (eller flere) av disse punktene, eller hvis det er noe du synes har behov for mer/bedre forklaringer, har forslag til endringer, etc.

Notasjon og skrivefeil: Notatene er forstøtt under utvikling, og det er flust av skrivefeil og annet rart; gi derfor gjerne beskjed hvis du ser noen feil eller mangler, samt om du har noen spørsmål eller forslag. Det vil jeg sette stor pris på!

Hva er pensum? Seksjoner som er markert med * er **ikke pensum**.

Egne notater: Ting som er markert i rødt er mine egne notater.



Merk: Disse notatene er fortsatt under utvikling (dette er versjon v1.5 (16.04.2024)). Figuren over indikere at noe ikke er ferdigstilt enda, men vil (forhåpentligvis) komme etterhvert.

Del I

Introduksjon og praktisk informasjon

1. Introduksjon

1.1. Kort om faget

Faget bygger videre på temaene du lærte i IELET2002 – Reguleringssteknikk.

Hovedmålet er å fylle på din “reguleringsstekniske verktøykasse”, med et fokus på praktisk og (for det meste) anvendbar reguleringssteknikk (ikke bare [Prosessregulering](#) men også mer generell [servoteknikk](#)).

Du skal mestre det grunnleggende innen følgende temaer relatert til dynamiske systemer:¹

i) **Modellering**; *ii*) **Simulering**; *iii*) **Analyse**; og *iv*) **Regulering**.

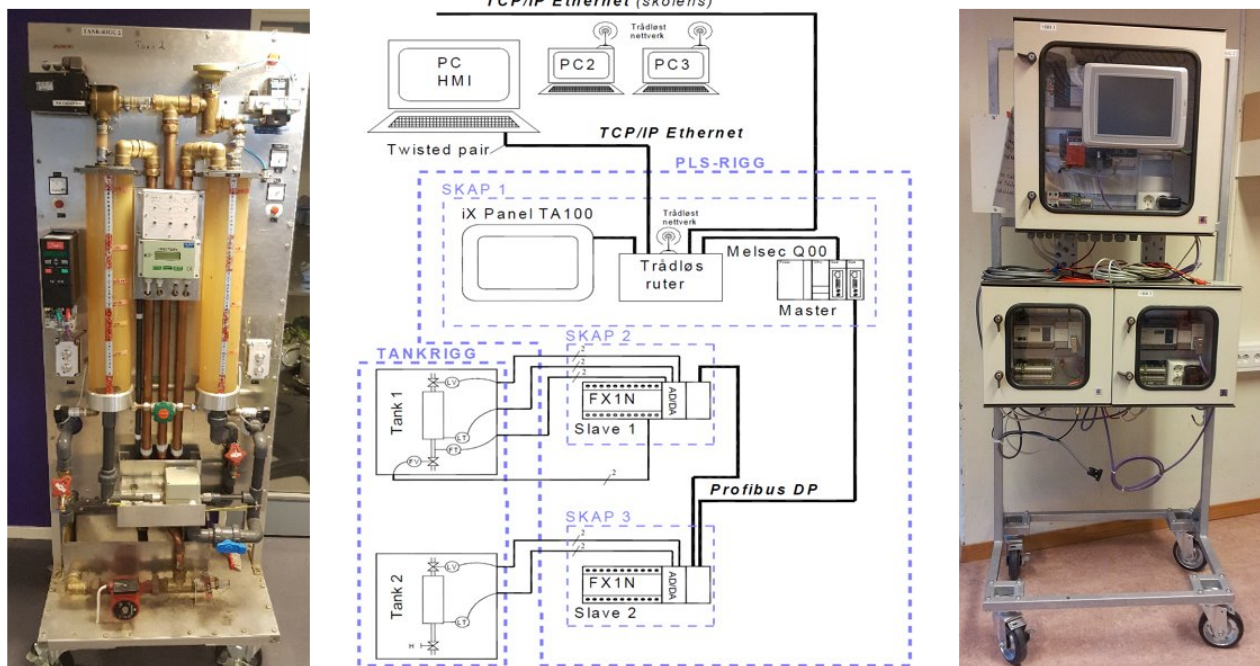
Omfang og relasjon til automatiseringsprosjektet: Selv om de fleste metodene vi skal se på er temmelig generelle, vil vi hovedsakelig fokusere på systemer og praktiske metoder relatert til prosessindustrien. Det er delvis historiske grunner til dette (Norge er fortsatt en oljenasjon), men også for å knytte faget tett opp mot [Automatiseringsprosjektet](#) hvor dere blant annet skal lage regulatorer som regulerer væsknivået i en tank (se figur 1.1).

Begrensninger: Vi skal hovedsakelig holde oss til lineære og kontinuerlige systemer.

Litt om notasjon, begreper og “tankemåte”:

- Minimere bruk av akronymer som ikke er super-duper kjente;
- Prøve å motivere og dra paralleler til andre (mer avanserte) deler av den reguleringsstekniske verden;

¹Andre viktige temaer vi ikke vil se så mye på inkluderer: **Systemidentifikasjon** – Det går ut på å bruke empiriske data (fra f.eks. eksperimenter) fra den fysiske prosessen man ønsker å regulere, til å finne både en passende matematisk modell, samt parametere for denne modellen som passer disse dataene (vi skal se litt på dette); **Adaptiv regulering** – Det går ut på å adaptivt finne eller estimere parametere for regulatoren og/eller prosessen online, noe som kan være nyttig når prosessen kan endre seg, spesielt med tanke på maksimal ytelse via foroverkobling; **Tilstandsestimering** – Estimere umålte tilstander vha. de målte (f.eks. [Luenberg observeren](#) og [Kalman filer](#)); **Optimal- og optimerings-basert regulering** – Regulatorer som er (analytiske eller numeriske) løsninger på et optimaliseringsproblem (feks. [LQR](#) og [MPC](#)).



Figur 1.1: Tanksystemet dere skal jobbe med i automatiseringsprosjektet.

- Vil som regel droppe instrumenteringsdelene av skjemaer (altså ingen rundinger med transmittere og kontrollere, etc.; viktig å ta med i praksis, men dere kan det jo allerede);
- Forsøk på intuitiv bruk av symboler, f.eks. $P(s)$ for overføringsfunksjonen til en prosess, $r(\cdot)$ for referansen, men også noe engelsk som $e(t)$ for feil (les “error”);
- Vi bruker hovedsakelig ikke desibel (dB), men det kan dukke opp;
- Gitt en funksjon av tid, si $y(t)$ ($y(0) = 0$), så vil vi prøve å bruke $Y(s) = \mathcal{L}\{y(t)\}$, altså stor bokstav for dens Laplace transformasjon ($y(s)$ er fy-fy i dette faget!);
- Selv om språkrådet sikkert grøsser av det, så vil jeg bruke “...” som anførselsteng, punktum for desimaltall (f.eks., $0.5 = 1/2$), og terminologi som “anti-windup” i stedet for en norskifiserte variant (anti-opptvinning eller -oppstilling?).

1.1.1 Arbeidskrav

For å gå opp til eksamen må du ha

- 8 av 10 øvinger godkjent

De fleste godkjennes av læringsassistenter ved muntlig utspørring og demonstrasjon i øvingstimer.

1.1.2 Pensum og og bok

Pensum er disse notatene.

⚠ NB! kapitler/seksjoner merket med “ * ” og alle vedleggene er **ikke pensum**.

Merk at notatene vil bli kontinuerlig oppdatert. Dette er versjon v1.5 (16.04.2024).

Et supplement til forelesningsnotatene er boken *Modeling, Simulation and Control* av Finn Aakre Haugen, som man kan laste ned gratis som PDF [her](#).

1.1.3 Annen litteratur og relevante kilder:

Norske bøker om reguleringsteknikk (det er noen valgmuligheter her):

- *Reguleringsteknikk* av Kåre Bjørvik og Per Hveem;
- [Reguleringsteknikk](#) av Jens G. Balchen, Trond Andresen og Bjarne A. Foss;
- [Reguleringsteknikk](#) av Finn Aakre Haugen;
- *Process Dynamics and Control* av Seborg et al. (tidligere pensumbok i emnet).
- *Multivariable Feedback Control* av Skogestad og Postlethwaite tar for seg det meste relatert til regulering av multivariable systemer.

Artikler: Teknisk ukeblad har en samling med artikler kalt [Praktisk reguleringsteknikk](#). F.eks. følgende artikler av Christian Svensson er temmelige relevante: [PID-regulatoren](#) og [Tuning av PID-regulatorer](#).

Videoer og relevante YouTube-kanaler:

- **Trond Andresen** underviste faget TTK4105 – Reguleringsteknikk i en årrekke, og bygde et rykte for å være en god foreleser og pedagog. I tillegg til boken over, har han tilgjengeliggjort opptak av flere relevante forelesninger og annet materiale; se: <https://folk.ntnu.no/tronda/regtek-kurs/>.
- **Brian Douglas** er et velkjent navn blant de fleste studenter som studerer reguleringsteknikk grunnet hans fantastiske YouTube-videoer, både via sin egen kanal og MATLAB sin; se følgende for en samling: <https://engineeringmedia.com/videos>. Han har også en lettleste bok, se: <http://bit.ly/2XLIAKL>.
- **Resourcium.org** har en rekke videoer og opplegg for flere av temaene i disse notatene.
- **Steve Brunton** har videoer om alt fra grunnleggende kontrollteori til mer avansert kontrollteori, se: <https://www.youtube.com/c/Eigensteve>
- **Kristin Y. Pettersen**, som er emneansvarlig for masteremnet TTK4150 - Ulineære systemer, har noen gode videoer relatert til stabilitet, og lineære- vs ulineære systemer, spesifikt L1.2-serien, se: <https://www.youtube.com/c/KYPTeachTech>.

1.2. Motivasjon for fagets innhold

Mye av det som følger lærte dere i IELET2002 – Regulerings-teknikk. Det blir dermed litt repetisjon, men repetisjon er jo bra!  [Jl4_CC3BnpE](#)

Steg i utviklingen av en regulator: Matematisk modell utledet fra førsteprinsipper, fenomenologiske betraktninger eller eksperimenter → systemidentifikasjon → regulator design via simulering, tuning og matematiske metoder → etterjustering.



Figur 1.2: Regulerings-teknikk finnes “over alt”: Eksempel på områder hvor avansert regulerings-teknikk er tatt i bruk inkluderer: Kraftsystemer; Prosesskontroll; Luftfart og romfart; Biler og fartøy; Bygninger og trafikksystemer; Robotikk og hvitevarer; Datasystemer og mobiltelefoner; Reklame og økonomi; Kunst og spill; Fysikk og biologi (figur fra [Samad et al., 2020]).

1.2.1 Hva er et dynamisk system? [GegJZ54KVpE](#)

Alternative kilder: §4.2 i [Haugen, 2023].

Følgende er en litt modifisert forenkling av teksten fra [Wikipedia](#):

Et **dynamisk system** er et system av matematiske ligninger (f.eks. differensialligninger) som beskriver hvordan systemets (indre) tilstander, som er punkter i systemets tilstand-rom, utvikler seg over tid (deres evolusjon).

Eksempler inkluderer de matematiske modellene som beskriver svingningene av en pendel, strømmen av vann i et rør, bevegelsene til en bil, eller hvordan været utvikler seg, etc.

Til enhver tid har et dynamisk system en tilstand som representerer et punkt i et passende tilstandsrom. Evolusjonsregelen for det dynamiske systemet er en funksjon som beskriver hvilke fremtidige tilstander som følger av den nåværende tilstanden.

Helt gresk: De greske bokstavene θ («theta») og τ («tau») dukker opp i det neste eksempelet. Det greske alfabetet er mye brukt på universitet, så det er bare å venne seg til og bli komfortabel med det. Du finner en liste i § A.1; eventuelt kan du se en catchy YouTube-video (3gaeIUsPJ-Y).

Eksempel 1.1. (Pendel)

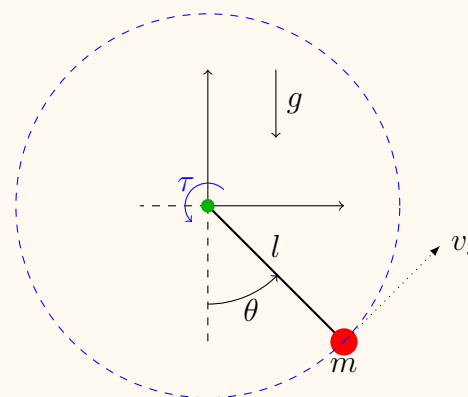
En pendel, som vist i figuren til høyre, er et av de enkleste eksemplene på et ulineært dynamisk system. Ligningene som bestemmer systemets **dy-**
namikk er på formen (disse skal vi utlede i kap. 3)

$$\dot{x} = \frac{d}{dt}x = f(x, u), \quad x = \begin{bmatrix} \theta \\ \dot{\theta} \end{bmatrix}, \quad u = \tau.$$

hvor

x ($=x(t)$) er tilstandvektoren;

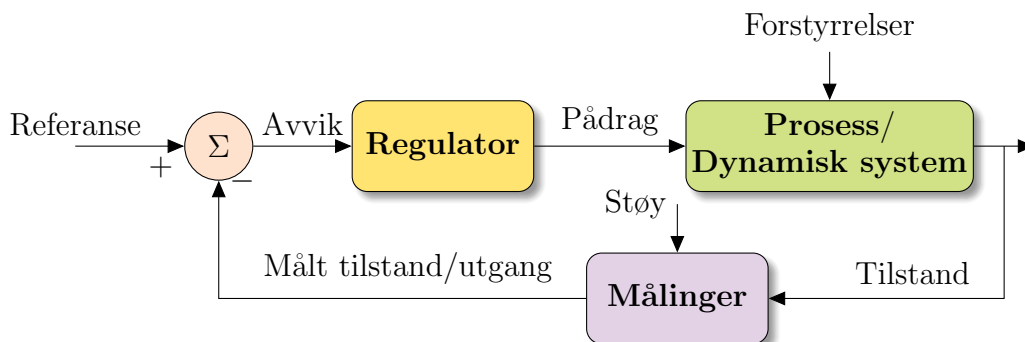
u ($=u(t)$) er pådraget (dreiemomentet om festet).



Det du skal lære i dette faget er essensielt metoder for å bruke pådraget til å “restrukturere” dynamikken til et slikt system, slik at det gjør (eller i hvert fall får visse egenskaper) som vi ønsker! Her spiller *tilbakekobling* en viktig rolle.

1.2.2 Hvorfor trenger vi tilbakekobling?

Vi skiller mellom et system *med* tilbakekobling, såkalt **lukket sløyfe** (eller “auto-modus” i industrien), og *uten* tilbakekobling, såkalt **åpen sløyfe** (manuell-modus).



Figur 1.3: Et enkelt reguleringsystem med tilbakekobling.

Fire viktige mål for en tilbakekoblingsløyfe:

- Få utgangen til å følge den ønskede referansen (kan være et konstant settpunkt);
- Gjøre den lukkede løyfen lite sensitiv til variasjoner i prosessen;
- Redusere effekten av eksterne forstyrrelser;
- Minimere effekten av målestøy.
- (Bonus) Gjøre systemet “bedre” enn det var uten tilbakekobling!

Grunner til å ha en regulator (stikkord-format):

- Stabilitets-design;
- Usikkerhets-kompensering;
- Forstyrrelses-fjerning;
- Støy-demping.

Dessverre kan forbedring av ett av disse punktene igjen føre til dårligere ytelse for ett eller flere av de andre. For eksempel kan man øke stabiliteten til et system ved å øke den kunstige dempingen (noe som ofte tilsvarer D-leddet i en PID-regulator), men dette kan til gjengjeld øke systemets sensitivitet til målestøy.

Vi må derfor ofte nøye oss med et **kompromiss mellom det kontrollerte systemets ytelse** (f.eks. hvor fort det responderer) **og robusthet** (f.eks. hvor sensitivt det er til støy og forstyrrelser).

Merk 1: Tilbakekobling kan gjøre en stabil prosess/system ustabil; det vil også mate målestøy inn i prosessen!

Merk 2: Man kan ofte øke ytelsen ved å kombinere tilbakekobling med, f.eks., fremoverkobling eller alternative reguleringsstrukturer; se kapittel 8.

Merk 3: reguleringsteknikk handler ikke bare om regulatorer og innstilling av disse, men også tilstands-estimering, systemidentifikasjon, maskin læring, etc.; se figur 1.4.

Merk 4: reguleringsteknikk kan variere veldig fra felt til felt; f.eks. har prosessregulering som regel ganske annerledes problemer og ytelseskriterier enn bevegelseskontroll.

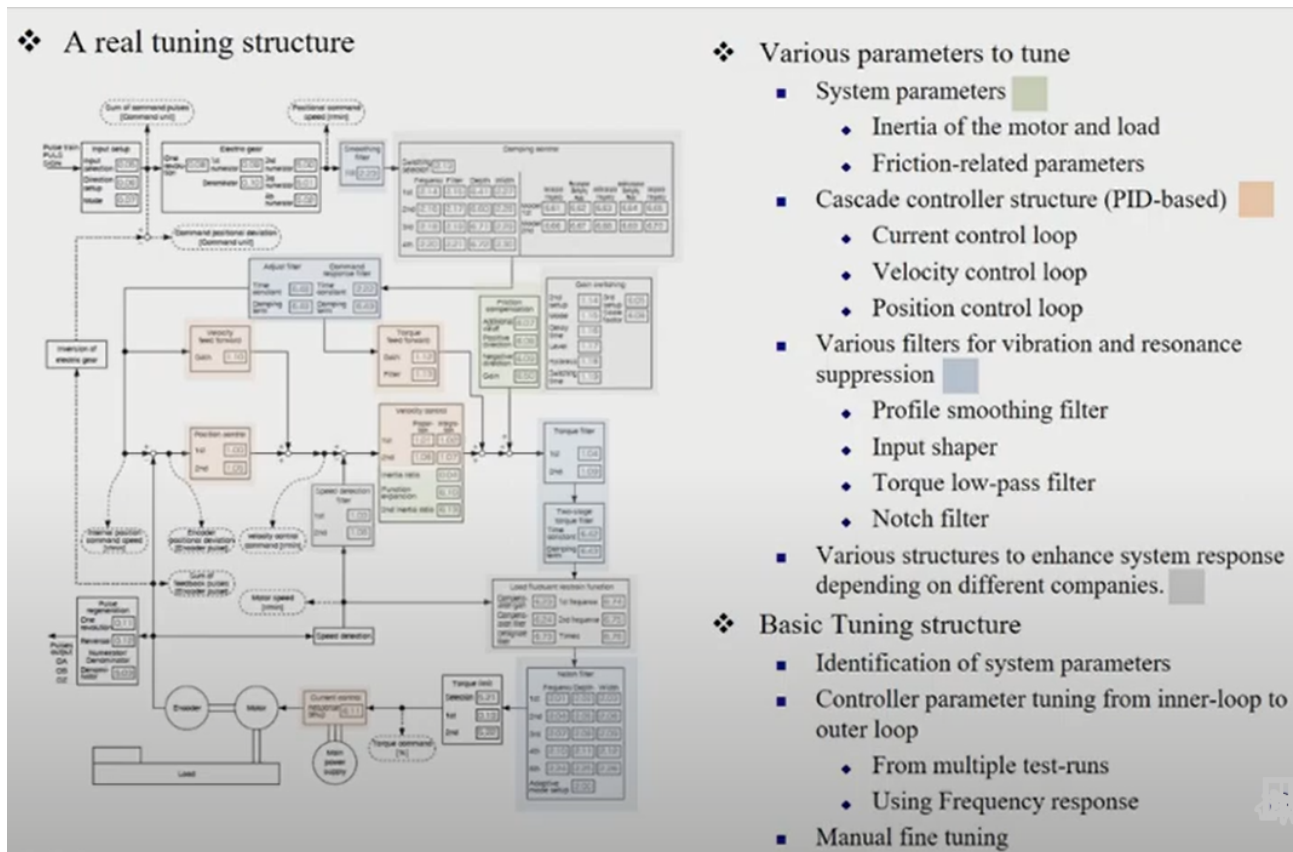
Merk 5: Det er ofte mye fokus på settpunktregulering, men ofte er forstyrrelsesresponsen viktigere!

1.2.3 Praktisk reguleringsteknikk og PID-regulatoren

Når ekspert-paneler skal rangere de viktige reguleringstekniske metodene (eller de metodene som har hatt, og vil fortsette å ha, størst betydning for industrien), så er PID-regulatoren alltid(?) ene og alene øverst på tronen (se f.eks. [Samad, 2017, Samad et al., 2020]).

Hierarki

PID-regulatoren og dens undervarianter (P, PI, og PD, se også 6 for andre varianter) spiller derfor en viktig rolle i dette faget. Mye av fagets innholds relateres til tuning/innstilling og analyse av slike regulatorer, samt andre “tilleggsmetoder” for å forbedre ytelse, robusthet, eller for å løse praktiske problemer. Noen slike metoder og temaer er listet under:



- ❖ Various parameters to tune
 - System parameters
 - ◆ Inertia of the motor and load
 - ◆ Friction-related parameters
 - Cascade controller structure (PID-based)
 - ◆ Current control loop
 - ◆ Velocity control loop
 - ◆ Position control loop
 - Various filters for vibration and resonance suppression
 - ◆ Profile smoothing filter
 - ◆ Input shaper
 - ◆ Torque low-pass filter
 - ◆ Notch filter
 - Various structures to enhance system response depending on different companies.
- ❖ Basic Tuning structure
 - Identification of system parameters
 - Controller parameter tuning from inner-loop to outer loop
 - ◆ From multiple test-runs
 - ◆ Using Frequency response
 - Manual fine tuning

Figur 1.5: Ekte regulatorer, og ikke minst justering («tuning») av disse, kan være kompliserte greier. Her er et eksempel på innstillingsstrukturen til et servokontroll-system. Figur av Dong-il Cho, <https://youtu.be/MkxqxBDjrC8>.

1.2.4 Innstilling/tuning (og etterjustering) av PID-regulatorer

Mål: Gitt en PID-regulator på formen (man kan også ha et første-ordens Derivat-filter)

$$u_{PID} = u_{nom} + k_P \left[e(t) + \frac{1}{T_I} \int_0^t e(\tau) d\tau + T_D \frac{d}{dt} e(t) \right], \tag{PID}$$

så skal vi se på forskjellige metoder for å finne kontrollparameterne k_P , T_I og T_D .

Vi skal i disse notatene blant annet se på enkle empiriske metoder hvor «tuningen» av regulatoren er basert på data generert fra det ekte systemet, automatiske (auto-tuning) metoder, samt strategier for hvordan man kan (manuelt) etterjustere/tune en regulator. Vi skal også se på mer matematisk-orienterte metoder som baserer seg på en matematisk modell av prosessen.

Hva er tuning?

Regulatorinnstilling (tuning) er å finne regulatorparametere (f.eks. k_P , T_I og/eller T_D) som gjør at systemet (i lukket sløyfe) har ønskede egenskaper (se listen under).

Behovet for tuning: En dårlig innstilt regulator (som det finnes mange av i industrien [Skogestad, 2003]) kan lede til trege responser, lav robusthet og unøyaktig regulering av ønsket settpunkt; i verste fall kan det til og med lede til en ustabil prosess.

Figur 1.6: Det finnes mange måter å stille inn PID-regulatorer, noe som innholdslistene til *Handbook of PI and PID controller tuning rules* [O’dwyer, 2009] tydelige illustrerer.

Hva er en godt innstilt regulator? Dette avhenger naturlig nok av forskjellige aspekter som problemet, prosessen og bruksområdet. Den følgende listen (tatt fra [Seborg et al., 2016]) gir dog noen vanlige kriterier og krav som ofte er nødvendig og/eller ønsket:

1. Den lukkede (prosess-) sløyfen er stabil;
2. Effekten av forstyrrelser (og støy) bli minimalisert;
3. Rask, glatt og responsiv regulering til ønsket settpunkt;
4. Eliminering av statiske avvik;
5. Minimerer unødvendige (store) pådrag;
6. Reguleringssystemet er robust mtp. usikkerhet og endringer i prosessen (modellen).

Hvorfor trenger man tuning-prosedyrer? Selv om en PID-regulator kun har tre parametere man kan endre (fire hvis vi tar med et derivat-filter), er det ikke ikke alltid lett å finne gode verdier (innstillinger) for dem ved bare prøving og feiling. En systematisk prosedyre, spesielt en med så få frihetsgrader (parametere) som mulig, er derfor av stor verdi (gitt at den fungerer da!).

Hva kjennetegner en bra tuning-metode? Fra [Skogestad, 2003] har vi at innstillingsreglene bør:

1. være godt motiverte, og gjerne modellbaserte og analytisk avledet;
2. være enkle og lette å huske;
3. fungere godt på et bredt spekter av prosesser.

Selv om punkt 1 over motiverer for bruken av modellbaserte metoder, så skal vi i disse notatene se på både modellfrie- og modellbaserte metoder, f.eks. SIMC for sistnevnte (se §7.3).

Skal man tune mtp. settpunkt-følging eller for å best motvirke forstyrrelser? Hvis settpunktet/referansen for det meste er konstant kan sistenevnt være den mest hensiktsmessige strategien.

Del II
Dynamiske systemer

2. Dynamiske systemer

I dette kapitlet skal vi se på aspekter ved dynamiske (regulerings-)systemer. Av spesiell interesse er representasjonsformene differensialligninger og overføringsfunksjoner, samt hvordan vi kan gå fra den ene representasjonsformen til den andre. I den forbindelse, skal vi se på differensialligninger tilstandsrom form, fra dette såkalte tilstandsrom-metoder (ofte kalt «moderne» reguleringsteknikk), samt lineære- vs ulineære systemer og linearisering. Vi skal så gå over til Laplace-omenet og overføringsfunksjoner (såkalt «klassisk» reguleringsteknikk). Andre ting som dukker så vidt opp inkluderer stabilitet, blokk- og instrumenteringsdiagram.

2.1. Matematisk notasjon

Vi starter med å repetere litt grunnleggende matematisk notasjon:

\mathbb{R} er de **reelle tallene**/tallinjen;

\mathbb{C} er de **komplekse tallene**/det komplekse plan;

\mathbb{R}^n er det n -dimensjonale **Euklidiske vektorrommet**;

\in betyr “i”, f.eks betyr $\mathbf{x} \in \mathbb{R}^n$ at \mathbf{x} tar verdier i \mathbb{R}^n ;

\dot{x} betyr tidsderivatet av $x = x(t)$, altså $\dot{x} = \frac{d}{dt}x$;¹

$\mathbb{R}^{n \times m}$ er settet av alle reelle $n \times m$ matriser;

$:=$ definert som, f.eks. $b := 2a$ betyr at b er definert som to a ;

$\mathbf{x} \in \mathbb{R}^n$ og $\mathbf{A} \in \mathbb{R}^{n \times m}$, altså bruker vi fete bokstaver for vektorer og matriser;

$j := \sqrt{-1}$ er den imaginære enheten;

$Y(s)$ er Laplace-transformasjonen til $y(t)$: $Y(s) = \mathcal{L}\{y\}(s)$.

¹Dette indikerer at $x(\cdot)$ er en **funksjon** av den uavhengig variabelen t (tid). Vi burde derfor egentlig skrive $x(t)$ (som betyr verdien av funksjonen $x(\cdot)$ ved tiden t). Vi vil dog ofte misbruke notasjonen og skrive bare x .

2.2. Størrelser, symboler og variabler i et reguleringsystem

Et dynamisk reguleringsystem «består» av tre ting: dets (interne) tilstander, ytre påvirkninger, og et sett med regler som forklarer hvordan systemets tilstander (dynamisk) vil endre seg med tiden:

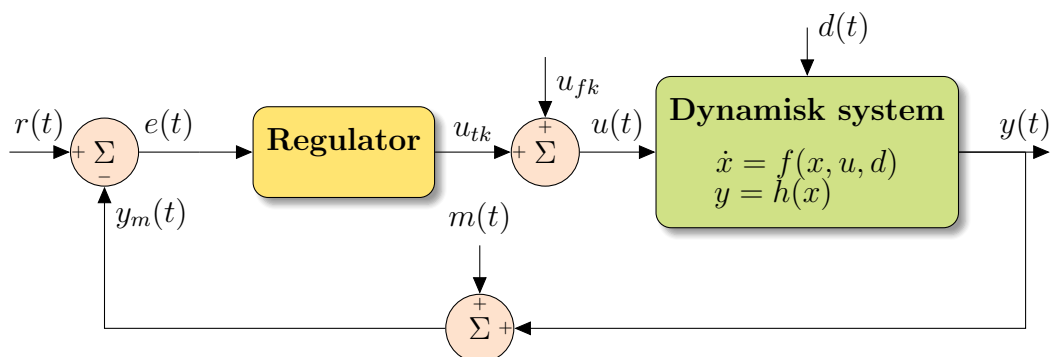
Tilstandene til systemet er de variablene man ved et hvert tidspunkt må vite for å 1) kjenne systemets daværende tilstand (duh!), og 2) forklare hvordan disse tilstandene vil utvikle seg med tiden hvis det ikke er noen ytre påvirkninger. Ta for eksempel et tog: Hvis du vil vite hva posisjonen til toget er fra et gitt punkt noen sekunder frem i tid, så er det ikke nok å bare vite nåværende posisjon (tilstand nr. 1), du må også vite togets hastighet (tilstand nr. 2).

Ytre påvirkninger har en effekt på hvordan systemet endrer seg, men er ikke en intern tilstand til systemet. Slike påvirkninger er typisk pådragsorgan, som moteren til et tog eller en væskestrøm inn i en tank. Det kan også være forstyrrelser, som for eksempel en ukjent innstrøm til en tank eller vind som påvirker toget.

Reglene som bestemmer hvordan tilstandene utvikler seg med tiden kan for eksempel være differensialligninger, differensligninger eller overføringsfunksjoner.

2.2.1 Symboler og variabler

I figur 2.1 har vanlige symboler for variablene bruk i figur 1.3 blitt satt inn. En beskrivelse av disse finner du i tabell 2.1. Disse er alle funksjoner av tiden t .




Figur 2.1: Symbolene som inngår i et enkelt reguleringsystem med tilbakekobling.

2.2.2 Aktuatorer og pådragsorganer

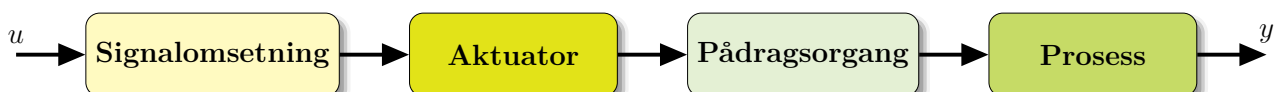
Pådragsignalet u , som blir beregnet av regulatoren i et system, inneholder informasjon som skal tilbakeføres til eller påvirke prosessen vi ønsker å kontrollere. Dette signalet må oftest omformes fra et lav-energi signal i en datamaskin til et høy-energi signal som kan påvirke prosessen direkte. For eksempel, i tilfellet med en automatisk styrbar heisekran, kan motoren påvirke lasten med krefter på mange tusen Newton. Det er derfor et behov for å oversette u til reelle, fysiske verdier som har direkte innflytelse på prosessen. Dette er illustrert i figur 2.2.

Først blir u omregnet i blokken Signalomsetning. Dette kan være en digital-til-analog omforming, omregning mellom strøm og spenning, skalering, eller annen form for signalbehandling.

Tabell 2.1: Symboler for vanlige variabler innen reguleringssteknikken.

Bregrep	Symbol	Eksempler
Tilstand	$x(t)$	Hastigheten til en bil, væskehøyden i en tank, varmen i en ovn.
Pådrag	$u(t)$	Dreiemoment fra en motor, væskestrøm gjennom en reguleringsventil.
Måling	$y(t)$	Turtall fra speedometer, væskehøyde fra flotør, temperatur fra termostat.
Målestøy	$m(t)$	
Referanse	$r(t)$	Ønsket hastighet til en bil, væsknivå i en tank eller temperatur i en ovn.
Avvik	$e(t)$	Differansen mellom ønsket verdi, altså referansen $r(t)$, og målt verdi, $y(t)$.
Forstyrrelse	$d(t)$	Vind, varierende innstrøm i en tank, temperatur utenfor en ovn.

En strøm på 4–20mA er et standard signal som ofte brukes til dette formålet, hovedsakelig på grunn av dens robusthet mot endringer i resistans - en grunn til at spenning sjelden brukes for signaloverføring. Deretter blir signalet sendt til en aktuator, som f.eks. en elektromotor. Motoren påvirker pådragsorganet, f.eks. en ventil, som igjen har direkte effekt på prosessen.

Figur 2.2: Signalkjede fra pådragssignal u til påvirkning (via pådragsorganet) på prosessen.

Hver av blokkene i figur 2.2 kan inneholde sin egen dynamikk (dette gjelder for den del også «Målinger»-blokken i figur 2.1). Denne dynamikken vil nødvendigvis påvirke ytelsen til reguleringsystemet som helhet. I dette faget antar vi at denne dynamikken er en del av prosess-blokken.

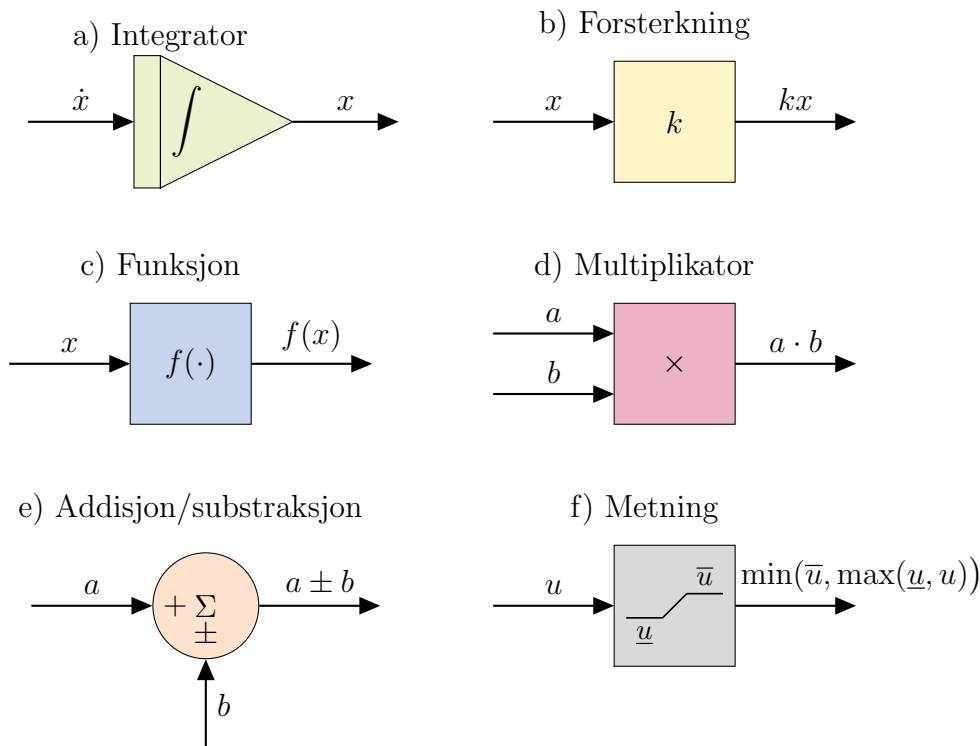
2.3. Blokk- og instrumenteringsdiagrammer

2.3.1 Blokkdiagrammer SdX_r4wFzpU

Alternative kilder: Kap. 5 [Haugen, 2023]; Kap. 2 i [Balchen et al., 2016]

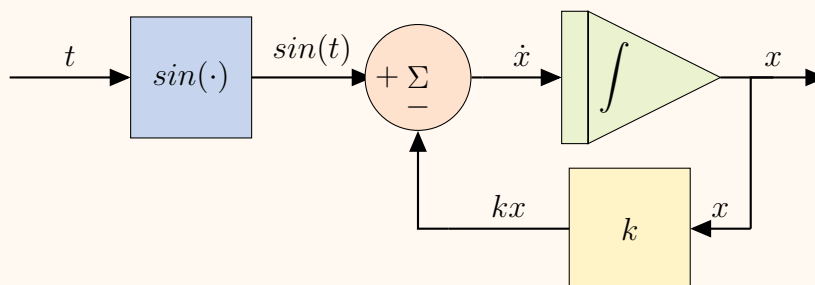
Blokkdiagrammer lar en visualisere flyten i dynamiske systemer ved representere systemet på en grafisk/skjematisk måte. En utvalg av de vanligste blokkene er vist i figur 2.3 (merk at fargene bare er lagt til for å skille blokkene).

Merk: blokkene brukt i **Simulink** er litt annerledes enn de tradisjonelle blokkene brukt her (f.eks. brukes Laplace-operatoren $\frac{1}{s}$ for integrasjon). Hvilken stil dere bruker er ikke så viktig (for min del i hvert fall), så lenge dere er konsekvente med det.



Figur 2.3: Eksempel på vanlige blokker i et blokkdiagram.

Eksempel 2.1. Blokkdiagrammet til systemet $\dot{x} = -kx + \sin(t)$ er som følger:



2.3.2 Instrumenteringsdiagrammer

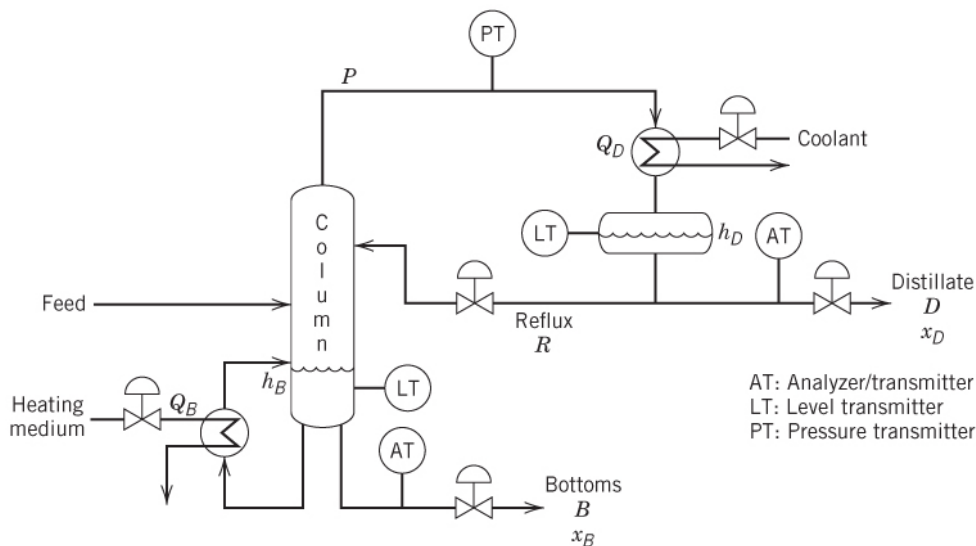
Alternative kilder: Kap. 2 i [Haugen, 2023]; §9.1 i [Seborg et al., 2016].

Et instrumenteringskjema for en destillasjonskolonne er vist i figur 2.4. I et slikt system inngår følgende viktige komponenter:

Sensorer måler en fysisk størrelse (variabel) og overfører dette til et elektriske eller mekanisk signal;

Transmittere konverterer et signal fra en sensor til et signal (normal sett et spennings- eller strøm-signal) som kan brukes av regulatoren;

Transdusere brukes til å omforme et signal fra en fysisk størrelse til en annen (f.eks. fra spenning til trykk eller omvendt).



Figur 2.4: Instrumenteringskjema for en destillasjonskollonne. Figur fra [Seborg et al., 2016].

2.4. Dynamiske systemer som differensialligninger

Alternative kilder: §3.2 i [Balchen et al., 2016]; §6.5 i [Seborg et al., 2016]; Brian Douglas video ; Steve brunton video.

2.4.1 Differensialligninger ▶ MAaWG9G3BHQ

En differensialligning (diff.ligning) er kort og greit en ligning som gir et gitt forhold mellom derivatene til en funksjon. For eksempel, så er

$$\dot{x}(t) = f(x(t)) \quad (2.1)$$

en **tidsinvariant**, **første-ordens** differensialligning på såkalt **tilsdandsromform** (mer om hva alle disse kule ordene betyr om litt). Dette er også en *ordinær* differensialligning, ofte abbreviert ODE pga. det engelske «ordinary differential equation».

Merk: jeg vil ofte, for enkelhets skyld, droppe t -argumentet og bare skrive $\dot{x} = f(x)$.

Hva betyr en slik ligning?

Betydningen av en slik ligning er som følger: Vi leter etter en funksjon (av tid) $x(t)$ som har et tidsderivat, $\dot{x}(t)$, som tilfredsstiller denne ligningen. Tilstanden til det tilsvarende dynamiske systemet ved tiden t er dermed $x(t)$, mens $f(x(t))$ tilsvarer den umiddelbare endringen (“hastigheten”) til tilstanden ved tiden t .

En annen måte å tenke på ligning er som følger: For en hver $x(t)$ kan man finne ut hvordan funksjonen $x(\cdot)$ endrer seg akkurat ved tidspunktet t . Dette kan vi se bedre ved å representere ligningen på en annen vanlig form:

$$\frac{dx}{dt} = f(x(t)).$$

Her kan man intuitivt tenke på *differensialene* (derav *differensial-ligning*), dx og dt , som knøttsmå endringer i henholdsvis x og t fra verdiene $x(t)$ og t . Ved å multiplisere med dt på begge sider og så integrere fra tiden t_0 til t , så finner vi at løsningen på denne differensialligningen er gitt ved

$$x(t) - x(t_0) = \int_{t_0}^t f(x(s)) ds. \quad (\text{Løsningen til en første-ordens differensialligning})$$

Ofta vet man hva **initialverdien** (også kalt startbetingelsene) $x(t_0) = x_0$ er; altså at man vet hvor systemet starter fra. Problemet hvor man ønsker å finne løsningen på differensialligningen fra den gitte initialverdien er da (utrolig nok) kalt et **initialverdiproblem**.

Første-ordens LTI systemer

Lineære og tidsinvariant differensialligninger (ofte abbreviert LTI) er en viktig klasse av systemer innen reguleringsteknikken. Et første-ordens LTI systemer er har følgende form:

$$\dot{x}(t) = -ax(t), \quad (2.2)$$

hvor a er et konstant tall ($a \in \mathbb{R}$). Som tidligere nevnt, så er den **lineær** fordi høyresiden avhenger lineært av $x(t)$, mens den sies å være **tidsinvariant** siden høyresiden bare indirekte avhenger av tiden t via funksjonen $x(t)$.

Eksempler på ulineære differensialligninger er $\dot{x} = ax^3$ eller $\dot{x} = a \sin(x)$.

Eksempler på tidsvarierende differensialligninger er $\dot{x} = (1+t)x$ eller $\dot{x} = -t^2$.

Løsning til en første-ordens differensialligning

Løsning: Anta at $x(0) = x_0$. Løsningen på differensialligningen (2.2) er da $x(t) = x_0 e^{-at}$.

La oss utlede dette. Vi begynner med å skrive (2.2) på følgende form

$$\dot{x} = -ax \quad \implies \quad \frac{1}{x} \dot{x} = -a$$

Merk at symbolet \implies betyr «impliserer». Vi kan da integrere med hensyn på tid på begge sider. Vi starter med venstresiden, altså leddet $\frac{1}{x}\dot{x}$:²

$$\int_0^t \frac{1}{x(s)} \dot{x}(s) ds = \left[\ln(|x(s)|) \right]_0^t = \ln(x(t)) - \ln(x_0) = \ln\left(\frac{x(t)}{x_0}\right).$$

Merk at jeg her i det første steget har brukt (**Kjerneregelen**), hvorfra vi har at $\frac{d}{dt}(\ln(x(t))) = \left(\frac{d}{dx} \ln(x)\right) \cdot \frac{d}{dt}x(t) = \frac{1}{x(t)}\dot{x}(t)$. Merk også at vi «fjerner» absoluttverdiene, altså $|\cdot|$, i den naturlige logaritmen siden $x(t)$ og x_0 må ha samme fortegn (det vil vi se om litt).

Ved å så integrere høyresiden finner vi at

$$\int_0^t -a \cdot ds = [-a \cdot s]_0^t = -at.$$

For å fjerne den naturlige logaritmen i leddet i midten gjør vi følgende:

$$e^{\ln(x(t)) - \ln(x_0)} = e^{\ln\left(\frac{x(t)}{x_0}\right)} = \frac{x(t)}{x_0} = e^{-at} \implies x(t) = x_0 e^{-at},$$

altså vi høynet Eulers konstant e i begge sidene, og brukte at $e^{\ln(y)} = y$.

⚠ Viktig! I stedet for e^{at} og lignende, vil vi ofte i stedet skrive $\exp(at)$, hvor $\exp(\cdot)$ er eksponentialfunksjonen. Dette er for å unngå notasjonsrot, siden e også brukes til å betegne et avvik.

Hvor fort $x(t)$ endrer seg vekk fra x_0 avhenger av tallet $\lambda = -a$, som har følgende navn:

Eigenverdi: $\lambda = -a$ kalles for eigenverdien til (2.2).

2.4.2 Differensialligninger på tilstandsromform

▶ N90KkoXqzFM

Alternative kilder: Kap. 6 i [Haugen, 2023].

Vi skal se på dynamiske systemer på **tilstandsromform**, det vil si at de er representert som et sett av ordinære differensialligninger (forkortet **ODE** fra eng. “ordinary differential equation”):³

$$\begin{aligned} \dot{x}_1 &= f_1(\mathbf{x}, \mathbf{u}, \mathbf{d}) \\ \dot{x}_2 &= f_2(\mathbf{x}, \mathbf{u}, \mathbf{d}) \\ &\vdots \\ \dot{x}_n &= f_n(\mathbf{x}, \mathbf{u}, \mathbf{d}) \\ y_1 &= h_1(\mathbf{x}) \\ y_2 &= h_2(\mathbf{x}) \\ &\vdots \\ y_p &= h_p(\mathbf{x}). \end{aligned}$$

²Legg merke til at jeg bruker en ny, midlertidig integrasjonsvariabel, s , i integranden siden vi integrerer til tiden t .

³I visse applikasjoner kan også $\mathbf{h}(\cdot)$ avhenge av pådraget \mathbf{u} (et typisk eksempel er hvis man kan måle akselerasjon vha. et akselerometer, samt er det nødvendig for visse typer filtre).

Vi vil normalt sett skrive disse på følgende mer kompakte **vektornotasjon-form**:

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \mathbf{u}, \mathbf{d}) \quad (2.3a)$$

$$\mathbf{y} = \mathbf{h}(\mathbf{x}) \quad (2.3b)$$

Her er

$\mathbf{x}(t)$: vektor av systemets n (interne) tilstander, $\mathbf{x} = [x_1, x_2, \dots, x_n]^T$;

$\mathbf{u}(t)$: vektor av systemets m pådrag, $\mathbf{u} = [u_1, u_2, \dots, u_m]^T$;

$\mathbf{y}(t)$: de p utgangene eller prosessvariabelene til systemet, $\mathbf{y} = [y_1, y_2, \dots, y_p]^T$;

$\mathbf{d}(t)$: vektor av systemets l eksterne forstyrrelser (“d” for “disturbance”), $\mathbf{d} = [d_1, d_2, \dots, d_l]^T$.

Både $\mathbf{f}(\cdot)$ og $\mathbf{h}(\cdot)$ er vektor-funksjoner:

$$\mathbf{f}(\mathbf{x}, \mathbf{u}, \mathbf{d}) = \begin{bmatrix} f_1(\mathbf{x}, \mathbf{u}, \mathbf{d}) \\ f_2(\mathbf{x}, \mathbf{u}, \mathbf{d}) \\ \vdots \\ f_n(\mathbf{x}, \mathbf{u}, \mathbf{d}) \end{bmatrix}, \quad \mathbf{h}(\mathbf{x}) = \begin{bmatrix} h_1(\mathbf{x}) \\ h_2(\mathbf{x}) \\ \vdots \\ h_p(\mathbf{x}) \end{bmatrix}. \quad (2.4)$$

Eksempel 2.2. La p betegne posisjonen (i meter) til et tog langs en rett og horisontal skinnegang, og la $v = \frac{dp}{dt} = \dot{p}$ være dets hastighet (“v” for “velocity”). Anta at vi kan styre toget frem og tilbake ved hjelp av en kraft u [N]. Anta videre at den eneste andre kraften som virker på toget er en negativ kraft fra vinden, d , som vi antar er uavhengig av hastigheten v .

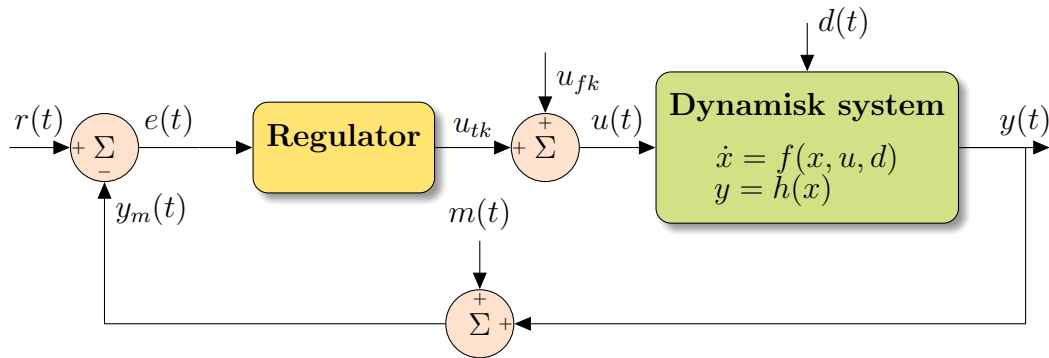
Fra Newtons andre lov vet vi da at $m\dot{v} = u - d$, hvor m er massen til toget (i kilogram). Ved å definere tilstandsvariablene $x_1 = p$ og $x_2 = v$, samt tilstandsvektoren $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$, få vi dermed følgende tilstandsromform:

$$\dot{\mathbf{x}} = \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} x_2 \\ \frac{1}{m}(u - d) \end{bmatrix} =: \begin{bmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}, u, d) \end{bmatrix} =: \mathbf{f}(\mathbf{x}, u, d).$$

2.4.3 Normal struktur for en tilbakekoblingsløyfe

Ved å sette inn symboler som representerer de forskjellige variablene som inngår i en reguleringsløyfe, så vil figur 2.5 tilsvare figur 1.3 for et mono-variabelt systemt (alstå én y og én u). En forklaring av de nye variablene er gitt under:

- $r(t)$ er referansen eller settpunktet, altså den verdien vi ønsker at $y(t)$ skal ha;
- $e(t) = r(t) - y_m(t)$ er avviket eller feilen (les “e” for “error”) — vi ønsker at denne skal være lik null;
- $m(t)$ er målestøy.



Figur 2.5: Reguleringsystem (mono-variabelt) med målestøy og eksterne forstyrrelser.

Merk: Det gir egentlig mer mening å definere avviket/feilen som $e(t) = y_m(t) - r(t)$, noe vi tidvis også vil gjøre. Så sjekk definisjonen!

Hovedmålene i en reguleringsløyfe er å

- Styre utgangen $y(t)$ mot en ønsket referanse $r(t)$;
- Motvirke forstyrrelser $d(t)$ ved å manipulere inngangen $u(t)$;
- Oppnå dette ved å minimisere reguleringsfeilen $e(t) = r(t) - y(t)$.

2.4.4 Lineære, tids-invariante (LTI) systemer ▶ 3eDDTFeSC_Y

Følgende er et lineært, tidsinvariant⁴ (LTI) system på tilstandsromform:

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} + \mathbf{E}\mathbf{d}, \tag{2.5a}$$

$$\mathbf{y} = \mathbf{C}\mathbf{x} + \mathbf{D}\mathbf{u}. \tag{2.5b}$$

Hvis $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{u} \in \mathbb{R}^m$, $\mathbf{d} \in \mathbb{R}^l$ og $\mathbf{y} \in \mathbb{R}^p$, så er dermed \mathbf{A} en $n \times n$ matrise (vi vil tidvis skrive $\mathbf{A} \in \mathbb{R}^{n \times n}$ for å indikere dette), $\mathbf{B} \in \mathbb{R}^{n \times m}$ er en $n \times m$ matrise, \mathbf{C} er en $p \times n$ matrise, $\mathbf{E} \in \mathbb{R}^{n \times l}$, og \mathbf{D} er en $p \times m$ matrise.

En illustrasjon av “oppbygningen” til systemt $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}$, for $\mathbf{x} \in \mathbb{R}^n$ og $\mathbf{u} \in \mathbb{R}^r$, er vist i figur 2.6.

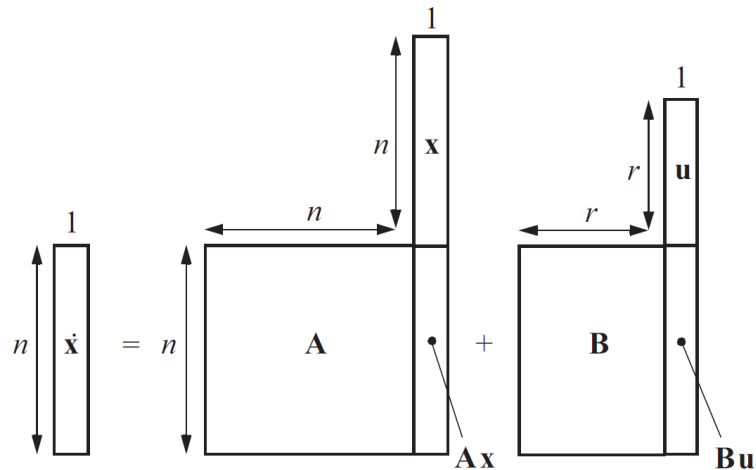
Eksempel 2.3. Gitt en andreordens differensialligning på formen

$$2\ddot{x} + 4\dot{x} + 6x = 2u$$

med utgang $y = x$. La $x_1 = x$ og $x_2 = \dot{x}$ slik at vi kan skrive om diff.ligningen på tilstandsromform:

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= -3x_1 - 2x_2 + u. \end{aligned}$$

⁴Tidsinvariant betyr bare at det ikke avhenger (varierer) med tiden t direkte, bare implisitt via $\mathbf{x}(t)$ etc.



Figur 2.6: “Oppbyggingen” av et LTI system på matrise-form; figur fra [Balchen et al., 2016].

Ved å ta $\mathbf{x} = [x_1 \ x_2]^T$ tilsvarer dette følgende matriseform:

$$\dot{\mathbf{x}} = \underbrace{\begin{bmatrix} 0 & 1 \\ -3 & -2 \end{bmatrix}}_{\mathbf{A}} \mathbf{x} + \underbrace{\begin{bmatrix} 0 \\ 1 \end{bmatrix}}_{\mathbf{B}} u$$

med $y = \mathbf{C}\mathbf{x} = [1 \ 0]\mathbf{x}$.

Egenskaper til Lineære, tids-invariante systemer (LTI)

Lineære, tids-invariant (LTI) systemer har noen viktige egenskaper som bør nevnes:

Superposisjons-prinsippet og homogenitet: En funksjon $f(x)$ som tilfredsstill *superposisjonsprinsippet* kalles for en lineær funksjon. Superposisjon kan defineres av to egenskaper: 1) additivitet: $f(x_1 + x_2) = f(x_1) + f(x_2)$; og 2) homogenitet: $f(ax) = af(x)$ for alle skalarer a .

Har vi for eksempel $f(x) = Ax$, så får vi jo $f(x_1 + x_2) = A(x_1 + x_2) = Ax_1 + Ax_2 = f(x_1) + f(x_2)$. Du kan også vise selv at vi da har $f(ax) = af(x)$ for alle skalarer a .

Willems’ lemma*: For styrebare LTI systemer, så sier Willems’ lemma [Willems et al., 2005, van Waarde et al., 2020] at man kan finne *alle* løsningene til systemet fra én enkelt tilstrekkelig eksiterende løsning (mer spesifikt, inngangen som skaper denne løsningen er tilstrekkelig eksiterende). Selv om vi ikke skal bruke dette direkte i disse notatene, er det nyttig å vite om dette ettersom det har stor verdi for data-drevet modellering og regulering for slike systemer.

2.5. Linearisering av ulineære systemer om et arbeidspunkt

Alternative kilder: §6.5 i [Haugen, 2023]; §4.3 i [Seborg et al., 2016]; §1.6 i [Bjørvik and Hveem, 2014]; §3.6 i [Balchen et al., 2016]; Wikipedia; Brian Douglas video.

2.5.1 Ulineære vs Lineære systemer

De fleste “ekte” systemer er ulineære, så hvorfor er vi så opptatt av lineære systemer? For lineære tidsinvariante (LTI) systemer på formen $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$ kan vi alltid finne en *unik* løsning analytisk: $\mathbf{x}(t) = e^{\mathbf{A}t}\mathbf{x}(0)$. For et unlineært system $\dot{\mathbf{x}} = \mathbf{f}(t, \mathbf{x})$ kan vi derimot generelt sett ikke finne en analytisk løsning (altså en vi kan representere vha. basisfunksjoner).⁵

Om reguleringstekniske grunner for dette.

Siden de fleste “ekte” prosesser er ulineære, kan det være hensiktsmessig å *linearisere* systemet om et ønsket arbeidspunkt. **Hvorfor?** Fordi vi har et stort arsenal av reguleringstekniske metoder (både for design, analyse og justering) for lineære systemer.

2.5.2 De grunnleggende ideene bak linearisering



Gitt et systemt med én tilstand, $x \in \mathbb{R}$, på følgende form:

$$\dot{x} = f(x).$$

Anta at

- $f'(x) = \frac{d}{dx}f(x)$ er en kontinuerlig funksjon;
- a er et **likevektspunkt** til systemet: $f(a) = 0$.

Det **lineariserte systemet** om punktet a er da

$$\frac{d}{dt}\delta x = A \cdot \delta x$$

hvor

- $A = f'(a) = \left. \frac{df}{dx} \right|_{x=a}$ (notasjonen $|_{x=a}$ betyr at vi setter $x = a$ etter vi har derivert funksjonen med hensyn på variabelen x);
- δx er “tilstandanden” til det lineariserte systemet, hvor man kan tenke at $\delta x \approx x - a$.

⁵Rent matematisk sett er generelt sett ikke løsninger til ulineære systemer unike, og hvis funksjonen $\mathbf{f}(t, \mathbf{x})$ ikke er kontinuerlig er det ikke engang sikkert en løsning eksisterer.

Oppgave 2.1. Gitt systemet over, vis at for $\Delta x = x - a$, så er $\Delta x = \delta x$ hvis og bare hvis funksjonen $f(\cdot)$ allerede er lineær (f har følgende form: $f(x) = \alpha x + \beta$ for konstante skalarer α og β).

Achtung! Man ser dessverre ofte at det lineariserte systemet er skrevet som $\frac{d}{dt}\Delta x = A\Delta x$ for $\Delta x = x - a$. Som du viste i oppgaven over er dette dog bare tilfelle for systemer som er lineære i utgangspunktet.

Grunnen til at disse generelt sett er forskjellig, er at det lineariserte systemet (som beskriver dynamikken til δx) kun approksimerer dynamikken til det ulineære systemet (som beskriver dynamikken til x og derav Δx) nært arbeidspunktet. Det lineariserte systemet er med andre ord et approksimasjonssystem som gir oss at $\delta x \approx \Delta x$ nært arbeidspunktet.

Motivasjonen bak linearisering om et likevektspunkt stammer fra [Taylors teorem](#), som sier at for all punkter a så har vi

$$f(x) = f(a) + f'(a)(x - a) + \gamma(x)(x - a), \quad \lim_{x \rightarrow a} \gamma(x) = 0.$$

Med andre ord, siden $\gamma(x)(x - a)$ normalt sett vil være mye mindre enn $f'(a)(x - a)$ når $(x - a)$ er liten (unntaket er hvis $f'(a) = 0$, uten at det ødelegger metoden på noen måte), så er $f'(a)(x - a)$ er grei approksimasjon av $f(x)$ nært et likevektspunkt a .

Følgende eksempel fra [Wikipedia](#) viser på en fin måte litt hvordan linearisering fungerer.⁶

Eksempel 2.4. Gitt funksjonen $f(x) = \sqrt{x}$, så er det jo er det jo lett å se at $f(4) = \sqrt{4} = 2$; men hva er $f(4.001)$? Vi kan anta at nærme $x_* = 4$, så har vi $f(x) \approx f(4) + f'(4)(x - 4) = 2 + \frac{1}{4}(x - 4)$, slik at $f(4.001) \approx 4 + 0.001/4 = 2.00025$, som jo er temmelig nærme den reelle verdien $f(4.001) = 2.00024998$.

Hva med pådrag? La nå systemet i stedet være på formen

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

altså med én tilstand, $x \in \mathbb{R}$, og ett pådrag, $u \in \mathbb{R}$. Anta at

- $\frac{\partial f(x,u)}{\partial x}$ og $\frac{\partial f(x,u)}{\partial u}$ er kontinuerlige funksjoner;
- (x_*, u_*) er et **arbeidspunkt/tvunget likevektspunkt**: $f(x_*, u_*) = 0$.

Det lineariserte systemet om (x_*, u_*) er

$$\frac{d}{dt}\delta x = A \cdot \delta x + B \cdot \delta u$$

hvor

- $A = \left. \frac{\partial f}{\partial x} \right|_a$ og $B = \left. \frac{\partial f}{\partial u} \right|_a$ med notasjonen $\left. \right|_a = \left. \right|_{\substack{x=x_* \\ u=u_*}}$;
- man intuitivt kan tenke at $\delta x \approx x - x_*$ og $\delta u \approx u - u_*$.

⁶Hvis du virkelig vil forstå det matematiske maskineriet som lar oss ta i bruk linearisering for reguleringstekniskeformål, så kan du f.eks. ta en titt på [Hartman–Grobman teoremet](#).

Eksempel 2.5. Systemet $\dot{x} = f(x, u) = \sin(x) + \cos(x)u$ har for $u = 0$ likevektspunkter som alle er multiplum av π , f.eks. $0, -\pi$ og π . La oss linearisere om punktet $(x_*, u_*) = (0, 0)$:

$$A = \left. \frac{\partial f}{\partial x} \right|_{\substack{x=0 \\ u=0}} = \cos(0) - \sin(0) \cdot 0 = 1 \quad \text{og} \quad B = \left. \frac{\partial f}{\partial u} \right|_{\substack{x=0 \\ u=0}} = \cos(0) = 1$$

Dermed er det lineariserte systemet $\frac{d}{dt} \delta x = \delta x + \delta u$.

2.5.3 Linearisering for multivariable systemer



Gitt et ulineært dynamisk system på formen på formen (2.3) uten forstyrrelser ($\mathbf{d} = 0$), altså

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \mathbf{u}), \quad (2.6a)$$

$$\mathbf{y} = \mathbf{h}(\mathbf{x}), \quad (2.6b)$$

hvor både $\mathbf{f}(\cdot)$ og $\mathbf{h}(\cdot)$ er kontinuerlige, differensierbare vektor funksjoner (se (2.4)). Vårt mål er å linearisere dette systemet om et arbeidspunkt $(\mathbf{x}_*, \mathbf{u}_*)$ som tilsvarer et tvunget *likevektspunkt*:

Likevektspunkter og arbeidspunkter:

Definisjon 2.1. Et punkt, \mathbf{x}_* , i tilstandsrommet \mathbb{R}^n er et **tvunget likevektspunkt**, eller **arbeidspunkt**, for systemet (2.6a) hvis det finnes et (nominelt) pådrag $\mathbf{u}_* \in \mathbb{R}^m$ slik at

$$\mathbf{f}(\mathbf{x}_*, \mathbf{u}_*) = 0.$$

Dermed endrer ikke tilstandene seg ved dette punktet ($\dot{\mathbf{x}} \equiv 0$) hvis pådraget er \mathbf{u}_* .

Mer om likevektspunkter: Et (dynamisk) system uten ytre påvirkninger (ingen forstyrrelser eller pådrag) er i **likevekt** hvis dets tilstand ikke endrer seg (kreftene som virker på systemet er i likevekt). Et punkt i tilstandsrommet hvor system er i likevekt kaller et **likevektspunkt** (eng.: equilibrium point).

Eksempel 2.6. $x = 2$ er et likevektspunkt for $\dot{x} = x - 2 + u$ siden $\dot{x} = 0$ når $x = 2$ og $u = 0$. Merk at for $u = 2$ så har vi $\dot{x} = 0$ når $x = 0$. For sistnevnte kaller vi $x = 0$ et **tvunget likevektspunkt** siden systemet bare er i likevekt for et pådrag som ikke er lik null.

Linearisering: Prosedyre for multivariable systemer

Gitt systemet (2.6a) og et arbeidspunkt $(\mathbf{x}_*, \mathbf{u}_*)$ (se definisjon 2.1), så har det lineariserte systemet følgende form:

$$\begin{aligned} \frac{d}{dt} \delta \mathbf{x} &= \mathbf{A} \delta \mathbf{x} + \mathbf{B} \delta \mathbf{u} \\ \delta \mathbf{y} &= \mathbf{C} \delta \mathbf{x}. \end{aligned}$$

Her kan man tenk at

$$\delta \mathbf{x} \approx \mathbf{x} - \mathbf{x}_*, \quad \delta \mathbf{u} \approx \mathbf{u} - \mathbf{u}_* \quad \text{og} \quad \delta \mathbf{y} \approx \mathbf{y} - \mathbf{h}(\mathbf{x}_*),$$

mens $\mathbf{A} := \left. \frac{\partial \mathbf{f}}{\partial \mathbf{x}} \right|_{\substack{\mathbf{x}=\mathbf{x}_* \\ \mathbf{u}=\mathbf{u}_*}}$, $\mathbf{B} := \left. \frac{\partial \mathbf{f}}{\partial \mathbf{u}} \right|_{\substack{\mathbf{x}=\mathbf{x}_* \\ \mathbf{u}=\mathbf{u}_*}}$ og $\mathbf{C} := \left. \frac{\partial \mathbf{h}}{\partial \mathbf{x}} \right|_{\substack{\mathbf{x}=\mathbf{x}_* \\ \mathbf{u}=\mathbf{u}_*}}$, hvor

$$\frac{\partial \mathbf{f}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \cdots & \frac{\partial f_n}{\partial x_n} \end{bmatrix}, \quad \frac{\partial \mathbf{f}}{\partial \mathbf{u}} = \begin{bmatrix} \frac{\partial f_1}{\partial u_1} & \frac{\partial f_1}{\partial u_2} & \cdots & \frac{\partial f_1}{\partial u_m} \\ \frac{\partial f_2}{\partial u_1} & \frac{\partial f_2}{\partial u_2} & \cdots & \frac{\partial f_2}{\partial u_m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial u_1} & \frac{\partial f_n}{\partial u_2} & \cdots & \frac{\partial f_n}{\partial u_n} \end{bmatrix}, \quad \frac{\partial \mathbf{h}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial h_1}{\partial x_1} & \frac{\partial h_1}{\partial x_2} & \cdots & \frac{\partial h_1}{\partial x_n} \\ \frac{\partial h_2}{\partial x_1} & \frac{\partial h_2}{\partial x_2} & \cdots & \frac{\partial h_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial h_p}{\partial x_1} & \frac{\partial h_p}{\partial x_2} & \cdots & \frac{\partial h_p}{\partial x_n} \end{bmatrix}. \tag{2.7}$$

Motivasjon: Tilsvarende det mono-variable systemet, så er motivasjonen bak dette igjen Taylors teorem, som sier at $\mathbf{f}(\mathbf{x}, \mathbf{u}) \approx \mathbf{A}(\mathbf{x} - \mathbf{x}_*) + \mathbf{B}(\mathbf{u} - \mathbf{u}_*)$ når (\mathbf{x}, \mathbf{u}) er nærme nok arbeidspunktet $(\mathbf{x}_*, \mathbf{u}_*)$; hvor nærme avhenger, grovt forklart, av hvor ulineært systemet er. Så selv om linearisering kan være veldig nyttig, må det brukes med forsiktighet; f.eks. kan arbeidsområdet til en regulator designet vha. det lineariserte systemet være veldig liten.

Merk: skal du designe en regulator som skal virke innen et gitt arbeidsområde, og ikke bare for et enkelt arbeidspunkt? Da kan kanskje parameterstyring (eng. “gain scheduling”) være et alternativ du bør se på (vi skal se nærmere på dette i §10.3).

Eksempel 2.7. (Pendel) Gitt $\ddot{\theta} = a \sin(\theta) + d\dot{\theta} + bu$. Ved å ta $x_1 = \theta$ og $x_2 = \dot{\theta}$ har vi

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= a \sin(x_1) + dx_2 + bu. \end{aligned}$$

Et likevektspunkt er dermed et punkt (\mathbf{x}^*, u^*) hvor $x_2^* = 0$ og $a \sin(x_1^*) + bu^* = 0$.

La oss linearisere om punktet hvor pendelen står rett opp, noe som her tilsvarer $x_1^* = \pi$, og dermed $u^* = 0$. Vi får da at

$$\mathbf{A} = \left. \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} \end{bmatrix} \right|_{x=x^*} = \left. \begin{bmatrix} 0 & 1 \\ a \cos(x_1) & d \end{bmatrix} \right|_{x=x^*} = \begin{bmatrix} 0 & 1 \\ -a & d \end{bmatrix} \quad \text{og} \quad \mathbf{B} = \left. \begin{bmatrix} \frac{\partial f_1}{\partial u} \\ \frac{\partial f_2}{\partial u} \end{bmatrix} \right|_{\substack{x=x^* \\ u=u^*}} = \begin{bmatrix} 0 \\ b \end{bmatrix}.$$

Eksempel 2.8. (*Robotligningen*) De dynamiske ligningene til en rekke systemer, deriblant robotmanipulatorer, har følgende form:

$$\mathbf{M}(\mathbf{q})\ddot{\mathbf{q}} + \mathbf{H}(\mathbf{q}, \dot{\mathbf{q}})\dot{\mathbf{q}} + \mathbf{G}(\mathbf{q}) = \mathbf{Q}\mathbf{u}.$$

Her er $\mathbf{q} \in \mathbb{R}^n$ såkalte generaliserte koordinater, $\dot{\mathbf{q}} \in \mathbb{R}^n$ er de tilsvarende generaliserte hastighetene, og $\mathbf{u} \in \mathbb{R}^m$ er pådragene som virker på systemet. Videre er $\mathbf{M}(\cdot)$ en ikke-singulær matrise for alle \mathbf{q} (tilsvarende masser og treghetsmomenter), matrisen $\mathbf{H}(\cdot)$ (som tilsvarer Coriolis- og sentrifugalkrefter) tilfredsstillers $\mathbf{H}(\mathbf{q}, \mathbf{X})\mathbf{Y} = \mathbf{H}(\mathbf{q}, \mathbf{Y})\mathbf{X}$, mens $\mathbf{G}(\cdot)$ tilsvarer gradient til systemets potensielle energi.

Merk at alle likevektspunkter for $\mathbf{u}^* = 0$ må tilfredsstillers $\mathbf{G}(\mathbf{q}^*) = 0$ og $\dot{\mathbf{q}}^* = 0$. La oss

linearisere systemet om et slikt punkt. Vi begynner med å skrive systemet på tilstandsromform:

$$\begin{aligned}\dot{\mathbf{x}}_1 &= \mathbf{x}_2 := \mathbf{f}_1(\mathbf{x}, \mathbf{u}) \\ \dot{\mathbf{x}}_2 &= \mathbf{M}(\mathbf{x}_1)^{-1} [-\mathbf{H}(\mathbf{x}_1, \mathbf{x}_2)\mathbf{x}_2 - \mathbf{G}(\mathbf{x}_1) + \mathbf{Q}\mathbf{u}] =: \mathbf{f}_2(\mathbf{x}, \mathbf{u})\end{aligned}$$

hvor $\mathbf{x}_1 := \mathbf{q}$ og $\mathbf{x}_2 := \dot{\mathbf{q}}$. Vi har dermed^a

$$\mathbf{A} = \begin{bmatrix} 0_{n \times n} & \mathbf{I}_n \\ \frac{\partial \mathbf{f}_2}{\partial \mathbf{x}_1} & \frac{\partial \mathbf{f}_2}{\partial \mathbf{x}_2} \end{bmatrix} \Bigg|_{\substack{x=x^* \\ u=0}} = \begin{bmatrix} 0_{n \times n} & \mathbf{I}_n \\ -\mathbf{M}(\mathbf{q}^*)^{-1} \mathbf{J}_G(\mathbf{q}^*) & 0_{n \times n} \end{bmatrix} \quad \text{og} \quad \mathbf{B} = \begin{bmatrix} 0_{n \times m} \\ \mathbf{M}(\mathbf{q}^*)^{-1} \mathbf{Q} \end{bmatrix} \Bigg|_{\substack{x=x^* \\ u=0}}$$

hvor $\mathbf{J}_G(\mathbf{q}) = \frac{\partial \mathbf{G}(\mathbf{q})}{\partial \mathbf{q}}$ er den såkalte **Jacobimatrisen** til $\mathbf{G}(\cdot)$ for en gitt \mathbf{q} .

^a1) Utrykket i klammeparentesene i \mathbf{f}_2 er lik null ved arbeidspunktet, slik at vi ikke trenger å differensiere $\mathbf{M}(\cdot)$; 2) Utrykket $\mathbf{H}(\mathbf{x}_1, \mathbf{x}_2)\mathbf{x}_2$ er "kvadratisk" i henhold til \mathbf{x}_2 , så vi kan se bort fra denne siden $\mathbf{x}_2^* = 0$.

2.6. Overføringsfunksjoner og Laplace-domenet

▶ 2X17-DF3g8

Alternative kilder: Kap. 8 i [Haugen, 2023]; Kap. 3 og 4 i [Seborg et al., 2016]; [Balchen et al., 2016].

Ved å representere lineære, tids-invariante systemet (se (2.5)) vha. overføringsfunksjoner, kan vi blant annet studere systemets respons i forhold til forskjellige frekvenser (noe vi skal se nærmere på i §11).

2.6.1 Reguleringsløyfe i Laplace-domenet

En kort repetisjon om hvordan man kan representere lineære dynamiske systemer ved hjelp av overføringsfunksjoner (også kalt transfer funksjoner) følger.

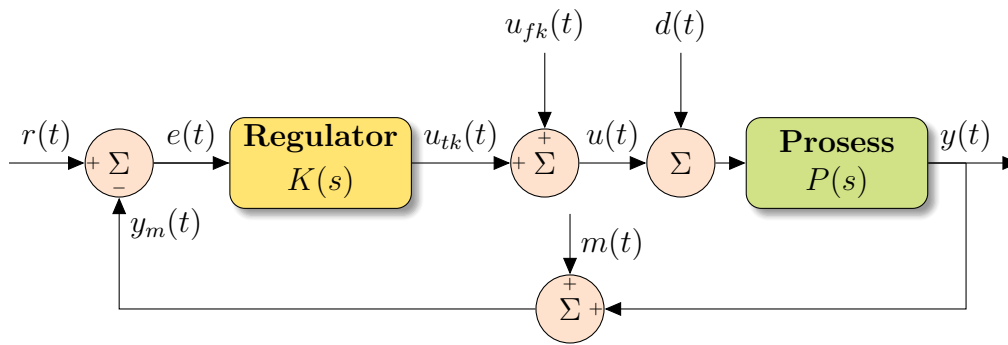
Gitt systemet i figur 2.7. **Anta at:** 1) det er et monovariabelt system med ett pådrag: $y, u \in \mathbb{R}$; 2) det er ingen fremoverkobling: $u_{fk} = 0$. Vi får da

$$Y(s) = \frac{P(s)}{1 + K(s)P(s)} D(s) + \frac{K(s)P(s)}{1 + K(s)P(s)} (R(s) - M(s)), \quad (2.8)$$

og dermed

$$E(s) = \frac{1}{1 + K(s)P(s)} R(s) - \frac{P(s)}{1 + K(s)P(s)} D(s) + \frac{K(s)P(s)}{1 + K(s)P(s)} M(s).$$

Vi vil se nærmere på noen av del-funksjonene som inngår i uttrykket over i §11 hvor vi skal se på nærmere på frekvensanalyse og robust regulering.



Figur 2.7: Reguleringsystem som overføringsfunksjoner i Laplace-domenet.

NB! Jeg misbruker her notasjonen litt for å gjøre diagrammet så tydelig som mulig, selv om det strengt tatt burde vært brukt $Y(s)$, $R(s)$ og $E(s)$, etc., i stedet for $y(t)$, $r(t)$ og $e(t)$.

2.6.2 Laplacetransformasjonen ▶ ZGPtPkTft8g

For å minne os på hvordan man utleder overføringsfunksjoner trenger vi å minne oss på definisjonen av den ensidige Laplace-transformasjonen.⁷

Laplacetransformasjon: Definert for en kontinuerlig (integrerbar) funksjon $f(\cdot)$ av tiden t som^a

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

Her et s et komplekst tall (altså $s \in \mathbb{C}$), ofte kalt *Laplace-variabelen*.

^aMerk at jeg er kanskje i overkant pedantisk når det kommer til notasjonen her, siden jeg bruker $f(\cdot)$ og ikke $f(t)$; grunnen til dette er at $f(t)$ betyr verdien til funksjonen $f(\cdot)$ ved tiden t .

Merk 1: Dette er det ensidige Laplace-transformasjonen, ettersom vi kun ser på tidsintervallet $[0, \infty)$ og ikke hele tallinjen. Dette kan vi gjøre ved å anta at $f(t) \equiv 0$ for alle $t < 0$.

Merk 2: Vi vil i disse notatene hovedsakelig bruke stor bokstav til å betegne Laplace-transformasjoner, f.eks. $F(s) = \mathcal{L}\{f(\cdot)\}$. Dette misbrukes tidvis i noen lærebøker med vilje for å “spare på symboler” (se f.eks. [Balchen et al., 2016]), men kan (etter min mening) fort føre til både dårlige vaner og misforståelser.

Merk 3: For våre formål er Laplacetransformasjonen et bytte av variabler (derav transformasjon) fra tiden t til Laplace-variabelen s . Variabelen s tar verdier i det komplekse planet \mathbb{C} ; det vil si $s = \sigma + j\omega$, hvor man (intuitivt sett) kan tenke på σ som en eksponentiell faktor og ω som frekvensen ($j := \sqrt{-1}$). Dermed har vi fra Eulers formel at $e^{st} = e^{(\sigma + j\omega)t} = e^{\sigma t} \cdot (\cos(\omega t) + j \sin(\omega t))$. Vi har derfor følgende intuitive «forklaring» på hva Laplacetransformasjonen gjør: For et hvert punkt $(\sigma, j\omega)$ i det komplekse plan, så ser vi på hvor mye et signal $f(t)$ avhenger av den eksponentielle faktoren, $e^{\sigma t}$, og en oscillerende del, gitt av $e^{j\omega}$, over all positiv tid. Siden løsningen $y(t)$ til lineære, tidsinvariante systemer alltid er bygd opp av ledd på formen $e^\sigma e^{j\omega}$, så vil dermed Laplacetransformasjonen av y , $Y(s)$, bli uendelig når er lik et slikt ledd, altså $s = \sigma + j\omega$, noe som da samsvarer med en av polene til systemets overføringsfunksjon.

⁷På engelsk bruker man Laplace transform til å bety selve operasjonen, mens Laplace transformasjon (“transformation”) betyr resultatet av en transform. Jeg har selv blitt opplært til det motsatte, så jeg er dermed litt forvirret og vil prøve å holde meg til transformasjon for enkelhetens skyld.

Vanlige transformasjoner:

La $f(t)$ betegne en glatt (deriverbar) funksjon av tid⁸, og la $F(s) = \mathcal{L}\{f(t)\}$. Vi bruker også et dot-symbol til å betegne tidsderivatet: $\dot{f}(t) := \frac{d}{dt}f(t)$. En liste med vanlige transformasjon er vist i tabell 2.2.

Tabell 2.2: Vanlige Laplacetransformasjoner ($f(t) = 0$ for $t < 0$).

$f(t)$	$F(s)$
$\delta(t)$	1
1	$\frac{1}{s}$
e^{-at}	$\frac{1}{s+a}$
te^{-at}	$\frac{1}{(s+a)^2}$
t	$\frac{1}{s^2}$
$\int_0^t f(\sigma)d\sigma$	$\frac{1}{s}F(s)$
$\dot{f}(t)$	$sF(s) - f(0)$
$\ddot{f}(t)$	$s^2F(s) - sf(0) - \dot{f}(0)$
$\sin(\omega t)$	$\frac{\omega}{s^2+\omega^2}$
$\cos(\omega t)$	$\frac{s}{s^2+\omega^2}$
$f(t - \theta)$	$e^{-\theta s}F(s)$

Eksempel 2.9. Tidsderivat: Som et eksempel, la oss vise at $\mathcal{L}\{\dot{x}(t)\} = sX(s) - x(0)$:

$$\begin{aligned}
 \mathcal{L}\{\dot{x}(t)\} &= \int_0^\infty \dot{x}(t)e^{-st}dt \\
 &= \int_0^\infty \frac{d}{dt}[x(t)e^{-st}] - (-x(t)se^{-st})dt \\
 &= [x(t)e^{-st}]_0^\infty + s \int_0^\infty x(t)e^{-st}dt \\
 &= 0 - x(0)e^0 + sX(s) \\
 &= sX(s) - x(0).
 \end{aligned}$$

2.6.3 Fra differensialligninger til overføringsfunksjoner



Alternative kilder: §8.10 i [Haugen, 2023].

Vi skal nå se hvordan man kan konvertere et mono-variabelt system fra tilstandsromform,

⁸Jeg misbruker her notasjonen litt, ettersom $f(t)$ strengt tatt er verdien av funksjonen $f(\cdot)$ ved tiden t .

rettere sagt LTI-systemer på matriseform, til overføringsfunksjoner.

Antagelser:

1. Prosessen er lineær:

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}u, \quad (2.9a)$$

$$y = \mathbf{C}\mathbf{x} + \mathbf{D}u. \quad (2.9b)$$

2. y og u er skalarer ($y, u \in \mathbb{R}$);^a

3. Initialbetingelsene er lik null: $\mathbf{x}(0) = 0$.

Tilsvarende overføringsfunksjon er da

$$\frac{Y(s)}{U(s)} = P(s) := \mathbf{C}(\mathbf{I}_n s - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D}. \quad (2.10)$$

^aProsedyren over er også gyldig for flere innganger og /eller utganger, altså når \mathbf{u} og/eller \mathbf{y} er vektorer.

Dette er enkelt å vise: vi har $\mathcal{L}\{y\} = Y(s) = \mathcal{L}\{\mathbf{C}\mathbf{x} + \mathbf{D}u\} = \mathbf{C}\mathbf{X}(s) + \mathbf{D}U(s)$, samt

$$\mathcal{L}\{\dot{\mathbf{x}}\} = s\mathbf{X}(s) - \mathbf{x}(0) = \mathcal{L}\{\mathbf{A}\mathbf{x} + \mathbf{B}u\} = \mathbf{A}\mathbf{X}(s) + \mathbf{B}U(s) \implies \mathbf{X}(s) = (\mathbf{I}_n s - \mathbf{A})^{-1}\mathbf{B}U(s).$$

Eksempel 2.10. Gitt

$$\begin{aligned} \dot{\mathbf{x}} &= \begin{bmatrix} 0 & 1 \\ -a & d \end{bmatrix} \mathbf{x} + \begin{bmatrix} 0 \\ b \end{bmatrix} u, \\ y &= \begin{bmatrix} 1 & 0 \end{bmatrix} \mathbf{x}. \end{aligned}$$

Vi har (se § B.3.1 for den inverse til en 2×2 matrise)

$$(\mathbf{I}_2 s - \mathbf{A})^{-1} = \left(\begin{bmatrix} s & -1 \\ a & s - d \end{bmatrix} \right)^{-1} = \frac{1}{s(s-d) + a} \begin{bmatrix} s - d & 1 \\ -a & s \end{bmatrix}.$$

Vi har dermed fra (2.10) at

$$\frac{Y(s)}{U(s)} = \frac{1}{s^2 - ds + a} \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} s - d & 1 \\ -a & s \end{bmatrix} \begin{bmatrix} 0 \\ b \end{bmatrix} = \frac{b}{s^2 - ds + a}.$$

2.6.4 Propre og strengt propre overføringsfunksjoner

Alle dynamiske systemer som kan representeres ved hjelp av lineære differensialligninger tilsvarer en proper overføringsfunksjon (som oftest en strengt proper en for fysiske prosesser). En overføringsfunksjon som ikke er proper kan ikke realiseres, og er derav generelt ikke av interesse for oss. Men hva vil det si at en overføringsfunksjon er proper?

Gitt en overføringsfunksjon, $G(s)$, la oss betegne polynomene i teller og nevner som henholdsvis $\mathcal{T}_G(s)$ og $\mathcal{N}_G(s)$:

$$G(s) = \frac{\mathcal{T}_G(s)}{\mathcal{N}_G(s)}.$$

Eksempel: hvis $G(s) = \frac{3s+2}{(s+1)^2}$ så er $\mathcal{T}(s) = 3s + 2$ og $\mathcal{N}(s) = (s + 1)^2$.

Properhet: Vi sier at $G(s) = \frac{\mathcal{T}_G(s)}{\mathcal{N}_G(s)}$ er en

- **proper overføringsfunksjon** hvis $\mathcal{N}_G(s)$ er av minst like høy orden som $\mathcal{T}_G(s)$, slik at $|G(s)| < \infty$ når $|s| \rightarrow \infty$.
- **strengt proper overføringsfunksjon** hvis $\mathcal{N}_G(s)$ har høyere orden enn $\mathcal{T}_G(s)$, slik at $|G(s)| \rightarrow 0$ når $|s| \rightarrow \infty$.

Eksempler:

- $\frac{(s+1)}{(s+2)}$ er proper siden graden til både nevneren, $s + 2$, og telleren, $s+1$, er 1;
- $\frac{(s+1)}{(s+2)^2} = \frac{(s+1)}{(s^2+4s+2)}$ er strengt proper siden $\mathcal{N}_G(s)$ har 1 høyere grad enn $\mathcal{T}_G(s)$;
- $\frac{(s+1)^2}{(s+2)} = \frac{(s^2+2s+1)}{(s+2)}$ er ikke proper siden tellers høyeste ledd er s^2 , mens nevners er s .

2.6.5 Fra overføringsfunksjoner til differensialligninger



Alternative kilder: §8.8-8.9 i [Haugen, 2023].

Det går også noen ganger an å gå fra en overføringsfunksjon til en differensialligning.

⚠ NB! Det er ikke en unik tilstandsromform for en hver overføringsfunksjon.

Prosedyre: Gitt en overføringsfunksjon fra $u(t)$ til $y(t)$ på formen

$$\frac{Y(s)}{U(s)} = G(s) = \frac{\mathcal{T}(s)}{\mathcal{N}(s)}. \tag{2.11}$$

Fra (2.10) har vi at

$$\mathbf{C}(\mathbf{I}_n s - \mathbf{A})^{-1} \mathbf{B} + \mathbf{D} = \mathbf{C}(\mathbf{I}_n s - \mathbf{A})^{-1} \mathbf{B} + \mathbf{D} = \frac{\mathbf{C} \text{adj}(\mathbf{I}_n s - \mathbf{A}) \mathbf{B} + \det(\mathbf{I}_n s - \mathbf{A}) \mathbf{D}}{\det(\mathbf{I}_n s - \mathbf{A})} = \frac{\mathcal{T}(s)}{\mathcal{N}(s)}.$$

hvor $\text{adj}(\mathbf{M})$ er den *adjointe/adjunkte* til matrisen \mathbf{M} og $\det(\mathbf{M})$ er determinanten. Ma prøver derfor å finne \mathbf{A} , \mathbf{C} og \mathbf{D} som tilfredsstill $\mathcal{N}(s) = \det(\mathbf{I}_n s - \mathbf{A})$ og $\mathcal{T}(s) = \mathbf{C} \text{adj}(\mathbf{I}_n s - \mathbf{A}) \mathbf{B} + \det(\mathbf{I}_n s - \mathbf{A}) \mathbf{D}$.

La oss illustrere dette et et eksempel

Eksempel 2.11. La

$$\frac{Y(s)}{U(s)} = k \frac{s+b}{s+a}.$$

La oss innføre $Y(s) = k(s+b)X(s)$ slik at

$$X(s) = \frac{1}{s+a}U(s) \implies (s+a)X(s) = U(s).$$

Dette gir $\dot{x} + ax = u$, mens fra $Y(s) = k(s+b)X(s)$ har vi $y = k(\dot{x} + bx)$, slik at

$$y = k(-ax + u + bx).$$

Dermed har vi $C = k(b-a)$, $D = k$, mens $A = -a$ og $B = 1$. På den annen side, så er ikke løsningen unik siden vi også kan ta

$$A = -a, \quad B = k, \quad C = k, \quad D = b-a,$$

slik at

$$\begin{aligned} \dot{x} &= -ax + (b-a)u \\ y &= kx + ku. \end{aligned}$$

Faktisk finnes det uendelig mange alternativer her.

Følgende prosedyre kan forenkle konverteringen for visse systemer:

Prosedyre for strengt-proper mono-variabelt system: Gitt en strengt proper overføringsfunksjon på følgende form:

$$G(s) = \frac{Y(s)}{U(s)} = \frac{b_m s^m + b_{m-1} s^{m-1} + \dots + b_0}{s^n + a_{n-1} s^{n-1} + \dots + a_0} = \frac{\mathcal{T}(s)}{\mathcal{N}(s)}$$

hvor $1 \leq m < n$.

La oss introdusere $X_1(s) = U(s)/\mathcal{N}(s)$, slik at $Y(s) = \mathcal{T}(s)X_1(s)$. Fra den inverse Laplace-transformasjon (hvor $\mathcal{L}^{-1}\{s^k X_1(s)\} = x_1^{(k)}(t) = \frac{d^k x_1(t)}{dt^k}$) får vi dermed

$$x_1^{(n)}(t) + a_{n-1}x_1^{(n-1)}(t) + \dots + a_0x_1(t) = u(t)$$

(antok der at $x_1(t)$ og dens derivater er lik 0 ved tiden 0) og

$$y(t) = b_m x_1^{(m)}(t) + b_{m-1} x_1^{(m-1)}(t) + \dots + b_0 x_1(t).$$

Ved å ta $x(t) = [x_1, x_2, \dots, x_n]^T$ kan dette skrives på formen

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

med

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & & \vdots \\ -a_0 & -a_1 & -a_2 & \cdots & -a_{n-1} \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}, \quad \text{og} \quad \mathbf{C} = [b_0 \quad b_1 \quad \cdots \quad b_m].$$

Eksempel 2.12. Gitt følgende overføringsfunksjon

$$\frac{Y(s)}{U(s)} = k \frac{as + b}{s^2 + cs + d}.$$

Fra prosedyren over har $Y(s) = (as + b)X_1(s)$ hvor $(s^2 + cs + d)X_1(s) = kU(s)$, og dermed

$$\ddot{x}_1 + c\dot{x}_1 + dx_1 = ku.$$

Dette gir

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= -dx_2 - cx_2 + ku \\ y &= bx_1 + ax_2. \end{aligned}$$

Legg merke til at vi får en annen tilstandsromform ved å i stedet ta $X(s) = \alpha \cdot U(s)/(s^2 + cs + d)$ for en hvilken som helst konstant koeffisient $\alpha > 0$

2.6.6 (Ikke-) Minimum-fase systemer

Alternative kilder: Eksempel 6.1 i [Seborg et al., 2016]; Brian Douglas video.

For systemer som ikke er minimum fase kan en for aggressiv reguleringsstrategi føre til lav ytelse, og i verste fall ustabilitet. Det er derfor viktig å kunne identifisere denne typen systemer.

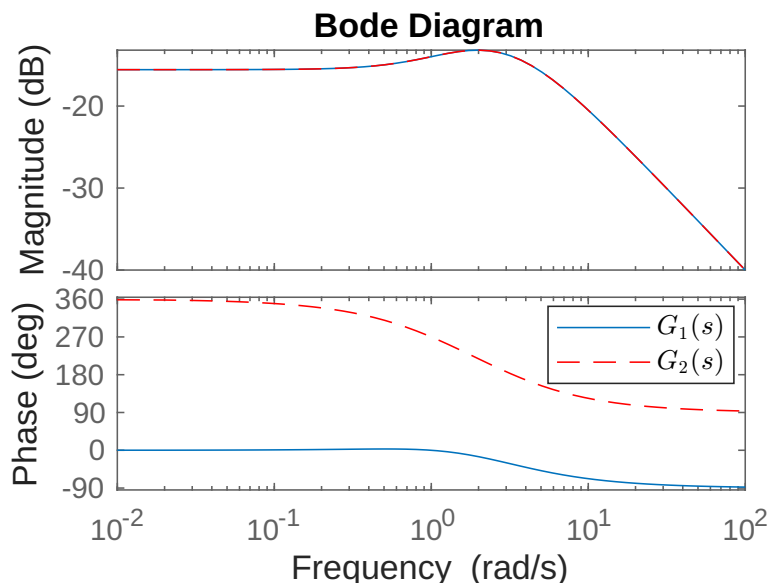
Eksempler på fysiske systemer som ikke er minimum-fase:

- Sykler; se <https://youtu.be/9cNmUNHSBac?t=141>
- Lukeparkering
- Fly
- Oppsving av invertert pendel på tralle

Som navnet tilsier, så er det fasen til et system som angir om det er minimum fase eller ikke, retttere sagt fasen til dets overføringsfunksjon. La oss illustrere konseptet med et eksempel: Legg merke til at magnituden til overføringsfunksjonene

$$G_1(s) = \frac{s + 1}{(s + 2)(s + 3)} \quad \text{og} \quad G_2(s) = \frac{-s + 1}{(s + 2)(s + 3)}$$

er de samme, altså $|G_1(j\omega)| \equiv |G_2(j\omega)|$. Slik som vist i figur 2.8 gjelder dog dette ikke for fasen deres, altså $\angle G_1(j\omega) \neq \angle G_2(j\omega)$ for alle $\omega > 0$.



Figur 2.8: Viser Bode-plottet til overføringsfunksjonene $G_1(s) = \frac{s+1}{(s+2)(s+3)}$ og $G_2(s) = \frac{-s+1}{(s+2)(s+3)}$. Det er tydelig at de har lik forsterkning (magnitudo), mens G_2 har betydelig større sprik i sine faseforskyving og er dermed ikke minimum fase.

Vi sier derfor at:

Overføringsfunksjonen $G(s)$ er **minimum-fase** hvis $G(s)$ hverken har nullpunkter eller poler i høyre halvplan, ei heller tidsforsinkelser, siden den da har minimum spenn i dets faseendring for alle overføringsfunksjon med samme magnituderespons.

Vi sier at $G(s)$ er **ikke-minimum-fase** hvis den har minst én pol og/eller et nullpunkt i høyre halvplan og/eller en tidsforsinkelse.

Eksempler:

- $\frac{1}{s+1}$ og $\frac{2s+1}{s+1}$ er minimum fase (poler og nullpunkt i venstre halvplan).
- $\frac{1}{s-1}$ og $\frac{2s-1}{s+1}$ er ikke minimum-fase pga. henholdsvis pol og nullpunkt i høyre halvplan;
- $\frac{1}{s+1}e^{-3s}$ er ikke minimum-fase pga. tidsforsinkelsen.

2.6.7 Kausalitet og Realiserbare overføringsfunksjoner

En viktig karakteristikk til en overføringsfunksjon $G(s)$ er om den er **realiserbar**; altså om det faktisk er mulig å implementere den (f.eks. i en datamaskin eller som en elektrisk krets). For dette må $G(s)$ hovedsakelige tilfredsstillende tre ting:

Krav for at en overføringsfunksjon skal være realiserbar:

1. **Stabilitet:** Den er stabil (ingen poler i høyre halvplan);
2. **Properhet:** Den er proper (høyeste ledd i nevner er større eller lik det i teller; se § 2.6.4);
3. **Kausalitet:** Den inneholder ingen inverteringer av tidsforsinkeler, altså prediktive ledd.

Eksempler på realiserbare overføringsfunksjoner:

- $\frac{(s+1)}{(s+2)}$ → proper, stabil og minimum fase.
- $\frac{(s-1)}{(s+2)^2}$ → strengt proper, stabil, men ikke minimum fase.
- $\frac{ke^{-2s}}{1+3s}$ → stabil med tidsforsinkelse.

Eksempler på overføringsfunksjoner som ikke er realiserbare:

- $\frac{(s+1)^2}{(s+2)}$ → stabil og minimum fase, men ikke proper.
- $\frac{(s+1)}{(s-2)(s+3)}$ → strengt proper, men ikke stabil og dermed ikke minimum fase.
- $\frac{ke^{2s}}{1+3s}$ → stabil, men med prediktivt element.

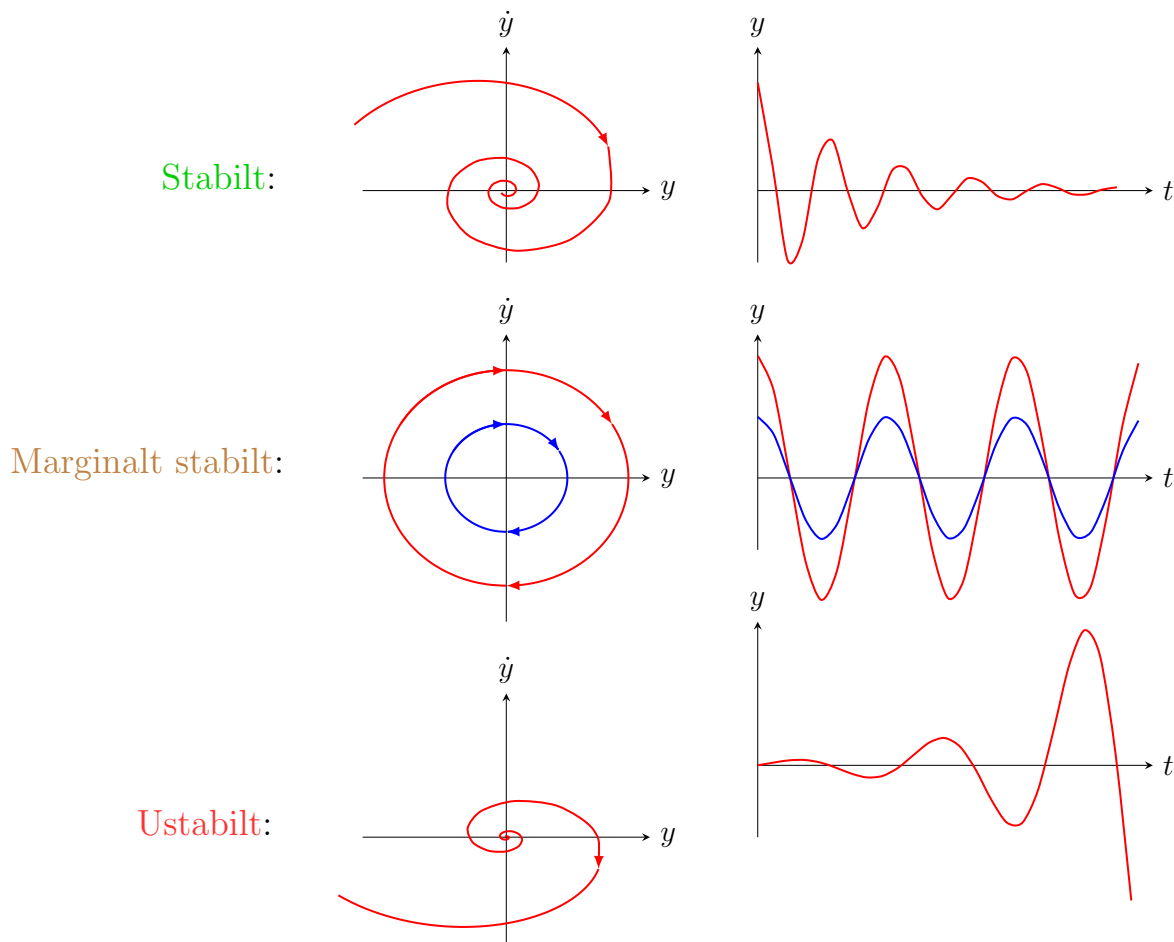
2.7. Stabilitet

Stabilitet er faktisk et ganske vidt begrep, og det finnes mange forskjellige varianter og definisjoner. For lineær, tids-invariant (LTI) systemer, som vi hovedsakelig er interessert i, er begrepet “stabilitet” heldigvis noe mer klart og avgrenset. Spesifikt, så er det to begreper som vi er interessert i: inngang–utgang stabilitet og stabiliteten til et punkt i systemets tilstandsrom.

2.7.1 Inngang-utgang-stabilitet

Inngang-utgang-stabilitet: La $u(t)$ være et signal med endelig varighet og begrenset amplitude (f.eks. en enhetspuls eller én enkelt firkantpuls). Et LTI system med én utgang, $y(t)$, og én inngang, $u(t)$, sies da å være

- **inngangs-utgangs-stabilt** dersom $|y(t)| < \infty$ for alle tidpunkt $t \geq 0$ (dette kalles også et **marginal inngang-utgang-stabilitet**);
- **asymptotisk inngangs-utgangs-stabilt** dersom det er inngangs-utgangs-stabilt og $y(t) \rightarrow 0$ når $t \rightarrow \infty$;



Figur 2.9: Illustrasjon av forskjellige stabilitets-karakteristikker til punktet $y = 0$; venstre = faseplanet; høyre = tidsplanet; se også: <https://youtu.be/KO6sonZb1c0?t=252>

- **ustabilt** dersom det ikke er stabilt.

Følgende type stabilitet er også viktig (se, f.eks., eksempel 7.2):

Intern stabilitet: En tilbakekoblingsløyfe på formen (2.8), hvor $K(s)$ er minimum fase (alle dens poler og nullpunkter er i venstre halvplan; se § 2.6.6) sies å være **internt stabil** hvis alle interne overføringsfunksjoner ($Y(s)/R(s)$, $Y(s)/D(s)$, etc.) er stabile (deres poler er i venstre halvplan).

2.7.2 Stabilitet i tilstandsrommet*



Del III

Modellering og simulering

3. Modellering

Alternative kilder: Kap. 2 i [Balchen et al., 2016], kap. 4 i [Haugen, 2023], Kap. 3 i [Bjørvik and Hveem, 2014].

I dette kapitlet skal vi se på hvordan man kan lage matematiske modeller av dynamiske systemer. Mer spesifikt, så ønsker vi å finne ett sett med ordinære differensiallikninger som beskriver systemets *dynamikk*, altså hvordan systemet (rettere sagt, dets *tilstander*) endrer seg over tid når det blir påvirket av eksterne krefter (f.eks. kontrollpådrag og forstyrrelser).

Slike modeller er selvsagt alltid forenklinger, og vil i varierende grad representere det virkelige systemet. Følgende velkjente (og nærmest obligatoriske) utsagn fanger dette godt:

«Alle modeller er feil, men noen er nyttige»

Så hvorfor er slike matematiske modeller så nyttige? Dynamiske modeller spiller en sentral rolle innen reguleringsteknikk, og har en rekke andre bruksområder:^a

- **Forbedre forståelsen av prosessen:** Dynamisk modeller og datasimuleringer lar en studere prosessatferd uten å måtte «forstyrre» den ekte prosessen.
- **Bygge simulatorer:** Matematiske modeller trengs for å lage realistiske simulatorer. Et godt eksempel på dette er flytrenings-simulatorer som brukes i luft- og romfartsindustrien.
- **Utvikle reguleringsstrategier:** En matematisk modell av en prosess tillater å evaluere mulige reguleringsstrategier.

^aMatematiske modeller er også viktig for tilstandsestimering, feildiagnose, ytelses- og robusthets-analyser, etc., etc.

Hvordan lage en matematisk modell? Det finnes tre måter å modellere et system:

- **Teoretisk modellering:** utlede modeller fra fysikkens lover («førsteprinsipper»).
- **Empirisk modellering:** tilpasse modeller fra å eksperimentere/data.
- **Semi-empirisk modellering:** en blanding av de to forrige.

Vi skal hovedsakelig se på teoretisk modellering i dette faget, selv om også empirisk modellering vil dukke opp i form av tilpasning av en modell fra en sprangrespons (se § 7.3).

Teoretisk modellering: Hva er vi ute etter? Vi er ute etter å bruke såkalte grunn-/første-prinsipper (lover fra fysikken) til å utlede (ordinære¹) differensialligninger som beskriver et systems dynamikk på en *god nok* måte. Det vil si at de gir en tilfredsstillende representasjon av systemet, med en passende balanse mellom enkelhet og realisme (variasjoner i tyngdekraften kan være viktig for en romraket, men kanskje ikke fullt så mye for en pendel).

Ved å se bort fra eksterne forstyrrelser, så skal vi utlede modeller på tilstandsromform:

$$\dot{x}_1 = f_1(x_1, x_2, \dots, x_n, u_1, \dots, u_m), \quad (3.1a)$$

$$\dot{x}_2 = f_2(x_1, x_2, \dots, x_n, u_1, \dots, u_m), \quad (3.1b)$$

$$\vdots \quad (3.1c)$$

$$\dot{x}_n = f_n(x_1, x_2, \dots, x_n, u_1, \dots, u_m), \quad (3.1d)$$

hvor $x_i = x_i(t) \in \mathbb{R}$, $i = 1, 2, \dots, m$, er systemets $n \geq 1$ tilstander. og $u_j = u_j(t) \in \mathbb{R}$, $j = 1, 2, \dots, m$, $m \geq 0$, er pådragene.

Slike modeller blir som regel funnet ved hjelp av bevaringslover (også kalt konserveringslover) for en eller annen fysisk størrelse (for eksempel masse, energi, bevegelsesmengde (momentum), elektrisk ladning, vinkelmoment, etc.). Vi skal se på noen slike metoder i dette kapitlet.

3.1. Masse- og energi-balanse

Alternative kilder: [Bjørvik and Hveem, 2014].

Masse- og energibalans baserer seg på prinsippene at enten masse eller energi (i et lukket system) forblir konstant (er bevart) hvis det ikke er noen tilførsel eller tap, samt at raten til en hver endring tilsvarer differansen mellom raten av tilførsel og tap.²

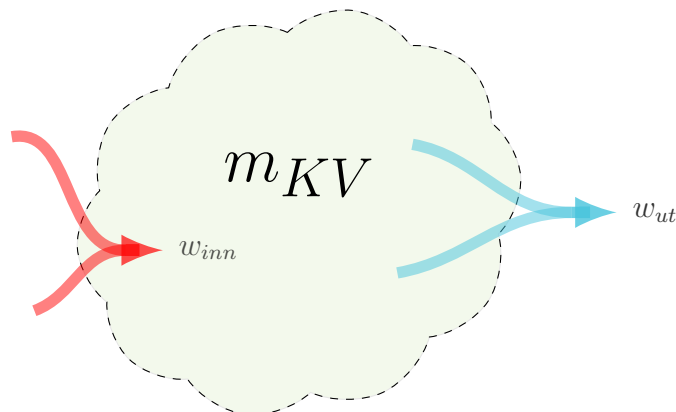
Hvorfor skal du lære dette? Spesielt innen prosessindustrien finner man som oftest teoretiske modeller ved hjelp av enten masse- eller energibalans.

3.1.1 Massebalans 90aAvUEoDzA

Massebalansen tilsvarer følgende meget avanserte konsept: Hvis det i løpet av 1 minutt strømmer 10 kg med væske inn i en tank og 2 kg ut, så vil det etter 1 minutt være 8 kg mer væske i tanken enn før!

¹Dynamikken til en rekke systemer kan kun modelleres ved hjelp av *partielle differensialligninger*, for eksempel, myke (soft) roboter, samt varme- og Navier–Stokes ligningene for henholdsvis varmfordeling og væskestrøm. For reguleringstekniske formål kan man noen ganger approksimere disse vha. ordinære differensialligninger.

²For å bevare den psykiske helsen til faglærer så skal vi holde oss til systemer hvor vi ikke trenger å ta høyde for Einsteins spesielle relativitetsteori, og derav hans velkjente formell, $E = mc^2$, som gir en sammenheng mellom masse og energi.



Figur 3.1: Illustrasjon av et kontrollvolum i grønt med ytterkant gitt av den stiplende linjen. Den akkumulerte massen i kontrollvolumet ved et gitt tidspunkt er m_{KV} , mens massestrømmene w_{inn} og w_{ut} flyter henholdsvis inn og ut av kontrollvolumet.

I dette eksempelet var tanken et såkalt **kontrollvolum** (KV). Det vil si, en (kunstig) avgrenset region med en klart definert, og som regel konstant, ytterkant. Et mer abstrakt kontrollvolum med massestrømmer inn og ut er vist i figur 3.1. Innen dette kontrollvolumet er den akkumulerte massen (målt i kg), altså all masse i volumet, ved et tidspunkt t gitt ved $m_{KV}(t)$. Ved å se på differansen mellom de samlede masstrømmene (målt i kg/s) inn og ut av volumet, betegnet henholdsvis $w_{inn}(t)$ og $w_{ut}(t)$, så kan man finne den momentane endringen i den akkumulerte massen i volumet ved dette tidspunktet ved hjelp av følgende:

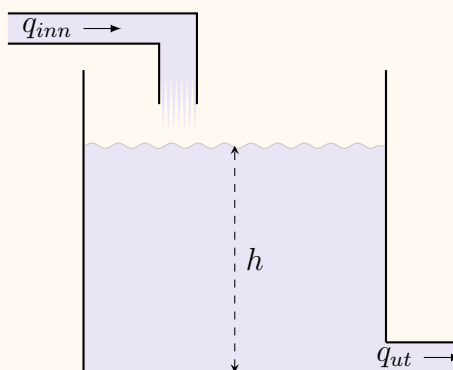
Massebalanse: For et lukket system med et gitt kontrollvolum (KV) har man

$$\underbrace{\frac{d}{dt}m_{KV}(t)}_{\text{Endringsraten til akkumulert masse}} = \underbrace{w_{inn}(t)}_{\text{rate av masse inn}} - \underbrace{w_{ut}(t)}_{\text{rate av masse ut}} \quad (3.2)$$

La oss se på nok et eksempel relatert til en tank:

Eksempel 3.1. Væskenivået i en tank:

En enkel tank er vist i figuren til høyre. Der er q_{inn} væskestrømmen inn og q_{ut} er væskestrømmen ut, begge med enhet $[m^3 s^{-1}]$. For volumet med væske i tanken (målt i kubikkmeter) bruker vi symbolet V . Det er dermed naturlig å definere tankens indre som kontrollvolumet. Hvis ρ $[kg/m^3]$ er massetettheten til væsken, så er den akkumulerte massen da $m_{KV} = \rho \cdot V$, mens $w_{inn} = \rho \cdot q_{inn}$ og $w_{ut} = \rho \cdot q_{ut}$. Fra masse-balanse-ligningen (3.2) får vi dermed at



$$\frac{d}{dt}(\rho V) = \rho(q_{inn} - q_{ut}).$$

La oss anta at 1) massetettheten ρ er (tilnærmet) konstant, 2) at avstanden mellom midten av utløpet og tankens bunn er neglisjerbart i forhold til væskeshøyden i tanken, og 3) at hvert

horisontalt tverrsnitt av tanken har et areal A [m²], Endring i høyden, h , til væsken i tanken målt i meter er dermed

$$\dot{h} = \frac{1}{A} [q_{inn} - q_{ut}].$$

Dette er et første-ordens system (bare ett derivat) med én enkelt tilstand, nemlig høyden h .

⚠ Massebevarelse betyr generelt sett ikke bevarelse av volum. Dette krever antakelsen at massetettheten (til væsken eller gassen) er konstant. I realiteten endres denne f.eks. både ved endring i temperatur og trykk (se f.eks. [Avogadros lov](#) for ideelle gasser).

Men: Konstant massetetthet (inkompressibilitet) er dog en god antagelse for de fleste væsker; se f.eks. [Cengel and Cimbala, 2013]. Jobber man med gasser, derimot, er det viktig å vurdere om dette er en fornuftig antagelse eller ikke for prosessen man jobber med siden denne antagelsen, tross alt, i ganske stor grad, forenkler den matematiske modellen.

Oppgave 3.1. I eksempel 3.1 kom vi fram til differensialligningen $\dot{h} = \frac{1}{A} [q_{inn} - q_{ut}]$.

Si at q_{ut} tilsvarer strømmen ut via et åpent rør og ut i friluft. Hvorfor er da ikke

$$q_{ut} = k_v h,$$

for en eller annen konstant $k_v > 0$, en veldig realistisk antagelse?

Den vanligste modellen til en slik utstrøm er (mer om dette straks)

$$q_{ut} = k_v \sqrt{h}.$$

Hvorfor er dette en mer realistisk modell?

Hint: Hva er løsningene på differensialligningene hvis $q_{inn} = 0$?

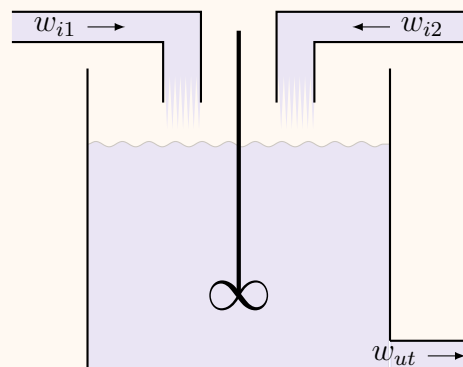
La oss se på et annet eksempel som er hakket mer komplisert:

Eksempel 3.2. Blandetank:

Vi skal nå finne de dynamiske ligningene som beskriver endringen til blandeforholdet mellom to stoffer, stoff F (f.eks. fanta) og stoff R (f.eks. rosévin), i en enkel blandetank. Blandetanken er vist i figuren til høyre, hvor

w_{i1} , w_{i2} , w_{ut} er massestrømmer [kg/s];

b_{F1} , b_{F2} , b_F er blandingsforholdet til stoff F for henholdsvis de to innløpene og utløpet.



La m [kg] betegne den akkumulerte massen i tanken, som igjen tilsvarer $m = \rho V$ hvor ρ kg m^{-3} er massetettheten til den blandede væsken og V er volumet (i kubikkmeter). Massebalanse-ligningen (3.2) gir oss dermed

$$\dot{m} = \frac{d}{dt}(\rho V) = w_{i1} + w_{i2} - w_{ut}. \quad (3.3)$$

Spørsmål: Før vi går videre, funder gjerne litt på følgende spørsmål:

1. Hva er aktuelle tilstandsvariabler? Og hvordan kan vi måle disse?
2. Hva er mulige pådrag?
3. Hva er mulige forstyrrelser?
4. Er konstant massetetthet en fornuftig antagelse?

En «fasit» er gitt i slutten av eksemplet.

Vi er også interessert i å finne hvordan blandingsforholdet b_F til stoff F endrer seg. Vi vet at den totale masseandel for stoff F i tanken er $m_F = V \rho b_F = m b_F$, og dermed

$$\dot{m}_F = \frac{d}{dt}(m b_F) = \dot{m} b_F + m \dot{b}_F = w_{i1} b_{1F} + w_{i2} b_{2F} - w_{ut} b_F,$$

hvor vi har brukt [Produktregelen](#). Ved å trekke fra $\dot{m} b_F$ på begge sidene og så sette inn for \dot{m} fra (3.3), så finner vi at

$$m \dot{b}_F = w_{i1}(b_{1F} - b_F) + w_{i2}(b_{2F} - b_F).$$

Dette kan vi igjen skrive som

$$\dot{b}_F = \frac{1}{m} [w_{i1}(b_{1F} - b_F) + w_{i2}(b_{2F} - b_F)]. \quad (3.4)$$

La oss nå anta følgende:

- tankens tverrsnittsareal, A , er konstant;
- utstrømmen w_{ut} er gitt av $w_{ut} = c_{ut} \sqrt{\rho g h}$, hvor h er væskehøyden i tanken.

Disse antagelsen betyr jo at $m = (\rho V) = \rho A h$ og $w_{ut} = c_{ut} \sqrt{m g / A}$.

Vi har dermed to naturlige kandidater til tilstander, nemlig $x_1 = m$ og $x_2 = b_A$. Deres dynamiske ligninger kan skrives på tilstandsromform som:

$$\dot{x}_1 = w_{i1} + w_{i2} - c_{ut} \sqrt{g/A} \sqrt{x_1} \quad (3.5a)$$

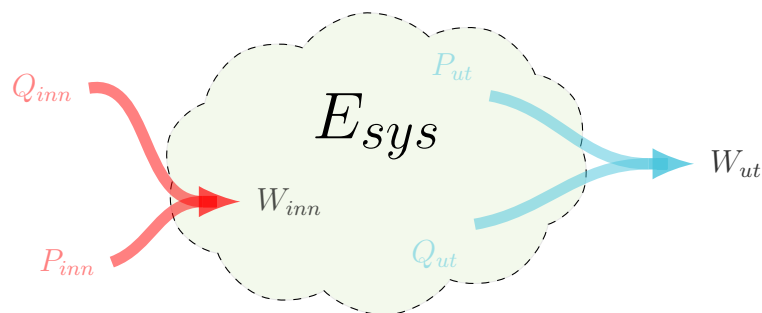
$$\dot{x}_2 = \frac{1}{x_1} [w_{i1}(b_{1F} - x_2) + w_{i2}(b_{2F} - x_2)]. \quad (3.5b)$$

Fasit: Svar på tidligere spørsmål:

1. Det er mange kandidater for prosessvariabler (og tilhørende målinger): for væskemengden er aktuelle kandidater massen m (veie tanken), høyden h eller volumet V (måle trykket i bunnen); for blandingsforholdet, så er blandingsforholdet b_F kanskje mulig (men hvordan måle det?), eventuelt kan man bruke tettheten ρ hvis denne endrer seg med blandingen (f.eks. via vekt- og volummålinger).
2. Pådragene er relatert til eventuelle reguleringsventiler eller pumper ved inn eller utstrømmene, og kanskje også konsentrasjonene inn. Når det gjelder ventilene, kan vi dermed se på pådraget som Strømningen w ut (evt. ventil åpningen l (evt. evt. spenning tilført reguleringsventilen)).
3. Forstyrrelser kan være blandingsforhold eller en inn-/utstrøm uten måling.
4. Nei, det er det trolig ikke, siden vi blander to stoff med muligens vidt forskjellige tettheter slik at massetettheten til blandingen kan variere.

3.1.2 Energibalanse ▶ aqdA6OofXoE

Energibalanse er egentlig akkurat som massebalanse, bare at vi bytter akkumulert masse i en kontrollvolum (nå også kalt et *lukket system*) med «akkumulert energi», samt bytter massestrømmer med «energi-strømmer» slik som vist i figur 3.2.



Figur 3.2: Illustrasjon av kontrollvolum (KV) for energibalanse: Energien «lagret» i kontrollvolumet (systemet) ved et gitt tidspunkt, $E_{sys}(t)$, har en endringsrate tilsvarende differansen mellom ratene av energi inn og ut.

Energibalanse følger direkte fra termodynamikkens lover (se, f.eks., [SNL](#)):

- **Termodynamikkens 1. hovedsetning:** Energi kan ikke forsvinne, men bare gå over fra en form til en annen.
- **Termodynamikkens 2. hovedsetning:** Overføring av varme skjer fra et sted med høyere temperatur til et sted med lavere temperatur («entropi øker med stor sannsynlighet»).

Vi kan bruke energibalanse til å utlede dynamiske modeller ved hjelp av følgende:

Energi-balanse: For et lukket system (SYS) med et gitt kontrollvolum (KV) har man

$$\underbrace{\frac{dE_{sys}}{dt}}_{\text{Endringsraten til et systems energi}} = \underbrace{W_{inn}}_{\text{rate av energi (altså effekt) inn i systemet}} - \underbrace{W_{ut}}_{\text{rate av energi ut av systemet}} \quad (3.6)$$

I boksen over betegner da E_{sys} energien (målt i Joule= $J=kgm^2/s^2$) «lagret» i kontrollvolumet (se fig. 3.2), mens W_{inn} og W_{ut} betegner energien som er tilført kontrollvolumet (det lukkede systemet) per sekund (målt i Watt = $W=J/s$).

For eksempel, hvis kontrollvolumet består av et medium med en konstant masse og uniform (altså likt fordelt) temperatur T og konstant **spesifikke varmekapasitet** C [$J K^{-1} kg^{-1}$], så vil en endring i systemets varmeenergi, $\Delta E_{sys} = E_{sys} - E_0$, være relatert til en endring i dets temperatur, $\Delta T_{sys} = T_{sys} - T_0$, via

$$\Delta E_{sys} = C_{sys} \cdot \Delta T_{sys} \quad (3.7)$$

hvor C_{sys} [J/K] er systemets spesifikke varmekapasitet skalert med massen i systemet, altså $\Delta E_{sys} = C \cdot m \cdot \Delta T_{sys}$ hvor m er massen i kilogram. Siden E_0 og T_0 er konstant, gir dette

$$\dot{E}_{sys} = C_{sys} \cdot \dot{T}_{sys}.$$

For vår del vil vi ofte bruke følgende form:

$$\frac{d}{dt} E_{sys} = P_{inn} + Q_{inn} - P_{ut} - Q_{ut} \quad (\text{Energibalanse})$$

hvor

- P_{inn} og P_{ut} er henholdsvis raten av arbeid utført på volumet og av volumet;
- Q_{inn} og Q_{ut} er henholdsvis varmestrømmer inn og ut av kontrollvolumet.

Dette er altså energibalansen for systemet/ kontrollvolumet. Arbeidet representerer en overgang fra og til mekanisk energi. Man kan også definere f.eks. P_{inn} som tilført elektrisk energi som utløser friksjonsvarme i en motstand. Alternativt kan man bare definere denne friksjonsvarmen som tilført varme Q_{inn} .

Relevante størrelser:

Mekanisk effekt (arbeid per tidsenhet):

- $P_{translasjon} = F \cdot v$ hvor F er kraft og v er hastighet;
- $P_{rotasjon} = \tau \cdot \omega$ hvor τ er et dreiemoment ω en vinkehastighet;
- $P_{trykk} = p \cdot q$ hvor p er et trykk og q en volumstrøm.

Elektrisk effekt:

- $P_{elektrisk} = U \cdot I$ hvor U er spenning og I er strøm.

Varmeoverføring og Newtons kjølelov

Alternative kilder: [YouTube-video](#); [Brian Douglas video](#); [Wikipedia](#).

Energi i form av varme kan overføres på tre måter, nemlig via [konveksjon](#), [stråling](#), og [konduksjon/varmledning](#). Vi skal hovedsakelig fokusere på sistnevnte.

Varmeoverføringen/-ledning mellom kalde til varme deler av et objekt, samt fra et medium til et annet kan beskrives av [Fouriers lov](#)³. Vi skal se på en enkel variant av denne loven som ofte antas i forbindelse med energy-balanse, nemlig Newtons avkjølingslov:

Newtons avkjølingslov: Varmeoverføring mellom to medium er direkte proporsjonal med temperaturforskjellen:

$$Q = h \cdot A \cdot \Delta T$$

hvor

ΔT : temperaturredifferansen [K].

Q : effekten tilsvarende varmeoverføringen [W] ;

h : [varmeovergangskoeffisienten](#) [W/m²K];

A : arealet mellom mediumene [m²];

Eksempel 3.3. Varmeskap:

Et varmeskap (VS) varmes opp elektrisk gjennom et varmeelement; se figuren til høyre. Varme lagres på grunn av skapets varmekapasitet. I tillegg kommer et varmetap til omgivelsene.

En naturlig kandidat for kontrollvolum er naturlig nok innsiden av varmeskapet. Energibalansen gir dermed:

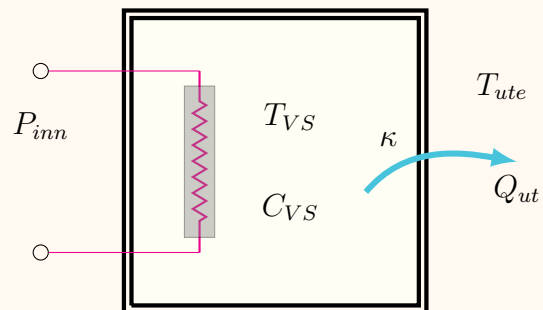
$$\frac{d}{dt} E_{VS} = P_{inn} - Q_{ut}.$$

La T_{VS} betegne temperaturen i varmeskapet, og la T_{ute} betegne temperaturen utenfor, som vi antar holder seg tilnærmet konstant. Fra Newtons kjølelov har vi dermed

$$Q_{ut} = \kappa \cdot (T_{VS} - T_{ute}),$$

hvor $\kappa = h \cdot A$ der h er varmeovergangskoeffisienten og A er skapets overflateareal.

Siden $E_{VS} = C_{VS}T_{VS}$, hvor C_{VS} er varmekapasiteten til mediumet i varmeskapet, har vi



³Tilsvarende Navier-Stokes-ligningen for væskeflyt, har man en kjent partiell-differensialligning for varmeoverføring, nemlig [varmeligningen](#).

dermed følgende fra energibalansen:

$$\dot{E}_{VS} = C_{VS}\dot{T}_{VS} = P_{inn} - \kappa \cdot (T_{VS} - T_{ute}) \implies \dot{T}_{VS} = \frac{1}{C_{VS}} [P_{inn} - \kappa(T_{VS} - T_{ute})].$$

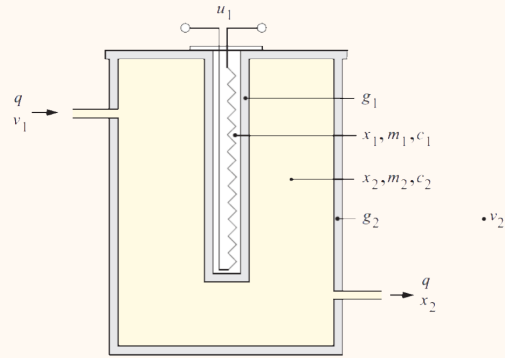
Spørsmål: Hva er naturlige tilstander, utganger, pådrag og forstyrrelser? Og hva er tilstandsromformen?

La oss ta et eksempelet som er hakket mer komplisert.

Eksempel 3.4. Varmtvannsbereder: (Eks. og figur fra 2.1 i [Balchen et al., 2016])

Gitt varmtvannsberederen vist i figuren til høyre, hvor

- u_1 : elektrisk effekt tilført kolben [kW];
- q : væskestrøm [m^3/s];
- x_i, v_i : temperaturer [K]
- m_i : masser [kg]
- c_i : spesifikk varmekapasitet [kJ/kgK];
- g_i : Varmeovergangstallet [kW/K];
- ρ : massetettheten til vannet [kg/m^3].



Antagelse: lik varmfordeling i tanken.

La oss først se på energibalanse for varmekolben. Tilført energi kommer fra pådraget, u_1 , mens vi antar at et “tap” av energi til tanken er proporsjonalt med temperaturforskjellen, altså tapet er $g_1(x_2 - x_1)$. Dette gir os følgende differensialligning for energi-endringen i varmekolben:

$$\dot{E}_1 = u_1 - g_1(x_1 - x_2).$$

Her har vi igjen fra Newtons avkjølingslov at $\dot{E}_1 = m_1 c_1 \dot{x}_1$, slik at temperaturendringen i varmekolben er gitt ved

$$\dot{x}_1 = \frac{1}{m_1 c_1} (u_1 - g_1(x_1 - x_2)).$$

For tanken kan energi flyte til og fra varmekolben og ut av tanken, samt tilføres fra det kalde vannet fra innløpet og tappes fra utløpet. Dette gir oss følgende energibalanse:

$$\dot{E}_2 = g_1(x_1 - x_2) - g_2(x_2 - v_2) + q\rho(c_2 v_1 - c_2 x_2).$$

Tilsvarende varmekolben, har vi at energiendringen i tanken er proporsjonal med temperaturøkningen: $\dot{E}_2 = m_2 c_2 \dot{x}_2$, og dermed

$$\dot{x}_2 = \frac{1}{m_2 c_2} (g_1(x_1 - x_2) - g_2(x_2 - v_2) + q\rho c_2 (v_1 - x_2)).$$

3.2. Elektro-mekaniske systemer

Hvorfor skal du lære dette? Som navnet tilsier, så er **elektromekaniske systemer** bygd opp av mekaniske og/eller elektriske komponenter. Eksempler på slike systemer inkluderer roboter og elektriske kjøretøy. Nuff said...

3.2.1 Newtons andre lov (kraftbalanse)

▶ 7YSuW5DKuBg

Alternative kilder: [Wikipedia](#); [SNL](#); En hvilken som helst fysikkbok.

Newtons andres lov gir en enkel formel for å beskrive dynamikken til én enkelt punktmasse (altså et objekt hvor man antar at all dens masse er lokalisert på ett enkelt punkt). Denne loven kan tenkes på som «kraft-balanse», som igjen egentlig stammer fra en bevaringslov, nemlig bevaring av bevegelsesmengde.

Newtons andre lov: Endringen i *bevegelsesmengden*, \vec{p} , til en punktmasse, er lik summen av kreftene, $\sum_i \vec{F}_i = \vec{F}_1 + \vec{F}_2 + \dots$, som virker på objektet: $\frac{d\vec{p}}{dt} = \sum_i \vec{F}_i$. Når massen, m , til objektet ikke endrer seg over tid (hvis den kan endre seg, se [denne](#)), tilsvarer dette:

$$m\vec{a} = m \frac{d\vec{v}}{dt} = \sum_i \vec{F}_i$$

hvor \vec{v} er objektets hastighet og $\vec{a} = \frac{d\vec{v}}{dt}$ dets akselerasjon.^a

^a**Merk:** \vec{p} , \vec{v} og \vec{F}_i er her normalt sett vektorer i \mathbb{R}^k for $k \in \{1, 2, 3\}$.

For å utlede bevegelsesligningene til flere enkle mekaniske systemer så er følgende størrelser høyst relevante:

Relevante størrelser:

Lineær fjær: $F_k = k \cdot \Delta p$ (Hookes lov) og $E_k = \frac{1}{2}k\Delta p^2$, hvor k [N m^{-1}] er fjærkonstanten og Δp [m] er komprimeringen av fjæren fra utgangsposisjonen.

Lineær demper: $F_d = d \cdot v$ hvor d [N s m^{-1}] er dempningskonstanten og v [m/s] er komprimeringshastigheten til demperen.

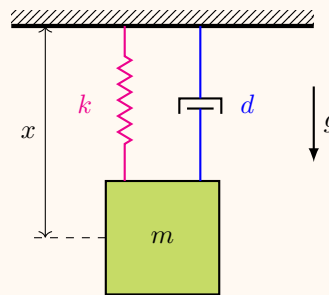
Kinetisk energi: $\mathcal{K} = \frac{1}{2}m \cdot v^2$ hvor m [kg] er massen og v [m/s] er hastigheten til et legeme.

Potensiell energi: $\mathcal{P} = m \cdot g \cdot h$ hvor m [kg] er massen, g ($\approx 9.81 \text{ m/s}^2$) er akselerasjonen fra tyngdekraften, og h [m] høyden et legeme fra et referansepunkt.

La oss ta en titt på en meget klassisk eksempel, nemlig et masse-fjær-demper system:

Eksempel 3.5. Hengende masse-fjær-demper:

Et hengende masse-fjære-demper system er vist i figuren til høyre. Det består av en hengende masse (en boks) som er festet til et tak via en lineær fjær og en lineær demper. Boksen har masse m [kg], fjæren har fjærkonstant k [N/m] og demperen har dempningskonstant d [Ns/m]. Boksens avstand målet fra taket betegnes med x [m]. Utgangsposisjonen til fjæren tilsvarer posisjonen $x = x_*$, det vil si posisjonen hvor systemet ville vært i likevekt hvis det ikke ble påvirket av akselerasjonen fra tyngdekraften, g .



Ved å bruke uttrykket for en lineær fjær gitt over (Hookes lov), så vet vi at kraften som virker på kassen fra fjæren er $F_k = -k \cdot (x - x_*)$, mens kraften fra demperen er $F_d = -d \cdot \dot{x}$. Vi må også ta med tyngdekraften, $G = mg$, hvor g ($\approx 9.81 \text{ m/s}^2$) er akselerasjonen fra tyngdekraften, siden kassens posisjon x ikke er målt relativt til fjærens utgangsposisjon.

Kraftbalanse fra Newtons andre lov gir dermed

$$ma = m\ddot{x} = \sum_i F_i = F_k + F_d + G = -k(x - x_*) - d\dot{x} + mg.$$

Anta at $m = 2$, $k = 10$ og $d = 3$, samt $g = 10$ og $x_* = 1$. Hva er da likevektspunktet til systemet? For å svare på dette, la oss først ta $x_1 = x$ og $x_2 = \dot{x}$ slik at vi kan skrive systemet om på tilstandsromform:

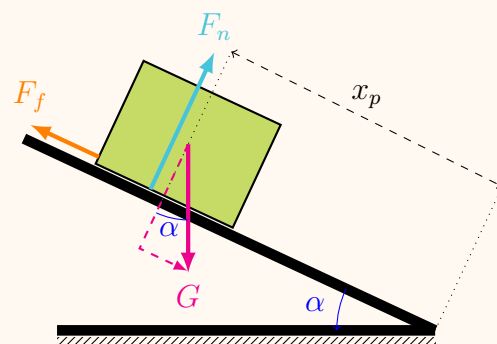
$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= -k \cdot (x_1 - x_*) - d \cdot x_2 + mg. \end{aligned}$$

Likevektspunkter tilsvarer punkter (\bar{x}_1, \bar{x}_2) slik at $\dot{x}_1 = \dot{x}_2 = 0$. Dermed har vi $\bar{x}_2 = 0$, mens fra den andre finner vi dermed at \bar{x}_1 må tilfredsstille

$$-10 \cdot (\bar{x}_1 - 1) + 2 \cdot 10 = 0 \implies \bar{x}_1 = \frac{20 + 10}{10} = 3.$$

Eksempel 3.6. Kasse på en skråning:

Gitt en kasse med vekt m [kg] som sklir ned en skråning med konstant helning α (se figur til høyre). La x_p betegne posisjonen til boksen målt parallelt med den skrå bakken, og la $x_h = \dot{x}_p$ betegne dens hastighet. Tilstandene er dermed x_p og x_h .



Anta at kassen påvirkes av kinetisk friksjon, gitt ved^a $F_f = \mu \cdot F_n$, hvor F_n er normalkraften mellom bakken og kassen, og hvor μ er friksjonskoeffisienten. Fra Newtons andre lov har vi dermed

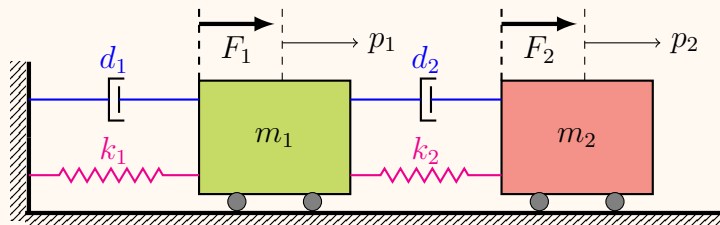
$$m\dot{x}_h = G_{\parallel} - F_f = G_{\parallel} - \mu F_n,$$

hvor $G_{\parallel} = G \sin(\alpha) = mg \sin(\alpha)$ er tyngkraften som virker på kassen og $g = 9.81 \text{ m s}^{-2}$ er gravitasjons akselerasjonen. Siden normalkraften er lik $F_n = |G_{\perp}| = mg \cos(\alpha)$, er de dynamiske ligningen for systemet som følger:

$$\begin{aligned} \dot{x}_p &= x_h \\ \dot{x}_h &= g [\sin(\alpha) - \mu \cos(\alpha)]. \end{aligned}$$

^aMerk at vi også bare kan ha $|F_f| \leq |G_{\parallel}|$, eller vil jo friksjonen kunne få kassen til å bevege seg oppover!

Eksempel 3.7. Koblede masse-fjær-demper systemer: Gitt to koblede masse-fjære-demper systemer som vist i figuren under.



Trallene har henholdsvis masse lik m_1 og m_2 , og påføres hver sin kraft ([N]), F_1 og F_2 . Fjærene, som tilfredsstiller Hookes lov med fjærkonstanter k_1 og k_2 , er i hvileposisjon for når henholdsvis $p_1 = 0$ og $p_1 - p_2 = 0$.

La $h_i = \dot{p}_i$ og $a_i = \dot{v}_i$. Kraftbalanse gir:

$$\begin{aligned} m_1 a_1 &= F_{k_1} + F_{d_1} + F_1 - F_{k_2} - F_{d_2} \\ m_2 a_2 &= F_2 + F_{k_2} + F_{d_2} \end{aligned}$$

Her er

$$\begin{aligned} F_{k_1} &= -k_1 p_1 \\ F_{d_1} &= -d_1 h_1 \\ F_{k_2} &= k_2 (p_1 - p_2) \\ F_{d_2} &= d_2 (h_1 - h_2) \end{aligned}$$

Ved å ta $(x_1, x_2, x_3, x_4) = (p_1, p_2, h_1, h_2)$ og $(u_1, u_2) = (F_1, F_2)$ kan vi bruke *matriseformen*:

$$\underbrace{\frac{d}{dt} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}}_{\dot{x}} = \underbrace{\begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -\frac{(k_1+k_2)}{m_1} & \frac{k_2}{m_1} & -\frac{(d_1+d_2)}{m_1} & \frac{d_2}{m_1} \\ \frac{k_2}{m_2} & -\frac{k_2}{m_2} & \frac{d_1}{m_2} & -\frac{d_2}{m_2} \end{bmatrix}}_{Ax} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} + \underbrace{\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ \frac{1}{m_1} & 0 \\ 0 & \frac{1}{m_2} \end{bmatrix}}_{Bu} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}.$$

3.2.2 Rotasjonsdynamikk (momentbalanse)



Alternative kilder: [Wikipedia](#).

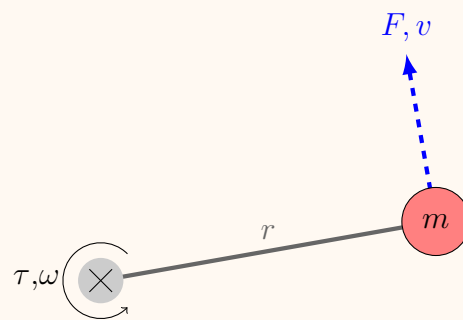
Hva med roterende legemer? Tilsvarende hvordan Newtons andre lov i tar for seg bevaring av et objekts bevegelsesmengde, kan man også se på bevaringen av vinkelmoment for roterende objekter. Mer spesifikt, ved å introdusere det roterende legemets *treghetsmoment* (tenk dets «roterende masse»), leder Newtons andre lov til Eulers rotasjonsligning⁴ (se [Wikipedia](#)).

Før vi hopper til denne ligningen, så starter vi med et eksempel hvor vi introduserer viktige størrelser.

Eksempel 3.8. Roterende stang:

Vi skal nå se hvordan man kan relatere kraft og hastighet til henholdsvis dreiemoment og vinkelhastighet.

Gitt systemet vist i figuren til høyre, bestående av en masse som er festet til enden av en stang. Stangen kan fritt rotere om en aksling. Anta at massen m [kg] er plassert ytterst på stangen/armen som har lengde r [m] (vi antar at stangen selv ikke har noen vekt).



Kraften F [N] som virker på massen (vinkelrett i forhold til stangen) genererer et **dreiemoment**, τ [N m] (målt positivt mot klokken), gitt ved

$$\tau = r \cdot F.$$

Hvis massen beveger seg med en hastighet v [m/s] som vist i figuren, så er tilsvarende **vinkelhastighet**, ω [rad/second], gitt ved

$$\omega = \frac{v}{r},$$

eller tilsvarende $v = r \cdot \omega$. Hvis lengden, r , til armen er konstant har vi dermed

$$\dot{v} = r \cdot \dot{\omega}.$$

Fra Newtons 2. lov vet vi jo at $m\dot{v} = F$ (hvis F er summen av kreftene som virker på massen). Ved å sette inn for \dot{v} og F får vi

$$m \underbrace{r \cdot \dot{\omega}}_{=\dot{v}} = \underbrace{\frac{\tau}{r}}_{=F} \implies J\dot{\omega} = \tau,$$

hvor

$$J = m \cdot r^2 \text{ [kg m}^2\text{]}$$

⁴Denne ligningene predikere faktisk den funky oppførselen du ser i <https://youtu.be/1n-HMSCDYtM>.

er **treghetsmomentet** (også kalt kraftmoment) til systemet/legemet. Den kinetiske energien til systemet er dermed

$$\mathcal{K} = \frac{1}{2}mv^2 = \frac{1}{2}m(r\omega)^2 = \frac{1}{2}(mr^2)\omega^2 = \frac{1}{2}J\omega^2.$$

Man kan derfor tenke på treghetsmomentet J som «massen» ved rotasjonsbevegelser, altså tilsvarende som m brukes for translasjonsbevegelser.

La oss nå ta det i mer detalj, ved å se på Eulers ligning (merk at denne nok er skrevet i overkant presist for dette faget per nå):

Eulers ligning for roterende objekter (momentbalanse): La $\vec{\omega} \in \mathbb{R}^3$ betegne rotasjonshastighet til et stivt legeme, og la $\mathbf{J} \in \mathbb{R}^{3 \times 3}$ være dets treghetsmoment-matrise^a målt relativ til en (konstant) verdensramme. Da

$$\mathbf{J} \frac{d\vec{\omega}}{dt} = \vec{\tau}, \quad (3.8)$$

hvor $\vec{\tau}$ er summen av eksterne dreiemoment som virker på objektet.

Rotasjon om én akse: Hvis legemet kun roterer om én akse, har man

$$J\dot{\omega} = \sum_i \tau_i, \quad (3.9)$$

hvor J er legemets treghetsmoment om denne aksen, τ_i er eksterne dreiemoment som virker om rotasjonsaksen, mens ω er vinkelhastigheten tilsvarende rotasjonen.

^aMerk: treghetsmoment-matrisen gitt i en verdensramme vil generelt sett ikke være konstant, og vil i stedet avhenge av objektets orientasjon.

Fun facts, bemerkninger og annet dill dall (you may skip)

Uten tyngdekraft kan spesielle treghetsmoment føre til ganske funky oppførsel, se <https://www.youtube.com/watch?v=1n-HMSCDYtM>

Relevante størrelser: (holder for små $\Delta\theta$, hvor r [m] er radius/arm fra rotasjonsaksen).^a

Lineær fjær: $\tau_f \approx k_f \cdot r^2 \Delta\theta$ hvor k_f [N m⁻¹] er fjærkonstanten og $\Delta\theta$ [m] er komprimeringen av fjæren fra en referansesvinkel i radianer.

Lineær torsjon: $\tau_t = k_t \cdot \Delta\theta$ hvor k_t [N m rad⁻¹] er torsjonskonstanten og $\Delta\theta$ [m] er rotasjonen fra et hvilepunkt.

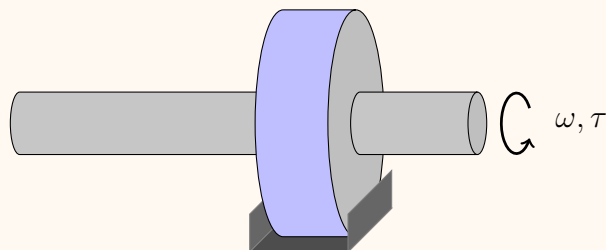
Lineær demper: $\tau_d = d \cdot r^2 \omega$ hvor d [N s m⁻¹] er dempningskonstanten og ω [m/s] er vinkelkomprimeringshastigheten til demperen.

Kinetisk energi: $\mathcal{K}_r = \frac{1}{2}J \cdot \omega^2$ hvor J [kg m²] er treghetsmomentet og ω [rad/s] er vinkelhastighet til det roterende legemet.

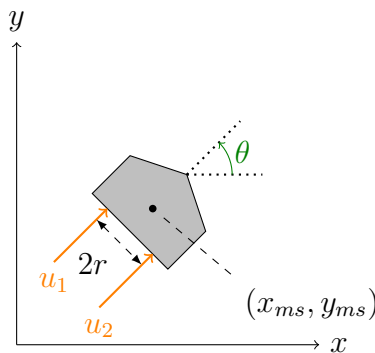
^aBruker at et utslag $\Delta p = r \sin(\Delta\theta) \approx r\Delta\theta$ for små $\Delta\theta$.

Eksempel 3.9. Svinghjul med friksjon:

Gitt et svinghjul med treghetsmoment J . Anta at mekanismen er utsatt for viskøs friksjon proporsjonal med vinkelhastigheten ω , altså $\tau_f = d\omega r$ hvor d er dempningskonstanten og r er radiusen hjulet. Anta videre at pådraget vårt (eller forstyrrelsen) er et påført dreiemoment τ . Eulers ligning (momentbalanse) gir dermed



$$J\dot{\omega} = \tau - d\omega r.$$



Figur 3.3: Hovercraft i planet.

Oppgave 3.2. Hovercraft: Figur 3.3 viser et hovercraft sett ovenfra. La (x_{ms}, y_{ms}) betegne posisjonen til massesenteret til farkosten: La dens masse være m og la treghetsmomentet om massesenteret være J . Retningsvinkelen θ (gitt i radianer) er målt positiv mot klokken fra x -aksen.

Det er to store vifter som kan brukes til å styre farkosten. Disse har samme avstand til massesenteret, og det $2r$ meter mellom dem. Vi bruker u_1 og u_2 til å betegne kraften fra disse viftene som virker på hovercraften (se figuren).

Finn de dynamiske bevegelsesligningene til dette systemet og skriv på dem på tilstandsromform, altså som et sett av førsteordens differensialligninger.

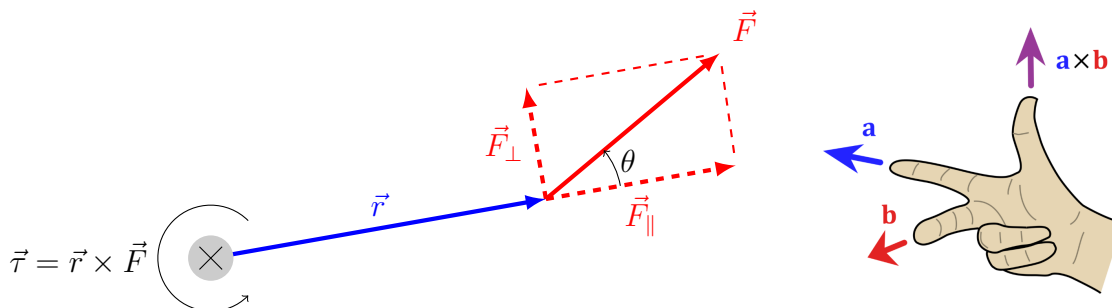
Dreiemoment og treghetsmoment

La oss se litt nærmere på størrelsene dreiemoment og treghetsmoment.

Dreiemomentet (og kalt kraftmomentet) $\vec{\tau}$ (SI-enheten er N m) er lik kryssproduktet av armen \vec{r} og kraften \vec{F} altså (se også figur 3.4)

$$\vec{\tau} = \vec{r} \times \vec{F}.$$

For å regne ut dreiemomentet, bør man derfor generelt sett tenke på alle størrelser som vektorer i rommet \mathbb{R}^3 . Ofte kan dette dog forenkles:



Figur 3.4: **Venstre:** En kraft \vec{F} påføres en stang en avstand \vec{r} fra stangens feste. Kraften fører til et dreiemoment $\vec{\tau} = \vec{r} \times \vec{F} = \|\vec{r}\| \|\vec{F}_\perp\| \vec{n} = \|\vec{r}\| \|\vec{F}\| \sin(\theta) \vec{n}$ om festet, hvor \vec{n} er en enhetsvektor som er normal på både \vec{r} og \vec{F} . **Høyre:** Høyrehåndsregelen for utregning av kryssprodukt; figur fra SNL.

Eksempel 3.10. Gitt en stang av lengde r slik som vist i figur 3.4. Det virker der en kraft \vec{F} vinkelrett på enden av stangen, slik at $\vec{F}_\parallel = 0$. Mer spesifikt, la

$$\vec{r} = \begin{bmatrix} r \\ 0 \\ 0 \end{bmatrix} \quad \text{og} \quad \vec{F} = \begin{bmatrix} 0 \\ F \\ 0 \end{bmatrix}$$

og dermed

$$\vec{\tau} = \vec{r} \times \vec{F} = \begin{bmatrix} 0 \\ 0 \\ r \cdot F \end{bmatrix}.$$

Det virker derfor et dreiemoment om enhetsvektoren $\vec{n} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$ med magnitudo $\tau = r \cdot F$.

Trehetsmomentet⁵ J (SI-enheten er kg m^2) kan sees på som det som tilsvarende til masse for rotasjoner. La oss fort vise for denne kvantiteten kommer fra:

Gitt en stang, som den i rødt i figur 3.4, som er $l = \|\vec{r}\|$ lang. Den totale massen til stangen er m , men denne er ikke jevnt fordelt; i stedet kan vi tenke oss at massen til systemet er bygget opp av en begrenset mengde av massepartikler, hver med masse m_i en avstand r_i fra rotasjonsaksen. Dreiemomentet er da gitt ved

$$J = \sum_i r_i \cdot m_i.$$

Enda mer presist: la massen til et hvert tversnitt langs stangen være gitt av $\rho(\cdot)$, slik at den totale massen er $m = \int_0^l \rho(r) dr$. Anta nå at stagen roterer om sitt feste med en

⁵Du finner en liste med trehetsmomentene til flere vanlige former på [Wikipedia](#).

vinkelhastighet ω [rad s⁻¹]. Hastigheten, $v_p(r)$ [m s⁻¹], til et punkt på stagen som er en avstand r fra rotasjonsaksen er $v_p(r) = r \cdot \omega$. Den kinetiske energien tilsvarende dette punktet er dermed $E_p(r) = \frac{1}{2} \rho(r) v_p^2(r)$, slik at den totale kinetiske energien til stangen er

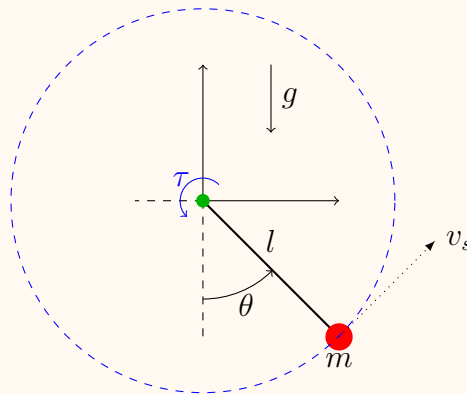
$$E = \int_0^l \frac{1}{2} \rho(r) v_p^2(r) dr = \frac{1}{2} \left[\int_0^l \rho(r) r^2 dr \right] \omega^2 =: \frac{1}{2} J \omega^2.$$

Altså er treghetsmomentet til legemet:

$$J := \int_0^l \rho(r) r^2 dr.$$

Eksempel: Anta at en stang av lengde l har jevn (og dermed konstant) massefordeling, slik at $\rho(r) = (m/l)$. Vi finner da treghetsmomentet til å være $J = \int_0^l \rho(r) r^2 dr = \int_0^l (m/l) r^2 dr = (m/l) \int_0^l r^2 dr = \frac{1}{3} ml^2$.

Eksempel 3.11. (En enkel pendel) Vi skal nå utlede den dynamiske ligningen til den enkle pendelen vist i figuren under.



Metode 1 – Eulers ligning: Anta at all massen til pendelen er lokalisert om et enkelt punkt på enden (merket rødt i figuren). Med andre ord antar vi at stangen ikke har masse. Siden stangen ikke har noen masse, har vi fra [parallellakseteoremet](#) at treghetsmomentet til pendelen er målt om rotasjonssenteret (se den grønne sirkelen) gitt ved $J = ml^2$ (se også [denne listen](#)).

Husk at [dreiemoment](#) er kryssproduktet av kraft og arm. Den eneste eksterne kraften som virker på pendel er tyngdekraften, gitt ved $\vec{G} = [0, -mg, 0]^T$. Armen fra rotasjonssenteret til pendelens massesenter er $\vec{r} = l[\sin(\theta), -\cos(\theta), 0]^T$. Dermed er dreiemoment forårsaket av tyngdekraften

$$\vec{\tau} = \vec{r} \times \vec{G} = [0, 0, -mgl \sin(\theta)].$$

Siden vinkelhastigheten til pendellen er $\vec{\omega} = [0, 0, \dot{\theta}]^T$, har vi fra Eulers ligning at

$$\ddot{\theta} = -\frac{g}{l} \sin(\theta).$$

Metode 2 – Newtons andre lov: La oss vise at Newtons andre lov resulterer i den samme bevegelsesligningen. Vi begynner ved å legge merke til at enden av pendelen bare kan bevege seg langs den blå, stiplede sirkelen vist i figuren. Vi vet derfor at endringen i hastigheten, $v_s = l\dot{\theta}$, til pendelen langs denne sirkelen er gitt av Newtons andre lov, hvor den eneste kraften som virker på den er tyngdekraften. Vi har dermed at

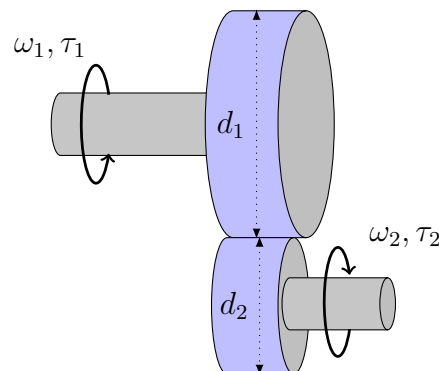
$$m \frac{d}{dt} v_s = (-mg \sin(\theta)) \implies \ddot{\theta} = -\frac{g}{l} \sin(\theta).$$

Her har vi brukt at hastigheten til enden av pendelen er $\vec{v}_p = v_s [\cos(\theta), \sin(\theta)]^\top$, slik at effekten av tyngdekraften på endringen til v_s er gitt av det indre produktet $\vec{G} \cdot \vec{v}_p / v_s = -mg \sin(\theta)$.

3.2.3 Gir (ideelle) ▶ wb0MWCU8t6s

Alternative kilder: [Wikipedia](#).

Hvorfor skal du lære dette? Gir er viktige komponenter i mange mekaniske og elektromekaniske systemer. Det er derfor viktig å kunne modellere disse, vite hensikten med, samt å ha en intuitiv forståelse av den grunnleggende virkemåten.



Figur 3.5: Illustrasjon av et ideelt gir, bestående av to (tann-)hjul med forskjellig diamenter.

Det finnes flere måter å implementere gir på, for eksempel vha. tannhjul. I virkeligheten fører dette til mange tap fra f.eks. friksjon og merkelige ulineariteter som «backlash». Vi skal dog holde oss til *ideelle* (tapsfrie) gir.

Hensikten med et gir er som regel å enten øke/reducere rotasjonshastigheten og/eller redusere/øke dreiemomentet. Den viktigste intuisjonen her er at for eksempel en økning i rotasjonshastigheten over giret fører til en reduksjon i dreiemomentet, og vice versa:

Giroverføring av vinkelhastighet og dreiemoment:

Gitt et ideelt gir som vist i figur 3.5. Anta at $n = \frac{d_1}{d_2} = \frac{r_1}{r_2}$, hvor r_1 og r_2 er positive heltall.^a

Vinkelhastighet: Vi har at hastighetene på enden av (tann-)hjulene er $v_1 = r_1\omega_1$ og $v_2 = r_2\omega_2$. Siden vi antar ideelle gir, hvor det ikke er noen form for tap (de glipper blant

annet ikke), så må vi ha $v_1 = v_2$, som igjen betyr $\omega_2 r_2 = \omega_1 r_1$, og dermed

$$\omega_2 = n\omega_1.$$

Dreiemoment: Anta at ω_1 holder seg konstant. Kreftene på enden av (tann-)hjulene er da gitt ved $F_1 = \tau_1/r_1$ og $F_2 = \tau_2/r_2$. På grunn av antagelsen om et ideelt gir, så har vi fra Newtons tredje lov at $F_1 = F_2$. Dermed $\tau/r_2 = \tau/r_1$, som igjen betyr

$$\tau_2 = \tau_1/n.$$

Konklusjon: for en girutveksling $n \geq 1$ (skriver ofte $1 : n$) så vil vinkelhastigheten (fra ω_1 til ω_2) øke, mens dreiemomentet (fra τ_1 til τ_2) vil synke.

^aGir er som regel basert på tannhjul. Man kan derfor tenke på r_1 og r_2 som antall tenner på tannhjulene.

Men skjer hvis systemet akselererer? La oss nå anta at vi har et ideelt slik som i figur 3.5, hvor $\omega_2 = n\omega_1$. Det virker et dreiemoment τ_1 på det venstre hjulet, og et dreiemoment fra en last, τ_l , på det høyre hjulet. Momentbalanse for det første hjulet resulterer i

$$J_1 \dot{\omega}_1 = \tau_1 - \hat{\tau}$$

hvor $\hat{\tau}$ er et dreiemoment som tilsvarende alt som skjer på høyresiden av giret pga. av (den **holonomske**) begrensningen $\omega_2 = n\omega_1$. Tilsvarende gir momentbalansen på høyresiden

$$J_2 \dot{\omega}_2 = \tau_2 - \tau_l = \hat{\tau}/n - \tau_l,$$

hvor da $\tau_2 = \hat{\tau}/n$. Siden $\dot{\omega}_2 = n\dot{\omega}_1$, så har vi derfor at

$$J_2 n \dot{\omega}_1 = \hat{\tau}/n - \tau_l \implies \hat{\tau} = n(J_2 n \dot{\omega}_1 + \tau_l).$$

Ved å sette dette inn i ligningen for venstresiden får vi

$$J_1 \dot{\omega}_1 = \tau_1 - n(J_2 n \dot{\omega}_1 + \tau_l) \implies (J_1 + n^2 J_2) \dot{\omega}_1 = \tau_1 - n\tau_l.$$

Fra dette kan vi også vise at

$$\left(\frac{J_1}{J_2} + n^2\right) \tau_2 = n\tau_1 + \frac{J_1}{J_2} \tau_l.$$

Hvis systemt ikke akselerer ($\dot{\omega}_1 = \dot{\omega}_2 = 0$) så får vi $\tau_2 = n\tau_1$ siden da $\tau_2 = \tau_l$.

Hvordan (og hvorfor) velge girutveksling? Det er ganske greit å tenke på en bil eller sykkel i denne sammenhengen:

- Tilstrekkelig stor trekraft og hastighet på utgangen (hjulene).
- Ikke for høyt turtall på motoren.
- Maksimal akselerasjon.

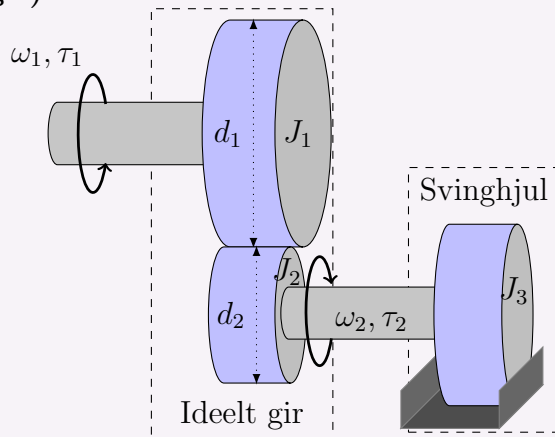
- Drivstofføkonomi.

Se også følgende: og [s3BsDF6UjCQ](#) og [pvE83NfbZ8g](#) .

I en bil kan man jo f.eks. skifte gir, og tilpasse det til situasjonen. I andre sammenhenger— ofte sånn for elektromotorer— bestemmer man girutvekslingen en gang for alle.

Oppgave 3.3. (Svinghjul med ideelt gir)

Et svinghjul med treghetsmoment J_3 skal styres ved hjelp av en motor (se figuren til høyre). Motoren generer et dreiemoment τ_1 , og er festet til en aksling som roterer med vinkelhastighet ω_1 . Akslingen er igjen festet til et ideelt gir med girutveksling $n = \frac{d_1}{d_2}$, hvor tannhjulene i giret har treghetsmoment henholdsvis lik J_1 og J_2 . Det virker et viskøst friksjonsmoment på svinghjulet lik $\tau_f = d_v \omega_2$.



Du skal: a) Vise at de dynamiske ligningene (på tilstandsromform) til systemet med ω_1 som tilstand (både ω_2 og τ_2 **ikke** skal dukke opp i ligningene) er

$$\dot{\omega}_1 = \frac{1}{J} [\tau_1 - d_v \cdot n^2 \cdot \omega_1]$$

hvor $J = J_1 + (J_2 + J_3) \cdot n^2$.

b) Anta at systemet er i ro ($\omega_1 = \omega_2 = 0$). Det er ønsket å starte det opp med maksimal vinkelakselerasjon $\dot{\omega}_2$ på svinghjulet. Hva skal girutvekslingen være for for å oppnå dette hvis $J_1 = 4$, $J_2 = 1$, $J_3 = 15$ og $d_v = 2$? Merk: dette trenger bare å holde i øyeblikket når $\omega_1 = 0$. ^a

^aHint: Husk at en funksjon $f(x)$ har et ekstremum (maksimum eller minimum) ved et punkt a hvis $f'(a) = 0$; girutvekslingen vil være på formen $n = a/b$ for noen positive heltall a og b (altså $2/3$, $6/2$, for å gi noen eksempler).

3.2.4 Krichhoffs lover for elektriske kretser

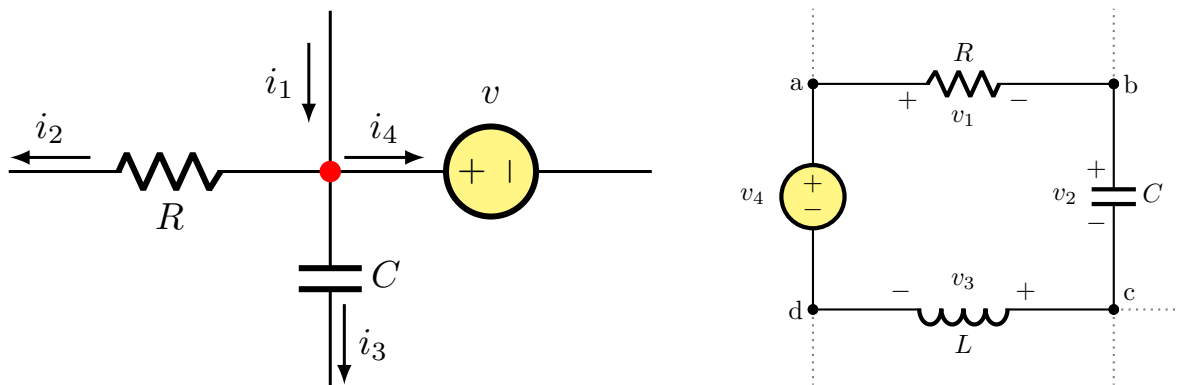


Alternative kilder: [Wikipedia](#).

Kirchhoffs lover er to grunnleggende prinsipper innen elektrisk kretsanalyse, oppkalt etter den tyske fysikeren Gustav Kirchhoff.

De to «lovene», strømløven (eng. - Kirchhoff's Current Law (KCL)) og spenningsloven (eng. Kirchhoff's Voltage Law (KVL)), er illustrert i henholdsvis **a**) og **b**) i figur 3.6.

Kirchhoffs strømløven: I et forgreningspunkt i en elektrisk krets med n foregreninger er summen av alle inngående strømmer lik summen av alle utgående strømmer: $\sum_{k=1}^n i_k = 0$.



a) Summen av strømmer inn og ute av den røde noden summerer til 0.

b) Summen av spenninger i den lukkede sløyfen a-b-c-d summerer til 0.

Figur 3.6: Illustrasjoner av Kirchhoffs lover for elektriske kretser.

Kirchhoffs spenningslov: Summen av spenninger i en lukket strømsløyfe med n komponenter er lik null: $\sum_{k=1}^n v_k = 0$.

Disse prinsippene er viktige verktøy innen elektrisk kretsanalyse og brukes for å beregne strømmer og spenninger i komplekse kretser. For våres del, som er ute etter å lage enkle matematiske modeller av dynamiske systemer vha. differensialligninger, er også følgende idealiserte relasjoner mellom strøm og spenning over basiskomponenter viktige:

La v betegne spenning i Volt [V] og i strøm i Ampere [A]. Vi har da følgende relasjoner:

- **Motstand (Ohms lov):** $v(t) = R \cdot i(t)$ hvor R er motstanden i Ohm [Ω].
- **Kondensator:** $i(t) = C \cdot \frac{dv(t)}{dt}$ hvor C er kapasitet/kapasitans i Farad [F].
- **Spole:** $v(t) = L \cdot \frac{di(t)}{dt}$ hvor L er induktansen i Henry [H].

Eksempel 3.12. RLC-krets: Vi skal nå utlede en differensialligning tilsvarende RLC-kretsen vist i figur 3.6-b). Fra kirchhoffs spenningslov har at

$$v_4 = v_1 + v_2 + v_3 = R \cdot i(t) + C \int_0^t i(\tau) d\tau + L \cdot \frac{di(t)}{dt}$$

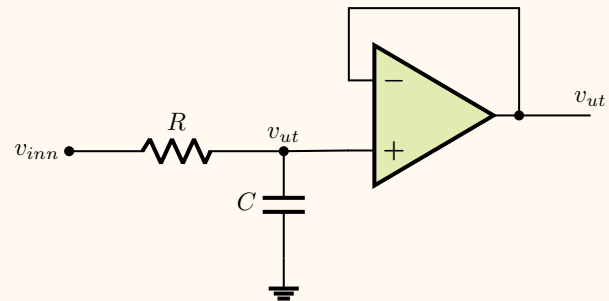
hvor $i(t)$ er strømmen gjennom kretsen.

Anta nå at spenningstilførslen v_4 holder seg konstant. Ved å da derivere med hensyn på tid på begge sider, og bruke at $\dot{v}_4 = 0$ (siden v_4 er konstant), så får vi følgende andre-ordens differensialligning for kretsen:

$$L \cdot \frac{d^2 i(t)}{dt^2} + R \frac{di(t)}{dt} + C \cdot i(t) = 0$$

Eksempel 3.13. Første-orden lavpassfilter fra RC-krets: Et lavpassfilter er en svært vanlig «komponent» med mange bruksområder, ikke minst innen reguleringsteknikken. Denne typen filter, som vi skal se nærmere på i § 12.2, brukes til å «glatte ut» et signal ved å filtrere vekk høyfrekvente deler av signalet.

Vi skal i dette eksempelet utlede differensialligningen tilsvarende et første-ordens lavpass filter gitt av kretsen til høyre. Dette er en RC-krets koblet til en såkalt spenningsfølger (implementert vha. en ideell operasjonsforsteker (opamp)). Spenningsfølgeren har som jobb å sørge for at spenningen v_{ut} er den samme uansett hva som kobles til etter spenningsfølgeren.



Vi trenger derfor bare finne hva spenningen v_{ut} over kondensatoren er. La oss denne gangen bruke strømloven i noden mellom motstanden og kondensatoren:

$$i_C(t) - i_R(t) = C \cdot \frac{dv_{ut}}{dt} - \frac{(v_{inn} - v_{ut})}{R} = 0 \quad \implies \quad \dot{v}_{ut} = -\frac{v_{ut}}{RC} + \frac{v_{inn}}{RC}.$$

Dette første-ordens lavpassfilteret tilsvarer derfor en første-ordens differensialligning på formen $\dot{y} = -ay + bu$ med $y = v_{ut}$, $u = v_{inn}$, og $a = b = 1/(RC)$.

3.2.5 *Euler–Lagrange-ligningene for stive legemer*

Gitt et system med n frithetsgrader, la

q_1, q_2, \dots, q_n betegne systemets *generelle koordinater*;

$\mathcal{K}(q, \dot{q})$ systemets kinetiske energi;

$\mathcal{P}(q)$ systemets potensielle energi;

$\mathcal{L}(q, \dot{q}) := \mathcal{K}(q, \dot{q}) - \mathcal{P}(q)$ er Lagrange-funksjonen;

$\tau_1, \tau_2, \dots, \tau_n$ generelle (ikke-konservative) krefter.

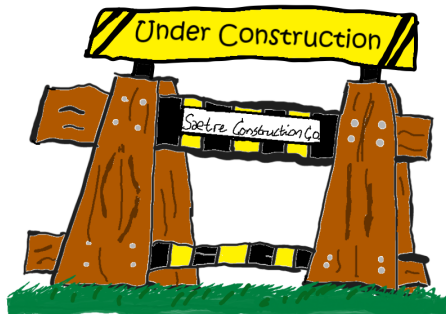
Systemets dynamikk er da gitt ved [Euler–Lagrange-ligningen](#):

Euler–Lagrange-ligningen:

$$\frac{d}{dt} \left(\frac{\partial \mathcal{L}}{\partial \dot{q}_i} \right) - \frac{\partial \mathcal{L}}{\partial q_i} = \tau_i, \quad i = 1, 2, \dots, n.$$

3.3. *Unlineariteter og fysiske fenomener*

Hysteres, stiction, dødbånd. s.149.



3.3.1 Metning og rate-begrensninger

Gitt et pådragsorgan u . Vi har da

- **Metning:** $u_{\min} \leq u \leq u_{\max}$
- **Ratebegrensning:** $\dot{u}_{\min} \leq \dot{u} \leq \dot{u}_{\max}$

På engelsk er metning “saturation”. Vil vil derfor bruke $\text{sat}(\cdot)$ til å betegne metningsfunksjonen:

$$\text{sat}_{\underline{u}}^{\bar{u}}(u) = \min(\bar{u}, \max(u, \underline{u})) = \begin{cases} \bar{u} & \text{når } u \geq \bar{u}, \\ u & \text{når } \underline{u} \leq u \leq \bar{u}, \\ \underline{u} & \text{når } u \leq \underline{u}. \end{cases} \quad (\text{Metningsfunksjonen})$$

Hvordan kan man implementere en ratebegrensning? Gitt en tidsvarierende variabel v , så kan dens ratebegrensninger beskrives vha. en metningsfunksjon:

$$\text{sat}_{\underline{\dot{v}}}^{\bar{\dot{v}}}(\dot{v}) = \min(\bar{\dot{v}}, \max(\dot{v}, \underline{\dot{v}})) = \begin{cases} \bar{\dot{v}} & \text{når } \dot{v} \geq \bar{\dot{v}}, \\ \dot{v} & \text{når } \underline{\dot{v}} \leq \dot{v} \leq \bar{\dot{v}}, \\ \underline{\dot{v}} & \text{når } \dot{v} \leq \underline{\dot{v}}. \end{cases} \quad (\text{Ratebegrensning})$$

Som med metningsfunksjonen, vil vi som regel droppe rategrensene $\bar{\dot{v}}$ og $\underline{\dot{v}}$, og skrive $\text{sat}(\dot{v})$.

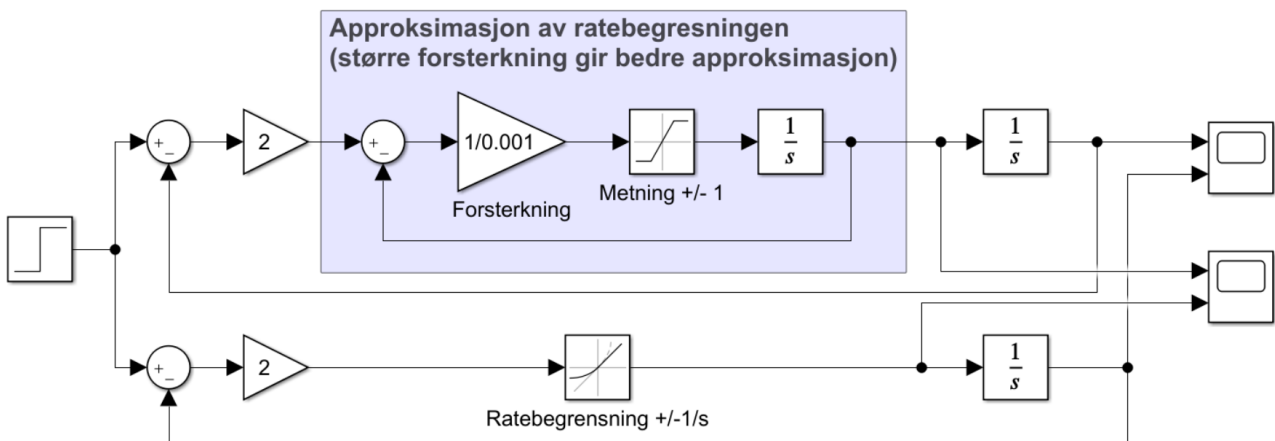
Anta at vi har ratebegrensninger, og at vi ønsker at variabelen v skal endre verdi til v_o . Dette er lett å implementere digitalt hvis vi har et tidssteg \mathfrak{h} :

$$v_{k+1} = v_k + \rho \cdot \mathfrak{h} \quad \text{hvor} \quad \rho = \text{sat}_{\underline{\dot{v}}}^{\bar{\dot{v}}} \left(\frac{v_o - v_k}{\mathfrak{h}} \right).$$

Dette tilsvarer implementasjonen til [ratebegrensnings-blokken i Simulink](#). Merk at man kan approksimere en ratebegrensning som vha. av en tilbakekoblingsløyfe med metning, som vist i figur 3.7.

3.3.2 Friksjon

Figur fra Ruderman?



Figur 3.7: Approksimasjon av ratebegrensning vha. tilbakekoblingsløyfe med metning.

Kinetisk friksjon

Statisk friksjon (“stiksjon”)

Tørr (Coloumb) friksjon

Hydrodynamisk friksjon

Andre friksjonsmodeller

3.3.3 Hysterese

Alternative kilder: [Wikipedia](#)

Hysterese er et fenomen hvor utviklingen til tilstanden til et system avhenger av dets historie, “det har minne”. Merk at dette ikke kan være tilfelle for en teoretisk modell av et system på tilstandsrom; f.eks. er endringer til tilstanden $x(\cdot)$ til det (tids-invariante) systemet $\dot{x}(t) = f(x(t))$ ved et tidspunkt t bare avhenge av tilstanden ved det tidspunktet, $x(t)$.

3.3.4 Dødbånd og backlash

Alternative kilder: [Wikipedia \(dødbånd\)](#) / [Wikipedia \(Backlash\)](#)

3.3.5 Grensesvingninger

Grensesvingninger (eng.: “limit cycles”).

Metning, friksjon (stic-slip + PI) og (tidsforsinkelser), kan føre grensesvingninger.

3.4. *Fluidmekanikk*

3.4.1 Bernoullis ligning:

Alternative kilder: [Wikipedia](#); [YouTube-video 1 \(Bernoullis-prinsipp\)](#) [YouTube-video 2^a](#)

^aMerk: selv om denne videoen er meget vellaget, er noen av de hevdede anvendelsene noe tvilsomme (det er generelt sett ikke pga. Bernoulli-prinsippet at flye flyr, for eksempel).

I oppgaven over ble det hevdet at $q_{ut} = k_v \sqrt{h}$ er en nogenlunde fornuftig modell for utstrømmen av en tank som i eksempel 3.1. Intuitivt sett, så gir det jo mening at væsken/massen i tanken strømmer på grunn av trykket fra mengden med væske som skapes av tyngrekräften. Men hvor kommer akkurat dette uttrykket med kvadratroten fra? Og hva er det i så fall som får masse/væske til å flyte gjennom horisontale rør og komponenter?

Hvis vi antar konstant væsketetthet, så er svaret gitt av det [Bernoulli-prinsippet](#) fra fluid-dynamikken:⁶

Prinsippet, så ligningen.

Bernoullis ligning: For en væske som ikke kan komprimeres (konstant massetetthet), så holder følgende forhold seg konstant langs en strømlinje i en jevn flyt:

$$\mathcal{B} = \rho \frac{v^2}{2} + \rho g z + p = \text{konst.} \quad (\text{Bernoulli-ligningen})$$

Her er ρ tettheten til væsken, z er høyden relativ til en referanse, v er væskens hastighet, g er gravitasjonsakselerasjonen, og p er trykket.

Antagelser: Bernoulli-ligningen krever i utgangspunktet at: 1) man har jevn flyt (altså ikke gyldig når det store endringer (transienter) i flyten), 2) væsken er tilnærmet inkompressibel, 3) det er minimalt med varmeoverføring langs strømlinjen (bør ikke brukes når det er store temperaturendringer, som f.eks. i visse varmevekslere), og 4) viskositets-, turbulens- og friksjons-effekter er neglisjerbare.

Hva er en strømlinje? Tegning

En [strømlinje](#) er linjen en partikkel i en væskestrøm vil ta. Vi sier at en væskestrøm har **jevn flyt** hvis strømlinjene forblir konstant med tiden. I en (teoretisk perfekt) jevn flyt vil derfor en lekebåt flyte langs akkurat samme linje hver gang hvis den starter fra samme punkt.

Hvor kommer ligningen fra? Bernoullis ligning gir bare en (som regel god) approksimasjon av forholdet mellom trykk, hastighet, og høydeforskjell. Selve uttrykket kommer fra bevaring av energi, retttere sagt bevaring a summen av kinetisk-, potensiell og strømnings-/trykk-energi. [Eksempel/enkel utledning her?](#)

⁶[Navier–Stokes ligningen](#) er en partiell differensialligning basert på Newtons andre lov (se § 3.2.1); denne tar også høyde for væskens viskositet, tiltater komprimerbare væsker, og gir dermed en nøyaktig beskrivelse av væskestrømmer. Selv om vi ikke skal se nærmere på denne ligningen i dette faget, er den vell verdt å vite om.

Fun facts, bemerkninger og annet dill dall (you may skip)

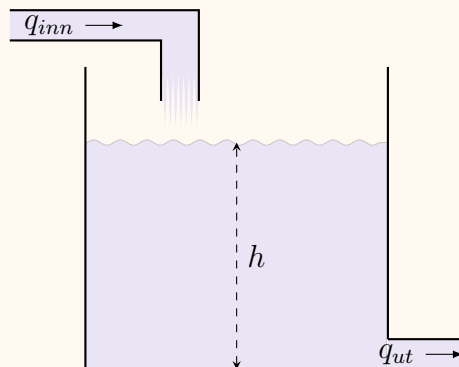
Bernoulli-prinsippet kan brukes til å forklare mange snodige fenomener: <https://youtu.be/eKEorBipbO8>. Det dukker også opp når du fyller bensin: <https://youtu.be/fT2KhJ8W-Kg>.

Merk: Bernoullis ligning er strengt tatt ikke gyldig gjennom en pumpe siden denne vil "ødelegge" en strømlinje. Det finnes dog modifikasjoner til ligningen som kan brukes for én eller flere av antagelsene til den ideelle ligningen (**Bernoulli-ligningen**) ikke holder; se f.eks. sek. 3.4.2.

Eksempel 3.14. Utstrøm av en tank:

Gitt tanken i figuren til høyre, med konstant tverrsnittsareal, A . Vi antar igjen at den inneholder en væske med konstant tetthet ρ .

La subscriptet t bety toppen av tanken ved væskehøyden h , og b bunnen.



1. Trykket i bunnen av tanken: Bernoullies prinsipp gir da følgende forhold:

$$\rho \frac{v_t^2}{2} + \rho g z_t + p_t = \rho \frac{v_b^2}{2} + \rho g z_b + p_b$$

La oss ta referansehøyden slik at $z_b = 0$ og $z_t = h$. Fra massebevarelse har vi $v_t = v_b$ siden tanken har konstant tverrsnitt. Dermed får vi

$$\rho g h + p_t = p_b \quad \implies \quad p_b - p_t = \rho g h \quad \xrightarrow{p_t=0} \quad p_b = \rho g h.$$

2. Væskestrømmen ut av tanken: Anta $q_{inn} = 0$. Hvis væsken strømmer ut i luft, hvor $p_{ute} = 0$, av tanken i et rør med diameter $A_r \ll A$, så får vi

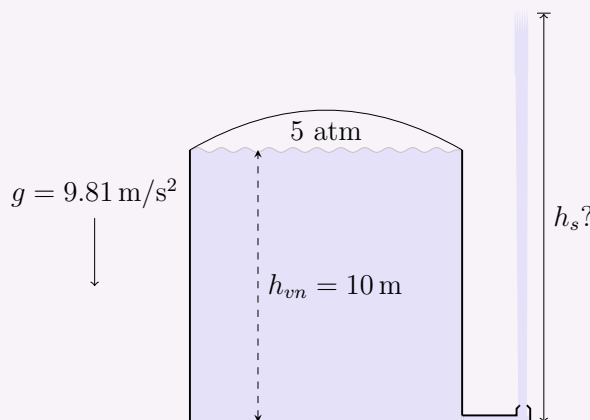
$$\rho \frac{(q_{ut}/A_r)^2}{2} = \rho \frac{(q_{ut}A)^2}{2} + \underbrace{\rho g h}_{=p_b} \quad \implies \quad q_{ut} = k_{ut} \sqrt{h}$$

hvor $k_{ut} = AA_r \sqrt{2g/(A^2 - A_r^2)}$.

Flere eksempler: vannkraft?

Oppgave 3.4. (Fontene)

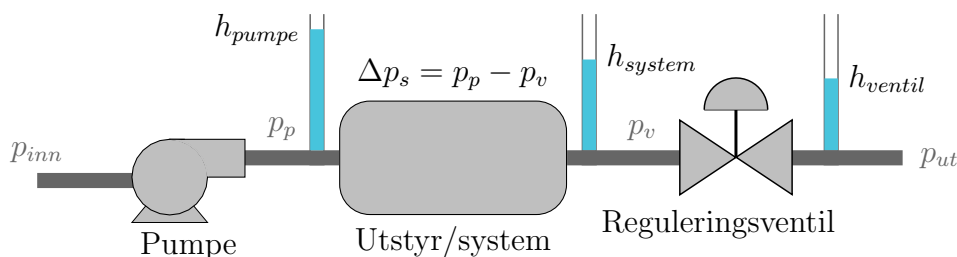
Tanken vist i figuren til høyre brukes tidsvis til å generere en vertikal vannstråle (fontene). Tanken befinner seg på havnivå, hvor det atmosfæriske trykket (atm) er tilnærmet $p_a = 1013 \text{ hPa}$. Tanken er fylt med vann (som har massetetthet $\rho = 1000 \text{ kg/m}^3$). Det antas at væskehøyden holdes konstant lik 10 m, og trykket i toppen av tanken er 5 atm. Hva er den maksimale høyden, h_s , strålen ut fra tanken kan få?



Du kan anta at diameteren til røret ut er ubetydelig sammenlignet med diameteren til tanken, slikt at utstrømmen også er ubetydelig sammenlignet med volumet av vann i tanken. Du kan også se bort fra effekter som friksjon og luftmotstand, etc.

3.4.2 Systemer med pumper og reguleringsventiler

Alternative kilder: §9.2 i [Seborg et al., 2016].



Figur 3.8: Illustrasjon av system med pumpe og reguleringsventil. Trykket ved forskjellige punkter tilsvarer en spesifikk væskehøyde, såkalt *head*, slik at et trykkfall tilsvarer et høydefall, ofte kalt *head loss*.

Witrant 2023 Teaching data-driven-FROM LINEAR DESIGN TO ADAPTIVE CONTROL WITH THROTTLE VALVES

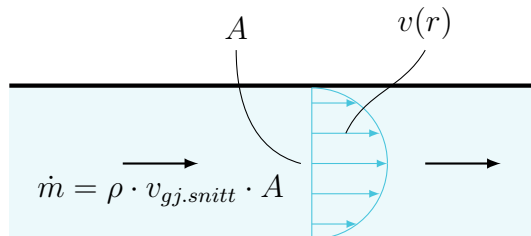
Når man skal regulere strømmer av væsker eller gasser i et system som vist i figur 3.8, bruker man som regel én (eller en kombinasjon) av følgende aktueringsmetoder:

- Reguleringsventil med pumpe som gir konstant trykk (modifiserer i prinsippet systemkarakteristikken: $R_s + R_v$ via ventilmotstanden R_v). Dette er nok den vanligste metoden, men har ulempen at man har et energitap pga. trykkfallet over ventilen (det er dette man bruker til å regulere strømmingen).
- Pumpe med variabelt trykk, styrt via frekvensomformer (modifiserer i prinsippet forsyningsstrykket: $\left(\frac{n}{n_0}\right)^2 p_0$ hvor n er turtallet til pumpen);

3.4.3 Head loss og den modifiserte Bernoulli-ligningen

Som nevnt, så reduseres en væskestrøm over en (regulerings-)ventil pga. trykkfallet over ventilen. Trykket før og etter ventilen kan representeres vha. av en væskehøyde⁷ (se figur 3.8), ofte kalt **head**, slik at trykkfallet kan representeres som et fall i høyden, såkalt **head loss**.

Utleddning av tap vha. Newton



Figur 3.9: Ved en (laminær) væskestrøm i et rør vil hastigheten være lik null ved rørkanten gitt heftebetingelsen. Massestrømmen gjennom røret vil derfor være gitt av gjennomsnittshastigheten $v_{gj.snitt}$.

Heftebetingelse (eng. **no-slip condition**) er en vanlig antagelse for væskestrømmer i rør. Denne antagelsen sier at hastigheten til væsken ved rørkanten er lik null, slik som vist i figur 3.9. Dette betyr at det er et tap, pga. friksjon, av energien til væskestrømmen, hvor kinetisk energi blir omgjort til varme. Dette fører dermed også til et trykkfall (head loss).

Bernoullis modifiserte ligning: Bernoulli-ligningen var basert på energibalanse uten tap. Hvis vi tar hensyn til energi tilførsel via pumper, samt tap pga. friksjon og head loss får vi den modifiserte ligningen:

Bernoullis modifiserte ligning:

$$\frac{p_1}{\rho g} + \alpha_1 \frac{v_1^2}{2g} + z_1 + h_{pumpe} = \frac{p_2}{\rho g} + \alpha_2 \frac{v_2^2}{2g} + z_2 + h_{tap} + h_{turbin} \quad (\text{Mod. Bernoullis ligning})$$

Hvor

- h_{pumpe} tilsvarer en trykkøkning fra pumper;
- h_{tap} tilsvarer trykktap pga. friksjon og reguleringsventiler etc. (head loss);
- h_{turbin} tilsvarer trykktap pga. turbiner og lignende;
- z_i er den vertikale høyden fra en referanse;
- v_i er gjennomsnittshastigheten;
- α_i er en korreksjonsfaktor pga. av hastighetsprofilen (se fig. 3.9).^a

Merk: hvert ledd tilsvarer en “høyde” (med benevning meter).

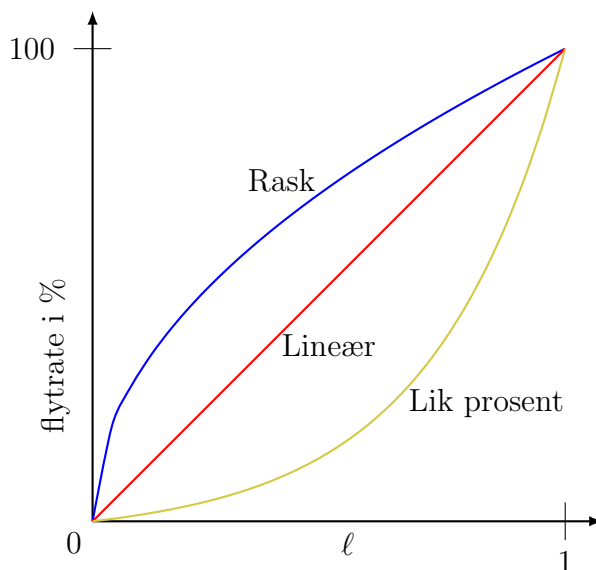
^aKorreksjonsfaktoren for en væskestrøm som i fig. 3.9 er gitt ved $\alpha = \frac{1}{A} \int_A \left(\frac{v(r)}{v_{gj.snitt}} \right)^3 dA$. Merk: $\alpha = 1$ ofte OK.

⁷Husk at for en væske med tetthet ρ , så kan trykket, p [N/m²] relateres til en høyde, h , via formellen $p = \rho gh$.

3.4.4 Reguleringsventiler

Åpne- og lukkekaraktistikk Utrolig nok, så kan åpningen til en ventil bare være et sted mellom helt lukket og helt åpent (et eksempel på metning, se § 3.3.1). Dette vil representeres som regel ved variabelen ℓ (merk krøllete ℓ gitt ved `\ell` i LaTeX) som kan ta verdier mellom 0 og 1 (evt. 0 – 100, tenk prosent), hvor $\ell = 0$ betyr helt lukket, og $\ell = 1$ betyr helt åpent.

Det er viktig å være klar over at $\ell = 0.5$ ikke nødvendigvis betyr at ventilen er 50% åpen. Dette avhenger av ventilens **lukkekaraktistikk**. Et eksempel på slike karakteristikker er vist i figur 3.10.



Figur 3.10: Forskjellige ventilkarakteristikker.

Væskestrømmen gjennom en ventil antas ofte å være gitt av **ventilligningen**:

Ventilligningen:

$$q_v = k_v f_v(\ell) \sqrt{\Delta p_v} / \sqrt{g_s} = c_v f_v(\ell) \sqrt{\Delta p_v} = C_v f_v(\ell) \sqrt{\Delta h_v} \quad (3.10)$$

hvor

q_v [$\text{m}^3 \text{s}^{-1}$] er væskestrømmen gjennom ventilen;

k_v [$\text{m}^3 / (\text{s} \sqrt{\text{Pa}})$] er ventilkoeffisienten;^a

$f_v(\ell)$ er ventilkarakteristikken (se figur 3.10);

$\ell \in [0, 1]$ er ventilåpningen;

Δp er trykkfallet over ventilen i Pascal (N m^{-2});

g_s er den dimensjonsløse **relative tettheten/spesifikke gravitasjonen** til væsken;

c_v samlekonstant som avhenger av tettheten ρ ;

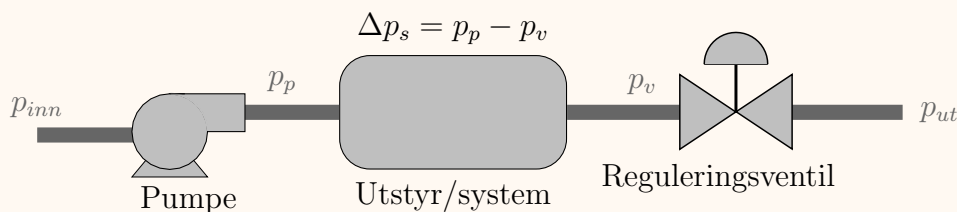
C_v samlekonstant som ikke avhenger av tettheten ρ ;

Δh_v er “head losset” over ventilen.

“I [Seborg et al., 2016] regnes ventilkoeffisient som dimensjonsløs, slik at man har en egen “enhetskonstant” N .”

Så hvorfor skulle man noen gang velge en annen karakteristikk enn en lineær en? La os ta et eksempel:

Eksempel 3.15. (Forenkling av eks. 9.3 i [Seborg et al., 2016]) Gitt systemet i figuren under, bestående en pumpe, et system og en reguleringsventil.



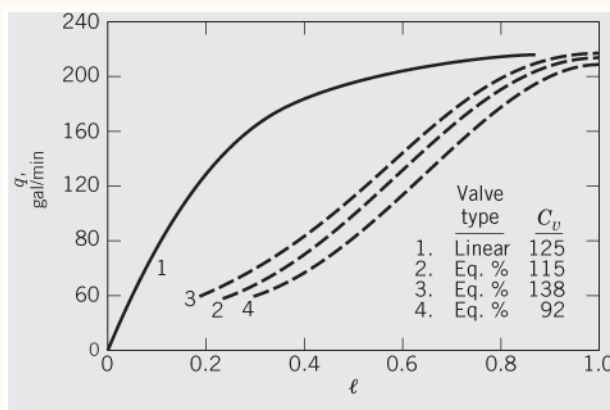
Anta at: 1) $p_{inn} = p_{ut} = 0$, 2) pumpen genererer et konstant trykk, p_p , og at 3) trykkfallet over systemet er proporsjonalt med med strømningsraten q igjennom det, altså $\Delta p_s = c_s q^2$. Anta videre at strømmingen gjennom ventil følger ventilligningen (3.10), slik at $q = c_v f_v(\ell) \sqrt{\Delta p_v}$. Vi har dermed

$$\Delta p = p_p - p_{ut} = p_p = \Delta p_s + \Delta p_v \implies \Delta p_v = p_p - c_s q^2.$$

Vi får derfor fra ventilligningen at

$$q = c_v f_v(\ell) \sqrt{p_p - c_s q^2} \implies q^2 = \frac{c_v^2 f_v^2(\ell)}{1 + c_v^2 f_v^2(\ell) c_s}.$$

Dette viser at ved hjelp av forskjellige ventilkarakterestikker kan man få et tilnærmet ønsket forhold (f.eks. lineært) mellom åpningen ℓ og strømmingen q om et ønsket arbeidspunkt; se følgende figur fra [Seborg et al., 2016] som viser ventilåpning mot strømmingen for forskjellige ventilkarakterestikker:



3.4.5 *Motorer*

figur i <https://ieeexplore.ieee.org/document/9795111>

Likestrømsmotor

Se 3.4.3 [Bjørvik and Hveem, 2014]. Bruke energibalanse.

3.4.6 Pumper

Dynamikk

Approximeres som regel som et første-ordens-pluss-tidsforsinkelse system om et gitt arbeidspunkt:

$$\dot{q}_v(t) = -aq_v(t) + bq_v^*(t - \theta),$$

hvor a og b er positive konstanter (for arbeidspunktet), $\theta > 0$ er en konstant tidsforsinkelse, og q_v^* er den ønskede værskestrømmer (satt ved tiden t).

3.5. *Aktuatorer og sensorikk*

Dere vil lære enda mer om dette i [IELET2106 - Industriell instrumentering](#).

Litt om valg av aktuatorer og sensorikk: Det er flere aspekter som må tas hensyn til når man skal velge dette. Den viktigste avveiningen er som oftest kostand vs ytelse, hvor man vil finne et “godt nok” utstyr til en rimelig penge. Eksempler på slike kriterier kan være:

- Hurtighet: Noen prosesser har veldig rask dynamikk (tenk en pendel), mens andre prosesser varierer sakte (f.eks. prosesser hvor tilstanden er varme) – utstyret bør som regel være betydelig raskere enn prosessen man skal regulere;
- Dimensjonering: utstyret må selvsagt være “kraftig” nok for prosessen (sterk nok pumpe, stort nok batteri, bredt nok måleområde, etc.), og man bør ofte ha greit med slingsmonn;
- Materialer og durabilitet: vann vs sterk syre, helium vs plasma, riste dun vs knuse stein – utstyret må tale prosessen;
- Presisjon: skal du regulere ting på nano-nivå, eller er det snakk om en buffer-tank? Bør aktuatoren/måleutstyret ha en mest mulig lineær karakteristisk?

Enheter: Pass på enhetene! Er ikke bare [SI-enheter](#) som blir brukt. Selv om feil kan skje de beste av oss ([se for eksempel denne](#)), så bør man alltid se godt etter i databladet.

4. Modellforenkling og -tilpasning

4.1. Første- og andreordens lineære reguleringsystemer

Vi skal nå se på lineære første- og andreordens systemer med én inngang og én utgang, såkalte monovariabel-/ SISO-systemer (eng. single input, single output).

4.1.1 Første-ordens systemer



Alternative kilder: §5.2 i [Seborg et al., 2016].

Et lineært første-ordens system har følgende tilstandsromform:

$$\dot{y} = -ay + bu. \quad (4.1)$$

Løsningen er¹

$$y(t) = e^{-at} \left(y(0) + b \int_0^t e^{a\sigma} u(\sigma) d\sigma \right). \quad (4.2)$$

Hvis $y(0) = 0$, så er tilsvarende overføringsfunksjon:

$$G(s) = \frac{Y(s)}{U(s)} = \frac{b}{s + a}.$$

Tidskonstant for første-ordens systemer

¹Multipliser begge sider av (4.1) med den integrerende faktoren e^{at} : $e^{at}\dot{y} = -e^{at}ay + e^{at}bu$. Dette kan skrives som $e^{at}\dot{y} + e^{at}ay = \frac{d}{dt}(e^{at}y(t)) = e^{at}bu(t)$, hvorfra vi får følgende ved å integrere mhp. tid på begge sider: $\int_0^t \frac{d}{d\tau}(e^{a\tau}y(\tau)) d\tau = [e^{a\tau}y(\tau)]_0^t = e^{at}y(t) - y(0) = \int_0^t e^{a\tau}bu(\tau)d\tau$. Ved å legge til $y(0)$ på begge sider, samt multiplisere med e^{-at} så får vi da (4.2).

Tidskonstant til første-ordens system: Tiden τ det tar å nå ca. 63% av stasjonærverdien etter et sprang i inngangen. For et system på formen (4.1) kan man finne denne fra

$$\tau = \frac{1}{a}$$

⚠ Det gir selvsagt kun mening å snakke om tidskonstanten hvis $a > 0$, altså når systemet er stabilt i åpen sløyfe.

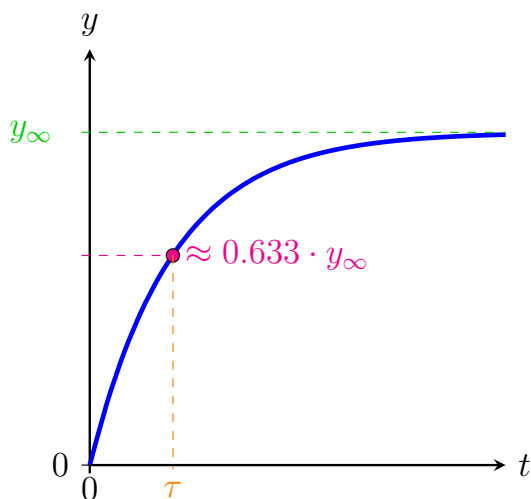
Så hvor kommer akkurat 63% fra? Anta at $y(0) = 0$, og la oss ta $\tau = 1/a$ og $k = b/a$ slik at

$$\dot{y} = -\frac{1}{\tau}(y - ku) \iff Y(s)/U(s) = \frac{k}{\tau s + 1}$$

Her kalles k ofte for den stasjonære forsterkningen. For et sprang i u fra 0 til $u_s = \text{konstant}$ har dette systemet løsningen

$$\int_0^t \dot{y}(\sigma) d\sigma = \int_0^t \left(-\frac{1}{\tau}(y(\sigma) - ku(\sigma)) \right) d\sigma \implies y(t) = (1 - e^{-t/\tau})ku_s$$

Når $t = \tau$ har man dermed $y(\tau) = (1 - e^{-1})ku_s = (1 - e^{-1})y_\infty \approx 0.63 \cdot y_\infty$; se figur 4.1.



t	$1 - e^{-t/\tau}$
0	0
τ	≈ 0.6321
2τ	≈ 0.8647
3τ	≈ 0.9502
4τ	≈ 0.9817
5τ	≈ 0.9933

- a) Respons etter sprang i u . b) Respons som faktor av tidskonstanten.

Figur 4.1: Sprangrespons og tidskonstant for første-ordens system.

4.1.2 Andre-ordens systemer



Alternative kilder: §5.4 i [Seborg et al., 2016].

Differensialligningen til et lineært andre-ordens system er

$$\ddot{x} + 2\omega_0\zeta\dot{x} + \omega_0^2x = bu, \tag{4.3a}$$

$$y = x. \tag{4.3b}$$

For $x_1 = x$ og $x_2 = \dot{x}$ er tilstandsromformen:

$$\begin{aligned}\dot{x}_1 &= x_2 \\ \dot{x}_2 &= -2\omega_0\zeta x_2 - \omega_0^2 x_1 + bu, \\ y &= x_1.\end{aligned}$$

Damping for andre-ordens systemer: Andreordens systemer på formen (4.3) har visse karakteristikkene basert på følgende størrelser:

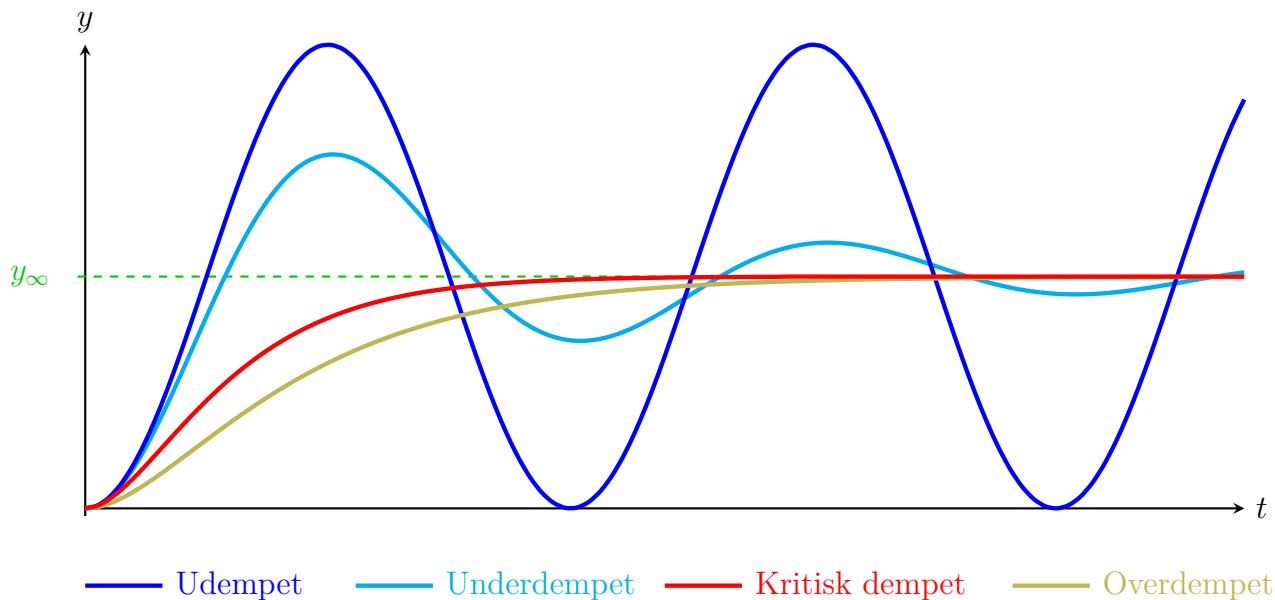
ω_0 er den udempede svinge-frekvensen;

ζ er den relative dempningsfaktoren.

Mer spesifikt, så sies systemet (4.3) med $u = 0$ og $\omega_0 > 0$ å være

- **ustabilt** når $\zeta < 0$;
- **udempet** (marginalt/kritisk stabilt) når $\zeta = 0$;
- **underdempet** når $0 < \zeta < 1$;
- **kritisk dempet** hvis $\zeta = 1$;
- **overdempet** når $\zeta > 1$.

Sprangresponsen til de forskjellige tilfellene er vist i figur 4.2.



Figur 4.2: Sprangresponser til andre-ordens systemer med forskjellig dempningsfaktor.

Løsning fra sprangrespons Gitt en sprangrespons $U(s) = u_s/s$, så kan overføringsfunksjonen til et andre-ordens system skrives på formen

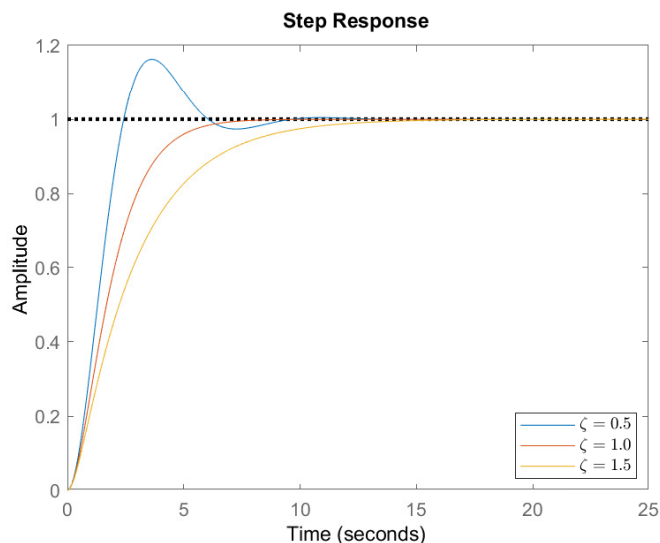
$$Y(s) = \frac{ku_s}{s(\tau^2 s^2 + 2\zeta\tau + 1)},$$

hvor da altså $\tau = 1/\omega_0$, $k = b/\omega_0^2$ og u_s er den (konstante) sprang-amplituden. Ved å betegne

$$\tau_1 = \tau\zeta + \tau\sqrt{\zeta^2 - 1} \quad \text{og} \quad \tau_2 = \tau\zeta - \tau\sqrt{\zeta^2 - 1}$$

slik at $\tau^2 s^2 + 2\zeta\tau + 1 = (1 + \tau_1 s)(1 + \tau_2 s)$, så er løsningen ved et slikt sprang for $y(0) = \dot{y}(0) = 0$ som følger for de forskjellige dempningsratioene:

- **udempet:** $y(t) = u_s k (1 - \cos(t))$;
- **underdempet:** $y(t) = u_s k \left(1 - e^{-\zeta t/\tau} \left[\cos\left(\frac{\sqrt{1-\zeta^2}}{\tau} t\right) \right] + \frac{\zeta}{\sqrt{1-\zeta^2}} \sin\left(\frac{\sqrt{1-\zeta^2}}{\tau} t\right) \right)$;
- **kritisk dempet:** $y(t) = u_s k \left(1 - \left(1 + \frac{t}{\tau}\right) e^{-t/\tau} \right)$;
- **overdempet:** $y(t) = u_s k \left(1 - \frac{\tau_1 e^{-t/\tau_1} - \tau_2 e^{-t/\tau_2}}{\tau_1 - \tau_2} \right)$.



Figur 4.3: Sprangresponser til andre-ordens systemer med forskjellig dempningsfaktor.

Figur 4.3 (som er generert vha. koden i kodesnutt 4.1) viser sprangresponsen til et andre-ordens system for forskjellige dempningsfaktorer

Kodesnutt 4.1: Sprangresponser for et AO system med forskjellige dempningsfaktorer.

```
s=tf('s');
w = 1;
zetas = [0.5, 1, 1.5];
figure(1); clf(1);
hold on;
for i=1:3
```

```

zeta=zetas(i);
G=1/(s^2+2*zeta*w*s+w^2);
step(G);
end
legend({'$\zeta=0.5$', '$\zeta=1.0$', '$\zeta=1.5$'}, 'Interpreter', 'latex', 'Location', 'best');

```

4.2. Modellforenkling

Vi skal nå se på metoder for å forenkle høyere-ordens modeller til enklere lavereordens modeller, og da hovedsakelig første- og andreordens prosesser med tidsforsinkelser (abbreviert hhv. FOPTF og AOPTF). For dette trenger vi dog først kjennskap til tidsforsinkelser og hvordan vi kan representere disse.

4.2.1 Systemer med tidsforsinkelser

▶ F15SRrBDHqU&t=270s

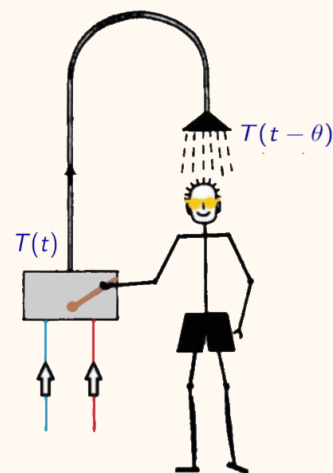
Alternative kilder: [Bjørvik and Hveem, 2014, §4.1 og §5.6], [Fridman, 2014]. Brian Douglas video.

Tidsforsinkelser (også kalt dødtid eller transportforsinkelse, avhengig sammenheng) er uunngåelig i del fleste reguleringssystemer grunnet tiden det tar å skaffe seg informasjonen (målinger) som trengs for beslutningstaking, for å skape reguleringsbeslutninger, og å gjennomføre disse beslutningene (sette et pådrag). Noen eksempler på årsaker til tidsforsinkelser er masse-transport i et rør, dødtid i en motor p.g.a. statisk friksjon, aktuator-dynamikk, kommunikasjon over nettverk (kremt, CAN-buss, kremt), samt sampling- og utregningstid. Siden tidsforsinkelser potensielt kan føre til en betraktelig begrensning i ytelsen til et reguleringssystem, er det viktig at man tenker på tidsforsinkelser når man utvikler en regulator.

Eksempel 4.1. (I dusjen, eksempel og figur tatt fra [Fridman, 2014])

Se for deg en dusjende person som ønsker å oppnå ønsket vanntemperatur, T_d , ved vha. blande-batteriet, som vist i figuren til høyre. La $T(t)$ betegne vanntemperaturen i blande-batteriets utgang ved tiden t , og la θ være den konstante tiden som trengs av vannet for å gå fra mikserutgang til personens hode. Anta at endringen av temperaturen er k -proporsjonal med rotasjonsvinkelen til håndtaket, mens hastigheten man kan rotere håndtaket er proporsjonal med $T(t) - T_d$. Ved tidspunktet t føler personen vanntemperaturen som forlater mikseren på tidspunktet $t - \theta$, noe som resulterer i følgende ligning med konstant forsinkelse θ :

$$\dot{T}(t) = -k(T(t - \theta) - T_d).$$



Hvis vi definerer avviket som $e(t) = T(t) - T_d$, så får vi følgende avviksdynamikk:

$$\dot{e}(t) = -k \cdot e(t - \theta).$$

Merk: uten tidsforsinkelsen kunne vi (teoretisk sett) gjort responsen så rask og responsiv vi bare ville ved å øke k siden løsningen da er $e(t) = e(0) \exp(-k \cdot t)$. Historien er dog en helt annen hvis vi har en tidsforsinkelse, siden for aggressive endringer (stor k) da kan ville føre til oversving, og iverste fall ustabilitet.

4.2.2 Tidsforsinkelser representert i Laplace-domenet



La oss minne oss på hvordan Laplacetransformasjonen til en tidsforsinket signal ser ut:

Teorem 4.1. Gitt en kontinuerlig funksjon $f(\cdot)$, $f(t \leq 0) = 0$, og en konstant tidsforsinkelse, $\theta \geq 0$, så er den ensidige Laplacetransformasjon av $f(t - \theta)$ ved tiden t gitt av^a

$$\mathcal{L}\{f(t - \theta)\} = \int_0^\infty f(\sigma - \theta)e^{-\sigma s} d\sigma = e^{-\theta s} F(s).$$

^aFor å komme fram til dette bruker vi $f(w) = 0$ for $w \leq 0$ og tar i bruk følgende bytte av integrasjonsvariabel: $w = \sigma - \theta$. Dermed $\int_0^\infty f(\sigma - \theta)e^{-\sigma s} d\sigma = \int_{-\theta}^\infty f(w)e^{-(w+\theta)s} dw = e^{-\theta s} \left[\int_{-\theta}^0 f(w)e^{-ws} dw + \int_0^\infty f(w)e^{-ws} dw \right] = e^{-\theta s} \int_0^\infty f(w)e^{-ws} dw = e^{-\theta s} F(s)$.

Eksempel 4.2. Fra «I dusjen»-eksemelet over hadde vi $\dot{e} = -k \cdot e(t - \theta)$. Ved å ta Laplace-transformasjonen på begge sider gir dette oss dermed $sE(s) - e(0) = -kE(s) \cdot \exp(-\theta s)$, slik at $E(s) = \frac{e(0)}{s+k \exp(-\theta s)}$. For avviksdynamikken skal være er stabil må denne overføringsfunksjonene ha alle sine poler i venstre halvplan, noe som tilsvarer at nullpunktene til $s + k \exp(-\theta s)$ er i venstre halvplan.

4.2.3 Forenkling/approksimasjoner av tidsforsinkelser



Første-orden (Taylor) approksimasjoner: Taylor ekspansjonen av tidsforsinkelsen er:

$$e^{-\theta s} = 1 + (-\theta)s + \frac{1}{2}\theta^2 s^2 + \underbrace{\text{deler av høyere orden}}_{=O(|s|^3)}$$

Første-ordens approksimasjonen (også kalt trunkert Taylor-rekke av orden 1) gir derfor

$$e^{-\theta s} \approx 1 - \theta s \tag{4.4}$$

altså en ikke-proper og ikke-minimumfase overføringsfunksjon (nullpunkt i høyre halvplan).

På den annen side kan vi også skrive

$$e^{-\theta s} = \frac{1}{e^{\theta s}} \approx \frac{1}{1 + \theta s}, \tag{4.5}$$

altså som en pol i venstre halv-plan (første-ordens lavpassfilter).

Padé approksimasjon: En **Padé-approksimasjon** finner man ved å ekspandere en funksjon som et forhold mellom to potensserier, hvor man bestemmer både teller- og nevnerkoeffisientene vha. funksjonens Taylor-approksimasjoner.

Første-ordens Padé-approksimasjonen til en tidsforsinkelse er

$$e^{-\theta s} \approx \frac{1 - \frac{1}{2}\theta s}{1 + \frac{1}{2}\theta s}, \tag{4.6}$$

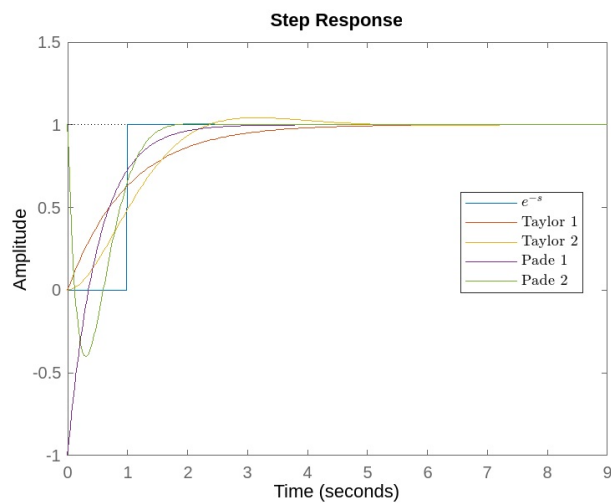
som i motsetning til (4.4) er proper, men fortsatt ikke-minimum fase pga. nullpunktet i høyre halvplan. Merk at approksimasjonen (4.6) i dette tilfelle (det holder dog ikke generelt sett) så tilsvarer

$$e^{-\theta s} = e^{-\frac{1}{2}\theta s} e^{-\frac{1}{2}\theta s} = \frac{e^{-\frac{1}{2}\theta s}}{e^{\frac{1}{2}\theta s}},$$

altså ved å ta første-ordens Taylor-approksimasjonen av både telleren og nevneren.

Se figur 4.4 for en sammeligning av disse approksimasjonene.

```
s=tf('s');
Ptf=exp(-s);
Pt1=1/(1+s);
Pt2=1/(1+s+0.5*s^2);
Pp1=(1-0.5*s)/(1+0.5*s);
Pp2=pade(Ptf,2);
figure(1); clf(1);
hold on;
step(Ptf);
step(Pt1);
step(Pt2);
step(Pp1);
step(Pp2);
legend({'$e^{-s}$', 'Taylor 1', 'Taylor 2', 'Pade 1', 'Pade 2'}, 'Interpreter', 'latex')
```



Figur 4.4: Høyre: Sprangrespons ved forskjellige Taylor- og Padé-approksimasjoner av en tidsforsinkelse ($\theta = 1$ s). Generert vha. MATLAB-koden til venstre.

4.2.4 Første-ordens-pluss-tidsforsinkelse (FOPTF) systemer



Første-ordens-pluss-tidsforsinkelse (FOPTF) systemer (eng. first-order plus time delay (FOPTD)) på tilstandsromform er som følger:

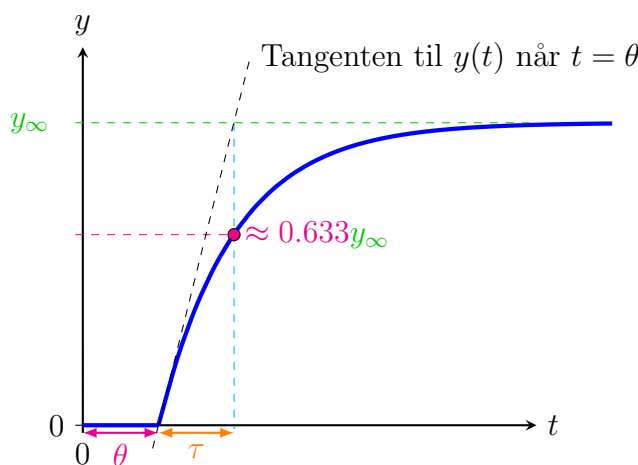
$$\dot{y} = -ay + bu(t - \theta)$$

hvor a, b er konstante parametere, og $\theta > 0$ er en konstant tidsforsinkelse. Hvis $y(0) = 0$, så er tilsvarende overføringsfunksjon

$$G_{FOPTF}(s) = \frac{k \cdot e^{-\theta s}}{1 + \tau s}, \tag{FOPTF}$$

hvor $\tau = 1/a > 0$ og $k = b/a$. Sprangresponsen til dette systemet er vist i figur 4.5. Du finner mer om dette i vedlegg C.3. Merk at hvis τ er veldig stor, så er dette tilnærmet en rent *integrerende prosess*.

En FOPTF-modell er tilstrekkelig for å modellere en rekke industrielle prosesser, hvor formålet med modellering er PID-regulering. Faseforsinkelsen introdusert gjennom forsinkelsen θ gjør FOPTF-modellen til en god tilnærming av høyere ordens dynamikk i faseområdet av interesse for PID-regulering. Et viktig aspekt her er den såkalte *relative tidsforsinkelsen*.



Figur 4.5: Sprangrespons av et første-ordens-pluss-tidsforsinkelse-system.

Relativ tidsforsinkelse: Den *relative tidsforsinkelsen* er definert som

$$\tau_r = \frac{\theta}{\theta + \tau} \tag{Relative tidsforinskelsen}$$

slik at $0 \leq \tau_r \leq 1$.

Dette forholdet gir en viktig karakterisering av dynamikken. En liten τ_r betyr at dynamikken er dominert av tidskonstanten $\tau \gg \theta$; en stor τ_r ($\gg 1$) betyr derimot at dynamikken er dominert av forsinkelsen θ . Vi sier at systemet er *lag-dominant* hvis τ_r er liten ($\tau < 0.3$) og *tidsforsinkelses-dominant* hvis $\tau_r \approx 1$.

4.2.5 Andre-ordens-pluss-tidsforsinkelse (AOPTF) systemer

Et andre-ordens-pluss-tidsforsinkelses (AOPTF) system kan skrives på en rekke former:

$$G_{AOPTF}(s) = \frac{k e^{-\theta s}}{\tau^2 s^2 + 2\zeta\tau s + 1} = \frac{k\omega_0^2 e^{-\theta s}}{s^2 + 2\zeta\omega_0 s + \omega_0^2} \tag{AOPTF}$$

hvor θ er tidsforsinkelsen/dødtiden, k er forsterkningen, $\omega_0 = 1/\tau$ er den udempede svingefrekvensen, og ζ er den relative dempningsfaktoren. Hvis $\zeta \geq 1$, så kan vi ta $\tau_1 = \tau\zeta + \tau\sqrt{\zeta^2 - 1}$

og $\tau_2 = \tau\zeta - \tau\sqrt{\zeta^2 - 1}$, slik at

$$G_{AOPTF}(s) = \frac{ke^{-\theta s}}{(1 + \tau_1 s)(1 + \tau_2 s)}. \quad (\text{AOPTF med reelle poler})$$

4.2.6 Modellforenkling ved sammenslåtte tidskonstanter



Nåværende problem: Gitt en overføringsfunksjon på formen

$$G(s) = \frac{k(1 - \ell_1 s) \cdots (1 - \ell_m s)}{(1 + \tau_1 s)(1 + \tau_2 s) \cdots (1 + \tau_n s)} e^{-\theta s} = k \frac{1 - (\ell_1 + \cdots + \ell_m)s + \mathcal{O}(s^2)}{1 + (\tau_1 + \tau_2 + \cdots + \tau_n)s + \mathcal{O}(s^2)} e^{-\theta s}, \quad (4.7)$$

hvor $\tau_1 \geq \tau_2 \geq \cdots \geq \tau_n \geq 0$ og $\ell_1 \geq \ell_2 \geq \cdots \geq \ell_m \geq 0$. Vi ønsker å approksimere systemet som en pte-ordens prosess med tidsforsinkelse.

En enkel måte å løse dette på er å øke den effektive tidsforsinkelsen ved å slå sammen visse tidskonstanter, samt eventuelle nullpunkt i høyre halvplan ved hjelp av approksimasjonene vi så i § 4.2.3:

Tidsforsinkelse fra sammenslåtte tidskonstanter: Approksimer (4.7) som en pte-ordens overføringsfunksjon med tidsforsinkelse som følger:

$$\tilde{G}_p(s) = \frac{k}{(1 + \tau_1 s)(1 + \tau_2 s) \cdots (1 + \tau_p s)} e^{-\tilde{\theta} s},$$

hvor den effektive tidsforsinkelsen er gitt ved

$$\tilde{\theta} = \theta + \tau_{p+1} + \tau_{p+2} + \cdots + \tau_n + \ell_1 + \cdots + \ell_m.$$

Eksempel 4.3. Vi vil approksimere

$$P(s) = \frac{2}{(2 + 2s)(1 + 0,2s)(1 + 0,02s)(1 + 0,01s)} e^{-0,5s}$$

som en første- og andre-ordens pluss tidsforsinkelses overføringsfunksjon.

Vi skriver om $(2 + 2s) = 2(1 + s)$ som dermed inneholder den største tidskonstanten. De resterende tidskonstantene legger vi til i den nye tidsforsinkelsen: $\tilde{\theta} = 0,5 + 0,2 + 0,02 + 0,01 = 0,73$, slik at

$$\tilde{P}_1(s) = \frac{1}{(1 + s)} e^{-0,73s};$$

mens en andre-ordens approksimasjon tilsvarer

$$\tilde{P}_2(s) = \frac{1}{(1 + s)(1 + 0,2s)} e^{-0,53s}.$$

4.2.7 Modellforenkling ved Skogestads halv-regel



Alternative kilder: §6.3 i [Seborg et al., 2016]; [Skogestad, 2003]

Når det kommer til regulering, så er tidsforsinkelser mer krevende enn “lag” fra tidskonstanter. Skogestad foreslo derfor en alternativ regel, hans såkalte *halv-regel* (eng. “half-rule”), hvor man jevnt fordeler den $p + 1$ største tidskonstanten mellom tidsforsinkelsen og tidskonstanten.

Halv-regelen: den største neglisjerte (nevner) tidskonstanten (lag) er fordelt jevnt på effektiv forsinkelse og den minste gjenværende tidskonstanten.

Det vil si, approksimer (4.7) som en p te-ordens overføringsfunksjon med tidsforsinkelse som følger:

$$\tilde{G}_p(s) = \frac{k}{(1 + \tau_1 s)(1 + \tau_2 s) \cdots (1 + (\tau_p + \frac{1}{2}\tau_{p+1})s)} e^{-\tilde{\theta}s},$$

hvor den effektive tidsforsinkelsen er gitt ved

$$\tilde{\theta} = \theta + \frac{1}{2}\tau_{p+1} + \tau_{p+2} + \cdots + \tau_n + \ell_1 + \cdots + \ell_m.$$

Merk: Skogestad foreslår også en måte å handtere negative ℓ , men vi vil ikke se på det i disse notatene.

Eksempel 4.4. Vi vil approksimere

$$P(s) = \frac{1}{(1 + s)(1 + 0,2s)(1 + 0,02s)(1 + 0,01s)} e^{-0,5s}$$

som en første- og andre-ordens pluss tidsforsinkelses overføringsfunksjon.

Vi har $\tilde{\theta} = 0,5 + \frac{0,2}{2} + 0,02 + 0,01 = 0,63$, slik at

$$\tilde{P}_1(s) = \frac{1}{(1 + (1 + \frac{0,2}{2})s)} e^{-0,63s} = \frac{1}{(1 + 1,1s)} e^{-0,63s};$$

mens en andre-ordens approksimasjon tilsvarer

$$\tilde{P}_2(s) = \frac{1}{(1 + s)(1 + 0,21s)} e^{-0,52s}.$$

Når bør man approksimere som en andre-ordens modell? Approksimer som en AOPTF fremfor FOPTF hvis τ_2 større enn θ .

Oppgave 4.1. Anta følgende prosess:

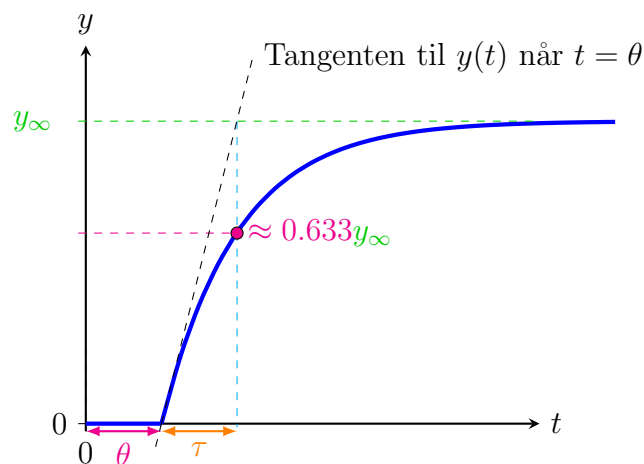
$$P(s) = \frac{3,6}{(3 + 0,3s)(1 + 0,3s)(2 + 0,3s)} e^{-0,7s}$$

Tilnærm overføringsfunksjonen som en første-ordens prosess med forsinkelse ved å slå sammen tidskonstantene. Bruk så Skogestads halv-regel til å tilnærme overføringsfunksjonen som **både** en første- og andre-ordens prosess med forsinkelse.

4.3. Modellering fra empiriske data

4.3.1 Kort om systemidentifikasjon

For reguleringstekniske formål trenger vi ikke alltid en global modell (en modell som er gyldig for alle tislntander). I stedet er ofte en modell som tilnærmer prosessen/systemet godt nært det ønskede arbeidsområdet bra nok. [Mer her](#)



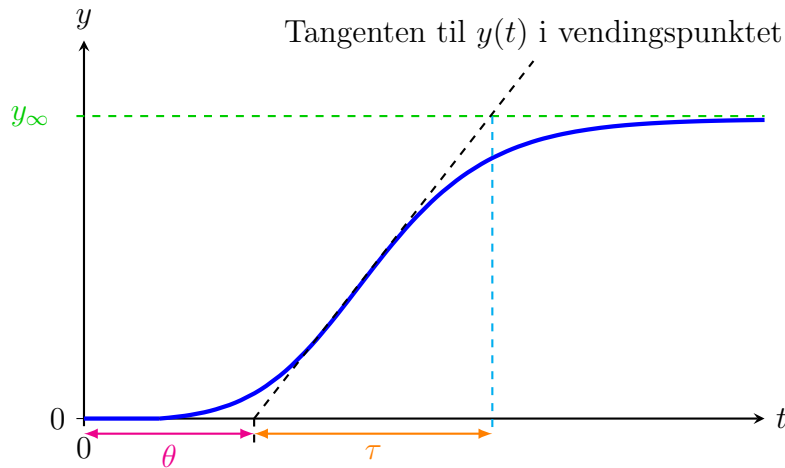
Figur 4.6: Sprangsrespons for tilpassing av FOPTF-modell.

4.3.2 Tilpasning av FOPTF-modell fra sprangrespons ▶ nFN5iaF2IzY&t=311s

Alternative kilder: §7.2 i [Seborg et al., 2016].

Vi ønsker å tilpasse et system vha. av en første-ordens-pluss-tidsforsinkelse (FOPTF) modell:

$$G_{FOPTF}(s) = \frac{k \cdot e^{-\theta s}}{1 + \tau s}. \quad (\text{FOPTF})$$



Figur 4.7: Sprangrespons for tilpassing høyere-ordens systemet til FOPTF-modell.

Prosedyre for tilpassing av FOPTF-modell fra sprangrespons:

1. Ta utgangspunkt i en enkel sprangrespons (u fra u_0 til u_s) i åpen sløyfe.
2. Tilpass en førsteordens modell med tidsforsinkelse se (FOPTF) til responsen:
 - (a) Bruk figur 4.6 hvis responsen «ser ut som» en FOPTF respons;
 - (b) bruk figur 4.7 hvis det er tegn til høyere-ordens dynamikk;
 - (c) ta $k = \Delta y / \Delta u = (y_\infty - y_0) / (u_s - u_0)$.

Merk: Mer detaljer rundt tilpassning av først- og andre-ordens modeller med tidsforsinkelse fra sprangresponser er gitt i vedlegg C.3

4.3.3 Visuell tilpassing av underdempet AOPTF-modell

Gitt et underdempet AOPTF system

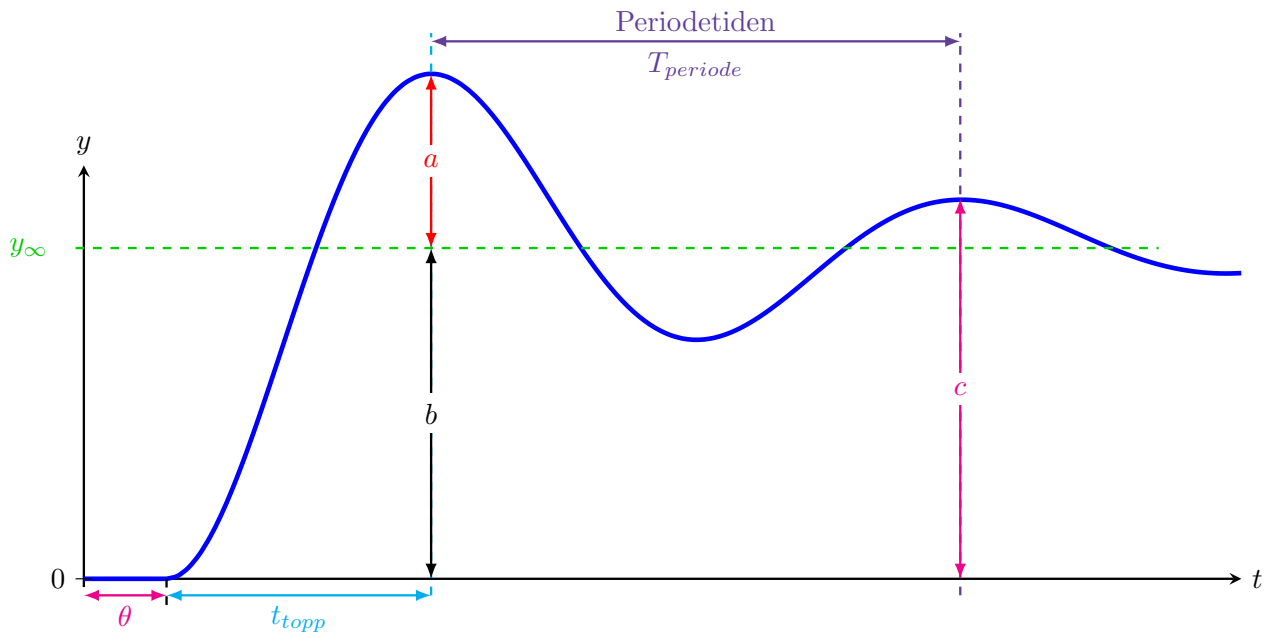
$$G_{AOPTF}(s) = \frac{ke^{-\theta s}}{\tau^2 s^2 + 2\zeta\tau s + 1}$$

med respons som vist i figur 4.8. Fra løsningen til et andre-ordens system (se § 4.1.2) så har vi

$$\begin{aligned}
 t_{topp} &= \pi\tau / \sqrt{1 - \zeta^2} && \text{(Tid til første topp)} \\
 a/b &= \exp\left(-\pi\zeta / \sqrt{1 - \zeta^2}\right) && \text{(Oversving)} \\
 c/a &= (a/s)^2 = \exp(-2\pi\zeta(\sqrt{1 - \zeta^2})) && \text{(Dempningsrate)} \\
 T_{periode} &= \frac{2\pi\tau}{\sqrt{1 - \zeta^2}} && \text{(Periodetid)}
 \end{aligned}$$

Vi skal nå bruke disse verdiene til å tilpasse en AOPTF-modell fra en slik sprangrespons.

Prosedyre:



Figur 4.8: Karakteristikker ved en underdempet sprangrespons

1. Ta utgangspunkt i en enkel sprangrespons (u fra u_0 til u_s) i åpen sløyfe.
2. Tilpass (AOPTF) til responsen:
 - (a) Ta θ som den opplagte tidsforsinkelsen du ser.
 - (b) Finn oversingsratioen $O = a/b$ og regn ut $\zeta = \sqrt{\frac{\ln(O)^2}{\pi^2 + \ln(O)^2}}$.
 - (c) Ta $\tau = \frac{T_{periode} \sqrt{1 - \zeta^2}}{\pi}$.
 - (d) ta $k = \Delta y / \Delta u = (y_\infty - y_0) / (u_s - u_0)$.

4.3.4 Smiths metode for tilpasning av AOPTF-modeller



Alternative kilder: §7.2.1 i [Seborg et al., 2016].

Vi vil nå tilpasse en andreordens-pluss-tidsforsinkelse (AOPTF) modell vha. av Smith metode. Det vil si, vi vil tilpasse en respons til en modell på formen

$$G_{AOPTF}(s) = \frac{ke^{-\theta s}}{\tau^2 s^2 + 2\zeta\tau s + 1} = \frac{ke^{-\theta s}}{(1 + \tau_1 s)(1 + \tau_2 s)} = \frac{k\omega_0^2 e^{-\theta s}}{s^2 + 2\zeta\omega_0 s + \omega_0^2} \quad (\text{AOPTF})$$

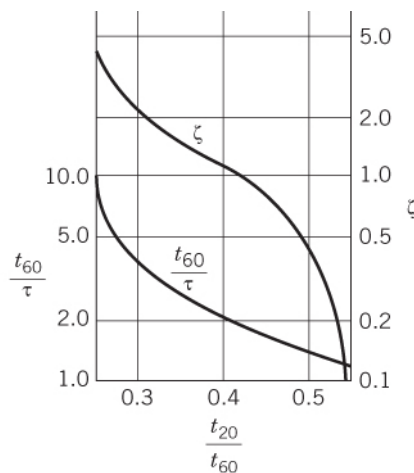
slik at $1/\tau = \omega_0$, hvor ω_0 er dempede svingefrekvensen og ζ er den relative dempningsfaktoren.

Smiths metode – Prosedyre:

1. Ta utgangspunkt i en enkel sprangrespons (u fra u_0 til u_s) i åpen sløyfe.

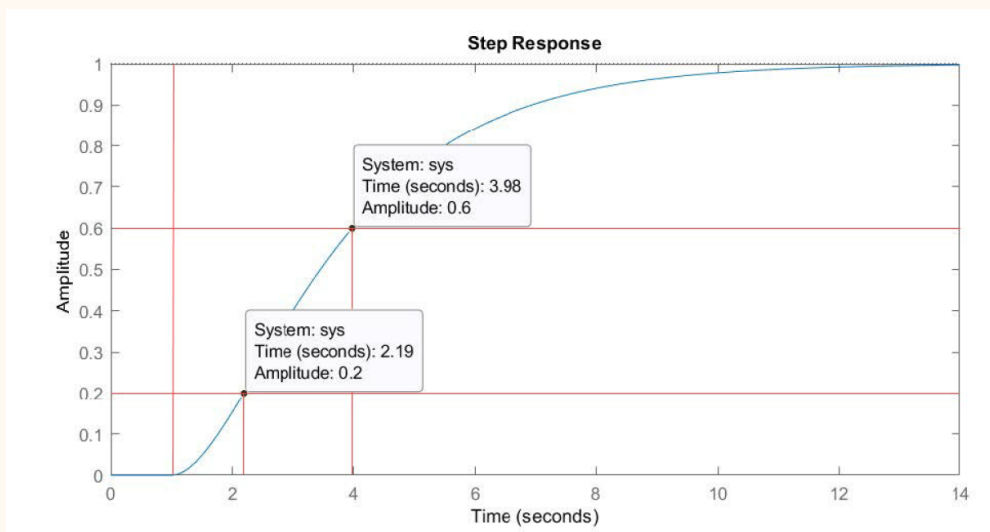
2. Tilpass (AOPTF) til responsen:

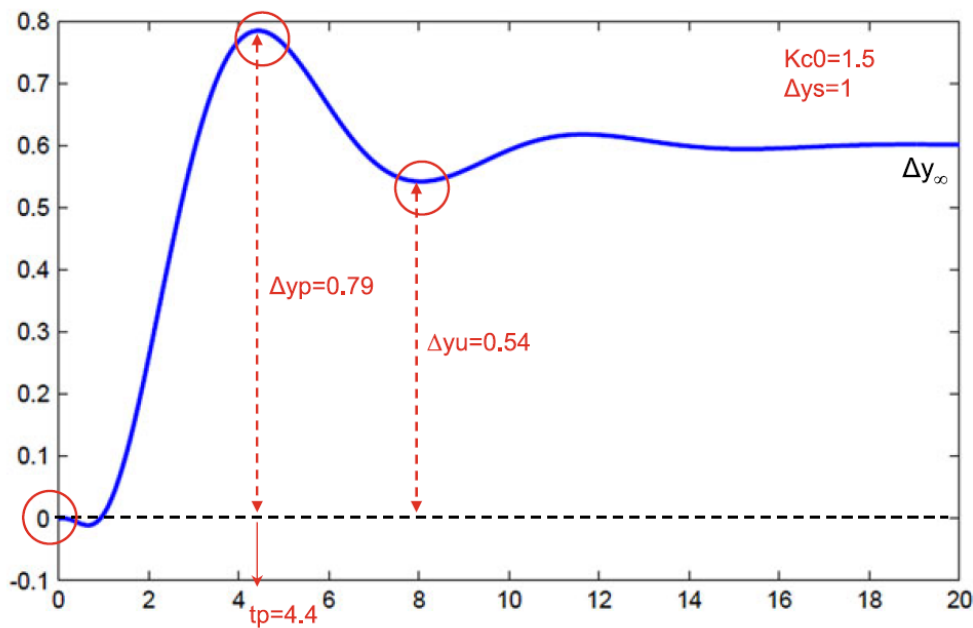
- (a) Ta θ som den opplagte tidsforsinkelsen du ser.
- (b) La t_{20} og t_{60} tidene målt fra etter tidsforsinkelsen til hvor responsen har nådd henholdsvis 20% og 60% av y_∞ relativt til startverdien y_0 ;
- (c) Finn τ og ζ fra figur 4.9;
- (d) ta $k = \Delta y / \Delta u = (y_\infty - y_0) / (u_s - u_0)$.
- (e) hvis $\zeta \geq 1$, ta $\tau_1 = \tau\zeta + \tau\sqrt{\zeta^2 - 1}$ og $\tau_2 = \tau\zeta - \tau\sqrt{\zeta^2 - 1}$.



Figur 4.9: Smiths metode: forholdet mellom (ζ, τ) og (t_{20}, t_{60}) ; figur fra [Seborg et al., 2016].

Eksempel 4.5. Gitt responsen i figuren under. Vi har $t_{20} = 1,19$, $t_{60} = 2,98$, slik at $t_{20}/t_{60} \approx 0,4$. Fra figur 4.9 har vi dermed at $\zeta \approx 1.1$ og $t_{60}\omega_0 = 2,25$. og dermed $\omega_0 \approx 0,76$.





Figur 4.10: Informasjon fra sprangrespons under lukket sløyfe med kun P-regulator for Skogestads modelltilpassningsmetode; figur fra [Skogestad and Grimholt, 2012].

4.3.5 *Skogestads lukkede-sløyfe-metode*



Alternative kilder: [Shamsuzzoha and Skogestad, 2010] og [Skogestad and Grimholt, 2012]

Det er ikke alltid mulig/lett å få en god respons egnet til modelltilpasning fra systemet i åpen sløyfe. I slike tilfeller kan Skogestads lukkede-sløyfe-metode være en mulighet.

Skogestads lukkede-sløyfe-metode er lik Ziegler-Nichols metoden (spesielt hvis den er kombinert med SIMC-regler (FOPTF)), men har den fordel at den ikke at man ikke trenger stående svingninger, men at istedet noe oversving er godt nok.

Lag egne figurer.

Prosedyre for Skogestads lukkede-sløyfe-metode:

1. Kjør systemet i lukket sløyfe med kun proporsjonal-regulator. Øk proporsjonal-forsterkningen k_{p0} til du får en respons lik den i figur 4.10;
2. Med utgangspunkt i figur 4.10, finn:
 - k_{p0} = proporsjonal-forsterkningen som er brukt (K_{c0} i figuren);
 - Δy_s = settpunktsendringen;
 - t_p = tid fra settpunktsendring til toppen av første oversving;
 - Δy_p = maksimal endring i utgangen;
 - Δy_u = endring ved første undersving;

- Merk: Hvis man ikke gidder å vente på at transienten tar slutt for å finne Δy_∞ , kan man bruke følgende estimat av det endelige statiske avviket: $\Delta y_\infty \approx 0.45(\Delta y_p + \Delta y_u)$

3. Regn ut:

- $\delta = \frac{\Delta y_p - \Delta y_\infty}{y_\infty}$ (oversving)
- $\beta = \left| \frac{\Delta y_s - \Delta y_\infty}{\Delta y_\infty} \right|$ (statisk avvik)
- $\alpha = 1.152\delta^2 - 1.607\delta + 1$
- $\rho = \alpha/\beta$

4. Ta $k = 1/(k_{p0}\beta)$, $\theta = t_p(0.309 + 0.209e^{-0.61\rho})$, og $\tau = \rho\theta$ i (FOPTF).

5. Simulering

Alternative kilder: Kap. 7 i [Haugen, 2023]

Hvorfor er simulering viktig for reguleringsteknikk? Muligheten til å numerisk simulere et dynamisk system er nyttig på mange måter. Uten fare for både helse, utstyr og omgivelser, kan man bruke simulering til å blant annet:

- teste ut både reguleringsstrategier og -parametere;
- undersøke sensitivitet og robusthet;
- få et estimat av ytelse og effekt.

Fun facts, bemerkninger og annet dill dall (you may skip)

Digitale tvillinger (se f.eks. [Sharma et al., 2022]) er en av de store “buzzword”-teknologiene i nyere tid. Den essensielle ideen bak digitale tvillinger er å sette opp en virtuell, simulert “dobbelgjenger/tvilling” av en virkelig prosess. Enkle målinger av den fysiske prosessen overføres så til simuleringen, hvor disse dataene brukes til å danne grenseverdier og startbetingelsene til simuleringsmodellen. Tidligere ukjent tilstandsinformasjon genereres gjennom simulering og blir deretter levert tilbake til enhetens reguleringsystem. Det er denne toveis datautveksling mellom den virkelige prosessen og dens digital motpart (simuleringen) i sanntid som kjennetegner denne teknologien.

For at dette skal kunne tas i bruk under drift (“in real time”), så innebærer dette naturlig nok at man må ha simuleringsresultater på den virtuelle siden som er minst like raske som den naturlige prosessen (“real time simulation”).

Fun facts, bemerkninger og annet dill dall (you may skip)

Sim2Real er et konsept som i disse dager er mye brukt innen maskinlæringsgrenen **reinforcement learning**. Ved hjelp av simulatorer kan man trene robotene i stor skala, på en rekke scenarier og forhold som sjelden oppstår i den virkelige verden. Man tar så i bruk

(sjeldent helt direkte) det som har blitt lært (hvilke handlinger som skal gjøres gitt sensor-data, og muligens tidligere tilstander og målinger) i den virkelige verden. Eksempler på dette inkluderer: Autonome kjøretøy (bl.a. Tesla), [robothender som løser rubiks kube](#), firbeinte roboter som tar seg en fjelltur ([ANYmal](#)), og [roboter som spiller bordtennis](#), etc.

5.1. Numerisk integrasjon og Runge–Kutta-metoder

Alternative kilder: [Wikipedia](#).

5.1.1 Hva er numerisk integrasjon?



For å kunne simulere et dynamiske system, må vi på en måte «løse» differensialligningene som forklarer systemets dynamikk. Generelt sett (et unntak er lineære systemer, som alltid kan løses analytisk) må dette gjøres ved hjelp av **numeriske metoder**. Mer spesifikt, så ønsker vi å finne en løsning—på formen av en (tidsvarierende) funksjon $x(t)$ —til et *initialverdiproblem* (IVP) :

Initialverdiproblem (IVP): Finn $\mathbf{x}(t)$ for $0 \leq t \leq T$ gitt

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t), t), \quad \mathbf{x}(0) = \mathbf{x}_0 \in \mathbb{R}^n. \quad (\text{IVP})$$

En løsning til IVPen må dermed tilfredsstill:

- 1) startbetingelsene (initialverdien) $\mathbf{x}(0) = \mathbf{x}_0$;
- 2) ODEen (differensialligning) $\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t), t)$ for alle tidspunkt i tidsintervallet $[0, T]$.

Merk: Selv om vi her ikke direkte tar hensyn til noe kontrollsignal $u(t)$ (ei heller forstyrrelse $d(t)$), kan man, hvis nødvendig, anta at dette er en del av funksjonen $\mathbf{f}(\cdot)$.

Hvorfor trenger vi numerisk integrasjon? Ved å integrere begge sider av (IVP) med hensyn på tid, så er det klart at en hver løsning på initialverdiproblemet må tilfredstille

$$\mathbf{x}(t) = \mathbf{x}_0 + \int_0^t \mathbf{f}(\mathbf{x}(\tau), \tau) d\tau.$$

Problem: vi kan bare i spesielle tilfeller finne en **analytisk løsning** på integralet $\int_0^t \mathbf{f}(\mathbf{x}(\tau), \tau) d\tau$, altså en løsning som kan uttrykkes eksplisitt ved hjelp av en bregrenset mengde av elementære funksjoner (altså summer og produkter av konstanter, polynomer, trigonometriske-, logaritmiske og eksponentielle funksjoner, etc).

Eksempler:

- $\dot{x}(t) = \cos(t)$, $x(0) = 0$, har løsningen $\int_0^t \cos(\tau) d\tau = \sin(t)$;

- $\dot{x}(t) = \cos(t^2)$, $x(0) = 0$, har ikke en analytisk løsning (fordi $\int_0^t \cos(\tau^2) dt$ ikke har det);
- $\dot{x}(t) = -2 \cdot x(t)$, $x(0) = 3$, har løsningen $x(t) = 3e^{-2t}$;
- $\ddot{x}(t) = -2 \cdot \sin(x(t))$, $x(0) = 3$ og $\dot{x}(0) = 0$, har ikke en analytisk løsning.

NB! Siste eksempelet har samme form som dynamikken til en enkel pendel (se eks. 3.11), så selv om det ikke finnes en *analytisk* løsning, så finnes det jo da en løsning.

Numerisk integrasjon: Finne en best mulig *approximasjon* av en løsning ved hjelp av numeriske metoder. Det er dog viktig at vi bruker metoder som tar høyde for at funksjonen $f(\cdot)$ kan avhenge av løsningen $\mathbf{x}(t)$ (som vi jo ikke vet hva er enda).

Vi skal se på noen slike metoder i den neste seksjonen, spesifikt noen såkalte *Runge–Kutta-metoder*, deriblant Eulers metoder og trapesmetoden.¹ Vi skal også se på hvordan Runge–Kutta-metoder kan videreutvikles slik at man kan bruke varierende steglengder for å øke nøyaktigheten og samtidig kunne løse ligningene relativt fort. I slutten av kapitlet skal vi også se på hvordan man kan bruke MATLAB og Simulink til å simulere dynamiske systemer ved hjelp av slike metoder.

5.1.2 Numerisk integrasjon uten tilstander

▶ V2KbyLPVq_g&t=413

La oss først se på hvordan man numerisk kan integrere en funksjon som bare avhenger av tiden t og ikke tilstandene \mathbf{x} , altså:

Nåværende problem: Finn en tilnærmet verdi (en approximasjon) av integralet

$$\int_a^b f(t) dt.$$

La oss dele tidsintervallet $[a, b]$ og i N deler av lengde \mathfrak{h} ,² altså

$$a = t_0 < t_1 < t_2 < \dots < t_{N-1} < t_N = b$$

hvor, for alle $k = 0, 1, 2, \dots, N - 1$ (evt. kunne vi her ha skrevet $\forall k = 0, 1, \dots$),

$$\mathfrak{h} = t_{k+1} - t_k.$$

Vi kan da dele integralet opp som følger:

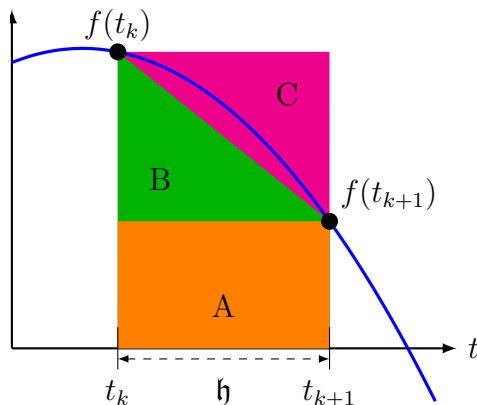
$$\int_a^b f(t) dt = \int_{t_0}^{t_0+\mathfrak{h}} f(t) dt + \int_{t_1}^{t_1+\mathfrak{h}} f(t) dt + \dots + \int_{t_{N-1}}^{t_{N-1}+\mathfrak{h}} f(t) dt = \sum_{k=0}^{N-1} \int_{t_k}^{t_k+\mathfrak{h}} f(t) dt.$$

Hvis \mathfrak{h} er liten nok, kan vi approksimere hvert av disse integralene ved én av følgende metoder (disse er grafisk illustrert i figur 5.1 for en skalar funksjon):

- **Eulers fremovermetode:** $\int_{t_k}^{t_k+\mathfrak{h}} f(t) dt \approx \mathfrak{h} \cdot f(t_k) = \mathbf{A} + \mathbf{B} + \mathbf{C}$;

¹Det finnes også en annen familie av metoder, såkalte *multistegs-metoder*, som vi ikke skal se på.

²Bruker gotisk “ h ” for å skille fra alle andre h -er i disse notatene. Alternativt kan man bruke Δt .



Figur 5.1: Areal som viser numeriske approksimasjoner av integralet $\int_{t_k}^{t_{k+1}} f(\tau) d\tau$: Eulers framover metode= $A+B+C=A+2B$; Eulers bakover metode= A ; Trapes metoden = $A+B$.

- **Eulers bakovermetode:** $\int_{t_k}^{t_k+h} f(t) dt \approx h \cdot f(t_k + h) = h \cdot f(t_{k+1}) = A$;
- **Trapesmetoden:** $\int_{t_k}^{t_k+h} f(t) dt \approx \frac{1}{2}h \cdot (f(t_k) + f(t_k + h)) = A+B$.

Legg merk til at Trapesmetoden er gjennomsnittet av de to første metodene.

5.1.3 Numerisk integrasjon med tilstander



Nåværende problem: For et IVP,

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t), t), \quad \mathbf{x}(0) = \mathbf{x}_0 \in \mathbb{R}^n, \tag{IVP}$$

finn en tilnærming av løsningen $\mathbf{x}(t)$ for $0 \leq t \leq T$, gitt ved

$$\mathbf{x}(t) = \mathbf{x}_0 + \int_0^t \mathbf{f}(\mathbf{x}(\tau), \tau) d\tau.$$

Antagelser og notasjon: Vi vil anta at $\mathbf{f}(\cdot)$ er en glatt funksjon, altså at den kan deriveres uendelig mange ganger. Vil vil også anta en konstant **steglengde** $h > 0$.

Vi vil bruke notasjonene $t_k = k \cdot h$, $\mathbf{x}(0) = \mathbf{x}_0$, og $\mathbf{x}[k] \approx \mathbf{x}(t_k)$, $k = 0, 1, \dots$, det vil si

$$\mathbf{x}[k + 1] \approx \mathbf{x}(t_k + h) = \mathbf{x}(t_k) + \int_{t_k}^{t_k+h} \mathbf{f}(\mathbf{x}(\tau), \tau) d\tau. \tag{5.1}$$

⚠ I år har notasjonen blitt endret fra $x_k \approx x(t_k)$ til $x[k] \approx x(t_k)$. Det kan derfor hende at x_k dukker opp i setdet for $x[k]$ noen steder.

I den følgende seksjonen skal vi se på såkalte *Runge–Kutta-metoder* for å approksimere en løsning til initialverdiproblemet (IVP). Vi begynner med de enkleste metodene, nemlig Euler metodene og trapesmetoden, som vi allerede har sett på for funksjoner som ikke avhenger av tilstandene.

Eulers (eksplisitte) forovermetode V2KbyLPVq_g&t=1041

Den enkleste metoden er Eulers (eksplisitte³) fremover metode:

Eulers fremovermetode:

$$\mathbf{x}[k + 1] = \mathbf{x}[k] + \mathfrak{h} \cdot \mathbf{f}(\mathbf{x}[k], t_k). \quad (5.2)$$

Idé: hvis $\mathfrak{h} > 0$ er liten nok, så er $\mathfrak{h}f(\mathbf{x}[k], t_k) \approx \int_{t_k}^{t_k+\mathfrak{h}} \mathbf{f}(x(\tau)) \cdot \tau d\tau$ en grei tilnærming. Vi finner altså $\mathbf{x}[k + 1]$ ved å gå én lengde $\mathfrak{h} \cdot \|\mathbf{f}(\mathbf{x}[k], t_k)\|$ fra $\mathbf{x}[k]$ i retningen til vektoren $\mathbf{f}(\mathbf{x}[k], t_k)$.

Fordeler:

- Rask, enkel og lett å implementere.

Ulemper:

- Unøyaktig (selv for små \mathfrak{h});
- Numerisk ustabil hvis \mathfrak{h} ikke er liten nok.

Nøyaktighet: Lokalt: $O(\mathfrak{h}^2)$; Globalt: $O(\mathfrak{h})$ (ganger med antall segment).

Numerisk ustabil betyr her at differansen mellom estimatet funnet fra Eulers metode og den ekte løsningen til differensialligningen øker, noe som kan resultere i at estimatets magnitudo går mot uendelig selv om den ekte løsningen holder.

Følgende eksempel fra [Wikipedia](#) illustrerer dette fenomenet:

Eksempel 5.1. Gitt følgende IVP:

$$\dot{x} = -3x, \quad x(0) = 1.$$

Vi vet at den ekte løsningen er $x(t) = e^{-3t}$, slik at $x(t) \rightarrow 0$ når $t \rightarrow \infty$. Lar vi $\mathfrak{h} = 1$, så får vi

$$\mathbf{x}[k + 1] = (1 - 3)\mathbf{x}[k] = -2\mathbf{x}[k]$$

fra Eulers metode. Dette viser tydelig at $|\mathbf{x}[k]| \rightarrow \infty$ når $k \rightarrow \infty$ ($\mathbf{x}[k + 1]$ har dobbelt så stor magnitudo i forhold til $\mathbf{x}[k]$). Tar vi derimot $\mathfrak{h} < 2/3$ vil også $\mathbf{x}[k]$ konvergere til null.

([Videolenke](#); mulig eksamensoppgave 2024? Husk å FJERNE!)

Eulers implisitte bakovermetode V2KbyLPVq_g&t=1339

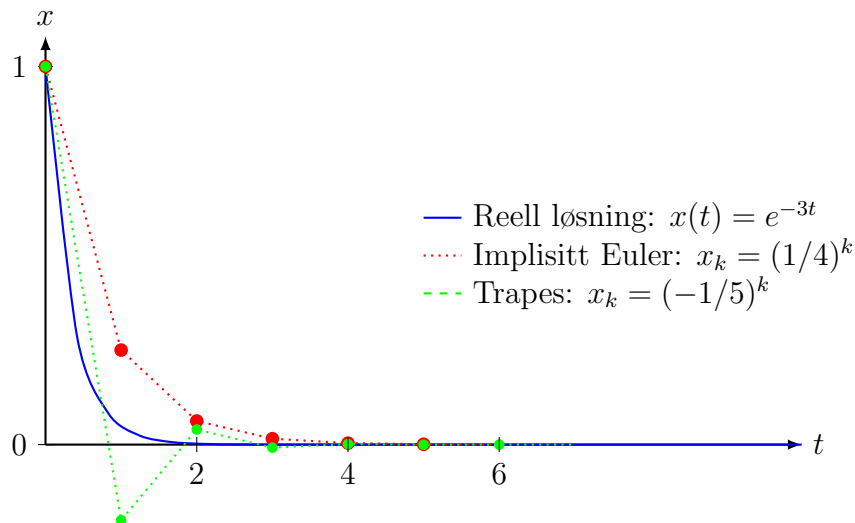
En annen metode, om enn hakket mer kompleks, med bedre numerisk stabilitet er Eulers (implisitte⁴) bakover metode:

Eulers bakovermetode:

$$\mathbf{x}[k + 1] = \mathbf{x}[k] + \mathfrak{h} \cdot \mathbf{f}(\mathbf{x}[k + 1], t_{k+1}). \quad (5.3)$$

³Metoden kalles også for den eksplisitte metoden siden høyresiden av (5.2) kun avhenger av $\mathbf{x}[k]$, ikke $\mathbf{x}[k + 1]$.

⁴Metoden kalles også for den implisitte metoden siden høyresiden av (5.3) også avhenger av $\mathbf{x}[k + 1]$.



Figur 5.2: Sammenligning av Euler (implisitte) bakovermetode og trapesmetoden for systemet $\dot{x} = -3x$, med $x_0 = \eta = 1$. Merk: Eulers bakover metode har en “overdempet” respons, mens trapesmetoden har en slags “underdempet” respons for dette eksempelet. Eulers forovermetode er ikke vist siden denne er ustabil for dette eksempelet.

Idé: I motsetning til fremovermetoden bruker man tangentvektoren $\mathbf{f}(\mathbf{x}[k+1], t_{k+1})$ i stedet for $\mathbf{f}(\mathbf{x}[k], t_k)$, noe som gjør det lett å regne ut $\mathbf{x}[k]$ gitt $\mathbf{x}[k+1]$, altså *bakover* i tid.

Fordeler:

- Relativt lett å forstå;
- Numerisk stabil for alle η (såkalt **A-stabil**)!

Ulemper:

- Unøyaktig (selv for små η);
- Må løse den implisitte ligningen (5.3) ($\mathbf{x}[k+1]$ på begge sider).

Nøyaktighet: Lokalt: $O(\eta^2)$; Globalt: $O(\eta)$ (ganger med antall segment).

Achtung! Eulermetodene introdusere noe man kaller for *numerisk dempning* (denne kan være både positiv eller negativ). Dette kan føre til at et (marginalt) stabilt system ser ustabil ut når vi simulerer det, eller vice versa. Dere vil se nærmere på dette i en av øvingene.

Eksempel 5.2. (Eulers bakover-metode med MATLABs fsolve) Mål: løse $\dot{x} = -\sin(x)$, $x(0) = 1$, vha. av Eulers bakovermetode (se (5.3)) i MATLAB med tidssteg $\eta = 0.1$. Ligningen fra bakover-Euler er $x_{k+1} = x_k + h \cdot (-\sin(x_{k+1}))$, som igjen tilsvarer $F(x_{k+1}, x_k) = x_{k+1} - x_k + h \cdot \sin(x_{k+1}) = 0$.

En enkel måte å løse dette på i MATLAB er vist i kodesnutt 5.1. Merk at $\mathbf{g}=\mathcal{O}(t,x) - \sin(x)*t$ betyr at \mathbf{g} er en funksjon av variablene t og x lik $-\sin(x)$, men $\mathbf{h}=\mathcal{O}(x)\mathbf{g}(2,x)$ gir en ny funksjon h som bare har variabelen x og tilsvarer da \mathbf{g} ved når $t = 2$, altså er h lik

funksjonen $@(x)-2*\sin(x)$.

Kodesnutt 5.1: Løse IVP med én tilstand i MATLAB vha. Eulers implisitte metode og fsolve.

```

h      = 0.1;           % Tidssteg
x0     = 1;           % Initialverdi
tEnd   = 10;         % Simuleringstid
N      = tEnd/h;     % Antall steg
t      = linspace(0, tEnd, N+1); % Liste med tidspunkt
x=zeros(N+1,1);      % Liste hvor vi skal lagre tilstandene
x(1)=x0;
% Systemets dynamikk:
f      = @(t, x) -sin(x);
% Funksjon tilsvarende ligning fra Implisitt Euler (z=x_{k+1}):
F      = @(t, z, xk) (z-xk-h*f(t, z));
opt    = optimset('Display', 'off', 'TolFun', 1e-8); % Til fsolve
senere
% For-lokke
for k=1:N
    x(k+1)= fsolve(@(z)F(t(k), z, x(k)), x(k), opt);
end
[tt, xt] =ode45(f, [0, tEnd], x0); % ODE45 til sammenligning
% Lage figur
figure(1); clf(1)
hold on
plot(t, x)
plot(tt, xt, 'r')
xlabel('Tid')
legend({'Implisitt (bakover) Euler', 'ODE45'}, 'Location', 'best')
```

Trapesmetoden



En mellomting mellom Eulers fremover- og bakover-metode:

Trapesmetoden:

$$\mathbf{x}[k+1] = \mathbf{x}[k] + \frac{h}{2} \cdot (\mathbf{f}(\mathbf{x}[k], t_k) + \mathbf{f}(\mathbf{x}[k+1], t_{k+1})). \quad (5.4)$$

Idé: Verdien til $\mathbf{x}[k+1]$ er gjennomsnittet av det man fikk fra Eulers fremover- og bakover metoder. Dette betyr at dette også er en implisitt metode.

Fordeler::

- OK-ish lett å forstå;
- Numerisk stabil for alle h (A-stabil).

Ulemper::

- Ikke veldig nøyaktig (selv for små h), men bedre enn Euler-metodene;
- Må løse den implisitte ligningen (5.4) ($\mathbf{x}[k + 1]$ på begge sider).

Nøyaktighet: Lokalt: $O(h^3)$; Globalt: $O(h^2)$ (ganger med antall segment).

Figur 5.2 viser enn sammeligning mellom trapesmetoden funnet via Newtons metode og Eulers baklengsmetode for systemet i eksempel 5.3.

Løse implisitte ligninger ved hjelp av Newtons metode



Hvordan kan vi (generelt sett) løse de implisitte ligningene (5.3) og (5.4) med hensyn på $\mathbf{x}[k + 1]$? En måte er Newtons metode for å finne nullpunkter til en funksjon.⁵

Nåværende problem: Finne et punkt \mathbf{z}_* slik at $\mathbf{F}(\mathbf{z}_*) = 0$, hvor $\mathbf{F} : \mathbb{R}^n \rightarrow \mathbb{R}^n$, $\mathbf{z} \mapsto \mathbf{F}(\mathbf{z})$, er en differensierbar funksjon som har en Jacobi-matrise^a, $\mathbf{J}_F(\mathbf{z}) = \frac{\partial \mathbf{F}}{\partial \mathbf{z}}(\mathbf{z})$, som er invertierbar for alle $\mathbf{z} \in \mathbb{R}^n$ nært \mathbf{z}_* .

^aHusk: Hvis $\mathbf{F}(\mathbf{z}) = [F_1(\mathbf{z}), F_2(\mathbf{z}), \dots, F_n(\mathbf{z})]^T$, hvor $\mathbf{z} = [z_1, z_2, \dots, z_n]^T$, så $\mathbf{J}_F(\mathbf{z}) = \begin{bmatrix} \frac{\partial F_1}{\partial z_1} & \dots & \frac{\partial F_1}{\partial z_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial F_n}{\partial z_1} & \dots & \frac{\partial F_n}{\partial z_n} \end{bmatrix}$.

Eksempler på slike funksjoner:

- For $F(z) = z + z^3$ har vi $F : \mathbb{R} \rightarrow \mathbb{R}$ og $J_F(z) = 1 + 3z^2$, med $z_* = 0$;
- For $\mathbf{F}(\mathbf{z}) = \begin{bmatrix} z_1 + z_2 + 1 \\ z_2 \end{bmatrix}$ har vi $\mathbf{F} : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ og $\mathbf{J}_F(\mathbf{z}) = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$, med $\mathbf{z}_* = \begin{bmatrix} -1 \\ 0 \end{bmatrix}$.

Newton metode for å finne et nullpunkt for en funksjon: Gjett \mathbf{z}_0 og sett

$$\mathbf{z}_{p+1} = \mathbf{z}_p - (\mathbf{J}_F(\mathbf{z}_p))^{-1} \mathbf{F}(\mathbf{z}_p). \tag{5.5}$$

inntil $\|\mathbf{F}(\mathbf{z}_{p+1})\| < \epsilon$ (evt. $\|\mathbf{z}_{p+1} - \mathbf{z}_p\| < \epsilon$) for en ønsket (veldig liten) toleranse $\epsilon > 0$.^a

^aHer er $\|\cdot\|$ den Euklidiske normen. Eks.: Gitt en vektor \mathbf{z} i \mathbb{R}^2 , $\mathbf{z} = [z_1, z_2]^T$, så er $\|\mathbf{z}\| = \sqrt{z_1^2 + z_2^2}$.

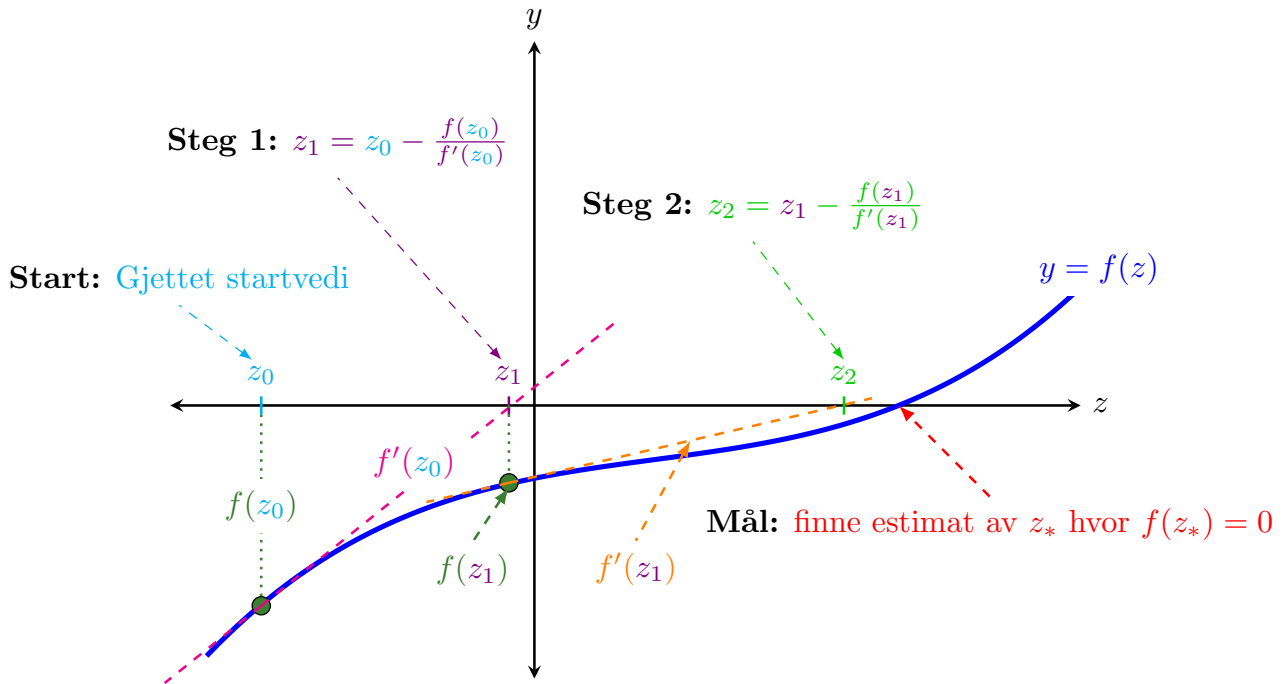
Alternativ: Det går an å i stede bruke MATLABs (noe tregere) `fsolve` funksjon; se [lenke](#).

Figur 5.3 viser de to første stegene av Newtons metode for å finne nullpunktet $z_* = 2$ til funksjonen $f(z) = (z - 2)(z^2 + 4)$. med startverdi $z_0 = 3$.

La oss bruke Newtons metode i forbindelse med Eulers bakovermetode for systemet i eksempel 5.1:

Eksempel 5.3. La $\dot{x} = -3x$, $x_0 = 1$ og $h = 1$. Fra (5.3) har vi $\mathbf{x}[k + 1] = \mathbf{x}[k] - 3\mathbf{x}[k + 1]$, som igjen tilsvarer $4\mathbf{x}[k + 1] - \mathbf{x}[k] = 0$. For en gitt $\mathbf{x}[k]$, la os derfor definere $F_k(z) := 4z - \mathbf{x}[k]$, slik at vi for hvert steg så ønsker vi å finne en $\mathbf{x}[k + 1] = z$ slik at $F_k(z) \approx 0$.

⁵Newtons metode, også kalt Newton–Rapshon metoden, er det en fordel å ha kjennskap til siden den har mange bruksområder. Den er for eksempel veldig viktig innen numerisk optimalisering; se .



Figur 5.3: Illustrasjon av to steg av Newtons metode for å finne estimat av et nullpunkt z_* til en skalar funksjon $f(z)$.

Newton's metode (5.5) gir følgende iterasjon:

$$z_{p+1} = z_p - \frac{4z_p - \mathbf{x}[k]}{4} = \frac{\mathbf{x}[k]}{4},$$

og dermed $\mathbf{x}[k + 1] = \mathbf{x}[k]/4 = (1/4)^k$. Vi har derfor at $\mathbf{x}[k] \rightarrow 0$ når $k \rightarrow \infty$, akkurat som den reelle løsningen $x(t) = e^{-3t}$ (se fig. 5.2).

Eksempel 5.4. (Simulere en pendel vha. Eulers bakovermetode) Gitt følgende IVP:

$$\dot{\mathbf{x}} = \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \mathbf{f}(\mathbf{x}) = \begin{bmatrix} x_2 \\ -4 \sin(x_1) \end{bmatrix}, \quad \mathbf{x}(0) = \begin{bmatrix} x_1(0) \\ x_2(0) \end{bmatrix} = \begin{bmatrix} \pi/2 \\ 0 \end{bmatrix}.$$

Mål: Simulere systemet vha. Eulers bakovermetode (EBM) med Newtons metode (NM).
Prosedyre:

- **Steg 1.** (Sett opp EBM) $\mathbf{x}[k + 1] = \mathbf{x}[k] + \mathfrak{h} \cdot \mathbf{f}(\mathbf{x}[k + 1])$
- **Steg 2.** (Konstruer vektor-funksjon for NM) $\mathbf{F}(\mathbf{z}) = \mathbf{z} - \mathbf{x}[k] - \mathfrak{h} \cdot \mathbf{f}(\mathbf{z}) = \begin{bmatrix} z_1 - x_k^1 - \mathfrak{h} \cdot z_2 \\ z_2 - x_k^2 + 4\mathfrak{h} \cdot \sin(z_1) \end{bmatrix}$
- **Steg 3.** (Finn dens Jacobimatrise) $\mathbf{J}(\mathbf{z}) = \begin{bmatrix} 1 & -\mathfrak{h} \\ 4\mathfrak{h} \cdot \cos(z_1) & 1 \end{bmatrix}$

- **Steg 4.** (Finn dens invers) $\mathbf{J}(\mathbf{z})^{-1} = \frac{1}{1+4\mathfrak{h}^2 \cos(z_1)} \begin{bmatrix} 1 & \mathfrak{h} \\ -4\mathfrak{h} \cdot \cos(z_1) & 1 \end{bmatrix}$
- **Steg 5.** (Velg tidssteg) Ta $\mathfrak{h} < \frac{1}{2}$ (bare sånn i tilfelle $z_1 = \pi$)
- **Steg 6.** (Velg initialbetingelse for NM) Ta f.eks. $\mathbf{z}_0 = \mathbf{x}_k$ ($\mathbf{x}_0 = [\pi/2, 0]^T$)
- **Steg 7.** (Kjør NM for hver iterasjon) $\mathbf{z}_{p+1} = \mathbf{z}_p - (\mathbf{J}_{\mathbf{F}}(\mathbf{z}_p))^{-1} \mathbf{F}(\mathbf{z}_p)$
- **Steg 8.** Ta $\mathbf{x}[l+1] = \mathbf{z}_{p+1}$ når $\|\mathbf{F}(\mathbf{z}_{p+1})\| < \epsilon$ for et lite, positivt tall (toleranse) ϵ
- **Steg 9.** Ta $k = k + 1$ og gå tilbake til steg 6 (gitt at ikke $t_k = k \cdot \mathfrak{h} = T$).

5.1.4 *Høyere ordens Runge–Kutta-metoder *

Husk vårt mål: å finne en tilnærmet løsning på initialverdiproblem $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, t)$, $\mathbf{x}(0) = \mathbf{x}_0$. Både metodene vi har sett på så langt (Eulers fremover- og bakover-metoder, samt trapesmetden) og kjente ODE-løsere som MATLABs `ode45` er (bygd opp av) såkalte *Runge–Kutta-metoder*. Disse metodene kan deles opp i deres orden (høyere orden gir høyere nøyaktighet), og om de er eksplisitte eller implisitte (som `ode45`, som faktisk består av to slike metoder, kan de også være adaptive / ha variable steglenger, mer om dette senere). For eksempel:

- Eulers fremovermetode er en første-ordens eksplisitt metode;
- Eulers bakovermetode er en første-ordens implisitt metode;
- Trapesmetoden er en andre-ordens implisitt metode;
- Dorman–Prince-metoden (`ode45`) består av både en fjerde- og en femte-ordens eksplisitt metode som kjøres i parallell.

Hvordan ser en Runge–Kutta-metode ut? For en gitt steglengde $\mathfrak{h} > 0$, samt et sett med *vekter*, $w_1, w_2, \dots, w_p \in \mathbb{R}$, og *noder*, $0 \leq \mu_1 < \mu_2 < \dots < \mu_p \leq 1$, $p \geq 1$, la oss anta at

$$\mathbf{x}[k+1] = \mathbf{x}[k] + \mathfrak{h} \sum_{i=1}^p w_i \mathbf{f}(\mathbf{x}(t + \mu_i \mathfrak{h}), t + \mu_i \mathfrak{h}), \quad (5.6)$$

er en god approksimasjon av høyresiden til (5.1). Her er da $\mathbf{x}[0], \mathbf{x}[1], \mathbf{x}[2], \dots$, tilnærmingen av $x(t)$ ved tidspunktene $t_0 = 0$, $t_1 = \mathfrak{h}$, $t_2 = 2\mathfrak{h}$, etc.

Fra et praktisk perspektiv, så er det dog et problem med denne tilnærmingen: vi vet jo ikke hva $x(t + \mu_i \mathfrak{h})$ er!

Hva kategoriserer en slik metode? Som tidligere nevnt, så er både Eulers fremover- og bakover-metode, henholdsvis (5.2) og (5.3), samt trapesmetoden (5.4), såkalte Runge–Kutta-metoder. Vi har dog enda ikke sagt hva som gjør disse til Runge–Kutta-metoder, samt motivasjonen og utledningen for slike metoder. Vi skal derfor nå prøve å gi en kort og, forhåpentligvis, en sånn halveis intuitiv forklaring på disse metodene.

Vi ønsker nå å modifisere (5.6) på en slik måte at høyresiden bare avhenger av $\mathbf{x}[k] = x(k\mathfrak{h}) = x(t)$:

$$\mathbf{x}[k+1] = \mathbf{x}[k] + \mathfrak{h} \sum_{i=1}^p b_i \kappa_i \quad \text{hvor} \quad \kappa_i := \mathbf{f}(\mathbf{x}[k] + \mathfrak{h} \sum_{j=1}^l a_{ij} \kappa_j, t + c_i \mathfrak{h}). \quad (5.7)$$

Dette er generelt sett en **implisitt metode**, siden hver κ_i også kan være “avhengig av seg selv”. Ved å ta $\kappa_1 = f(\mathbf{x}[k], t)$ og $l = i - 1$, så får vi dog en **eksplisitt metode**, siden κ_i da avhenger av $\kappa_1, \dots, \kappa_{i-1}$.

For eksempel, tar vi $p = b_1 = 1$ og $c_1 = a_{11} = l = 0$, så får vi Eulers fremovermetode (5.2), som er en første-orden metode siden Taylor-serien av orden én er $\mathbf{x}(t + \mathfrak{h}) \approx \mathbf{x}(t) + \mathbf{f}(x(t), t)\mathfrak{h}$. Tar vi derimot $p = b_1 = 1$ og $c_1 = a_{11} = l = 1$, så får vi Eulers bakovermetode, som også er en første-ordens (en implisitt) metode. Dette kan man se ved å legge merke til at $\mathbf{f}(\mathbf{x}[k+1], t_{k+1}) \approx \mathbf{f}(\mathbf{x}[k], t_k) +$ ledd av høyere orden mtp. \mathfrak{h} .

Merk: En ganske utfyllende liste av Runge–Kutta-metoder finnes på [Wikipedia](#).

Vi kan nå gi definisjon på en Runge–Kutta metode:

Runge–Kutta-metode: En metode på formen (5.7) (se også (5.6)) er en Runge–Kutta-metode av orden p hvis dens Taylor-serie tilsvar $\mathbf{x}(t + \mathfrak{h})$ sin Taylor-serie, altså

$$\mathbf{x}(t + \mathfrak{h}) \approx \mathbf{x}(t) + \sum_{r=1}^{\infty} \frac{1}{r!} \mathbf{x}^{(r)}(t) \mathfrak{h}^r, \quad \text{hvor} \quad \mathbf{x}^{(r)}(t) = \frac{d^r}{dt^r} \mathbf{x}(t),$$

opp til et ledd av orden p .^a

^a**Konklusjoner:** *i*) En Runge–Kutta metode av orden p kan nøyaktig integrere en ODE på formen $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, t)$ hvor $\mathbf{f}(\cdot)$ bare er et polynom av orden $p-1$ i den uavhengige variabelen t , altså $\mathbf{f}(\mathbf{x}, t) = \sum_{i=1}^{p-1} a_i t^i$; *ii*) den lokale trunkeringsfeilen er av orden $\mathcal{O}(\mathfrak{h}^{p+1})$, hvor jeg har brukt [store O-notasjon](#).

5.1.5 Aspekter ved numerisk integrasjon og ting som bør vurderes

 [V2KbyLPVq_g&t=2595](#)

Huskeregler for valg av ODE-metode og -løserer:

- Høyere-orden: høyere nøyaktighet, men også høyere kompleksitet og flere nødvendige steg;
- Eksplisitt: enkle å implementere men dårligere stabilitet;
- Implisitt: mer kompliserte å implementere (trenger f.eks. noe a la Newtons metode) men A-stabile (og derfor egnet til stive ODEer);
- Adaptiv/variabel: tillater en ønsket balansegang mellom hurtighet og ønsket nøyaktighet.

Numerisk integrasjon med variable/adaptive steglengder

Nøyaktigheten til en integrasjonsmetode kan alltid økes ved å korte ned steglengden h . En kortere steglengde fører dog igjen til flere diskretiseringpunkter, og dermed til en lengre integrasjonstid.

Et alternativ som lar en finne et ønsket kompromiss mellom nøyaktighet og integrasjonstid er å ta i bruk adaptive løsere. To vanlige strategier:

1. Ta samme løser og sammenlign resultatet for to forskjellige steglengder; f.eks. h vs $\frac{h}{2}$;
2. Ta to forskjellige løsere å sammenlign deres resultatet, slik at man approksimere feilen.

Eksempel: Bak MATLABs velkjente ODE-løser `ode45` er [Dormand–Prince metoden](#) [Dormand and Prince, 1980]. Denne metoden bruker totalt seks funksjonsevalueringer for å beregne fjerde- og femteordens approksimasjoner. Forskjellen mellom disse brukes som et estimat på feilen til disse approksimasjonene. Dette feilestimatet er veldig praktisk for integrasjonsalgoritmer med adaptive trinnstørrelser.

Stive vs ikke-stive differensialligninger

Intuitivt sett, så er [stive differensialligninger](#) bare differensialligninger som det er vanskelig å integrere numerisk. Eksempler på dette er systemer som innehar både veldig treg og veldig rask dynamikk.

Tips: Bruk implisitte ODE-løsere.

Kaotiske systemer, kaosteori og kontrollering av kaos

Kaotiske dynamiske systemer kjennetegnes av at evolusjonen til systemets tilstander er meget sensitive i forhold til initialbetingelsene. Små feil vil derfor raskt blåses opp (eksponentiell feilpropergering), noe som gjør at man må ta resultatene man får etter en viss tidsperiode med en klype salt. Læren om slike systemer kalles [kaosteori](#).

Kjente eksempler på kaotiske systemer: værmodeller, en [dobbel pendel](#), og solsystemet vårt.

Merk: Når vi designer en regulator for et kaotisk system (for å f.eks. stabilisere et gitt punkt), så er vi faktisk ute etter å fjerne/eliminere all kaos fra systemet og i stedet gjøre dets oppførsel prediktivt og forutsigbart; med andre ord, for å få det til å høres mest mulig kult ut, så vil vi kontrollere kaos.

Trunkeringsfeil og numeriske feil

TODO

Flyttall og avrundingsfeil.

Trunkeringsfeil i fermover Euler.

Solver	Problem Type	Accuracy	When to Use
ode45	Nonstiff	Medium	Most of the time. ode45 should be the first solver you try.
ode23		Low	ode23 can be more efficient than ode45 at problems with crude tolerances, or in the presence of moderate stiffness.
ode113		Low to High	ode113 can be more efficient than ode45 at problems with stringent error tolerances, or when the ODE function is expensive to evaluate.
ode78		High	ode78 can be more efficient than ode45 at problems with smooth solutions that have high accuracy requirements.
ode89		High	ode89 can be more efficient than ode78 on very smooth problems, when integrating over long time intervals, or when tolerances are especially tight.
ode15s	Stiff	Low to Medium	Try ode15s when ode45 fails or is inefficient and you suspect that the problem is stiff. Also use ode15s when solving differential algebraic equations (DAEs).
ode23s		Low	ode23s can be more efficient than ode15s at problems with crude error tolerances. It can solve some stiff problems for which ode15s is not effective. ode23s computes the Jacobian in each step, so it is beneficial to provide the Jacobian via odeset to maximize efficiency and accuracy. If there is a mass matrix, it must be constant.
ode23t		Low	Use ode23t if the problem is only moderately stiff and you need a solution without numerical damping. ode23t can solve differential algebraic equations (DAEs).
ode23tb		Low	Like ode23s, the ode23tb solver might be more efficient than ode15s at problems with crude error tolerances.
ode15i		Fully implicit	Low

Figur 5.4: Liste over ODE-løserene i MATLAB. Hentet fra [lenke](#).

5.2.1 MATLABs numeriske ODE-løsere (ODE45 og ODE23)

En liste av alle ODE-løserene i MATLAB er vist i figur 5.4, hvor man også kan se hvilke typer systemer de er egnet til, samt deres nøyaktighet.

Kodesnutt 5.2: Simulering av pendel vha. ode45 i MATLAB.

```

%% Simulere en pendel vha. ode45
%% Init
x0 = [pi/2;0]; % Initialverdier
t0 = 0; % Starttid for simulering [s]
tend = 20; % Sluttid for simulering [s]
g = 9.81; % Gravitasjonsakselerasjonen [m/s^2]
l = 2; % Pendelens lengde [m]
params.a = g/l; % Modellparameter
%% Simulere:
opts = odeset('RelTol',1e-3,'AbsTol',1e-5);
[t,x] = ode45(@(t,x)pendDyn(t,x,params),[t0,tend],x0,opts);
%% Plotte:

```

```

figure(1); clf(1);
hold on;
subplot(2,1,1);
plot(t,x(:,1),'r','LineWidth',2);
ylabel('$x$ [rad]','Interpreter','latex','FontSize',12);
hold on;
subplot(2,1,2);
plot(t,x(:,2));
plot(t,x(:,1),'-o','LineWidth',1);
xlabel('$t$ [s]','Interpreter','latex','FontSize',12);
ylabel('$\dot{x}$ [rad/s]','Interpreter','latex','FontSize',12);
%% Dynamikken: (slutten av scriptet/egen funksjon)
function dxdt = pendDyn(t,x,params)
a = params.a;
dxdt=[x(2);-a*sin(x(1))];
end

```

Hva med tidsforsinkelser? Løseren `dde23` kan brukes ved konstante tidsforsinkelser.

Man kan også håndtere tidsforsinkelser ved hjelp av [transport delay-blokken](#) i Simulink.

5.2.2 Innstillinger: Relativ- og absolutt toleranse

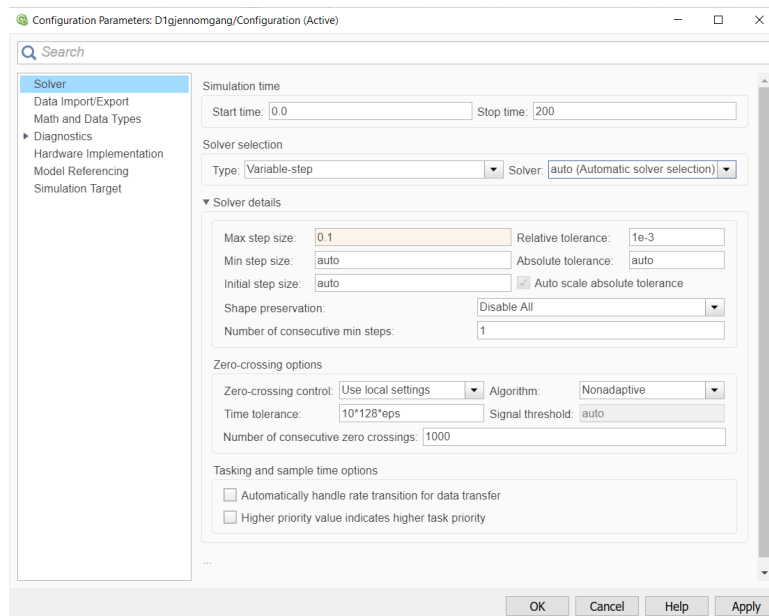
MATLAB: De forskjellige ODE-løserene har mange innstillinger man kan endre vha. `odeset-`kommandoen, se [denne lenken for full oversikt](#). Følgende tre er det viktig å ha kontroll på:

MaxStep: Den største steglengde ODE-løseren får lov å ta. Med andre ord: selv om løseren mener den kan ta et lengre steg uten at det går utover ønsket nøyaktighet (gitt av størrelsen under), får den ikke lov å gå over verdien til `MaxStep`.

Relativ toleranse (RelTol): Verdien `RelTol` spesifiserer den tillatte feiltoleransen i forhold til tilstandsvektoren ved hvert simuleringstrinn. Hvis du setter `RelTol` til $1e-2$ ($=0.01$), spesifiserer du at en feil på 1 % i forhold til hver tilstandsverdi er akseptabel ved hvert simuleringstrinn. Intuitivt sett, kontrollerer den antall signifikante sifre i en løsning, bortsett fra når det er mindre enn den såkalte *absolutte toleransen*.

Absolutt toleranse (AbsTol): `AbsTol` brukes til å bestemme den største tillatte absolutte feilen på ethvert trinn i en simulering. Denne toleransen “tar over for” den relative toleransen når en løsning er liten. Intuitivt sett, når løsningen nærmer seg 0, er `AbsTol` terskelen hvor du ikke lenger bekymrer deg for nøyaktigheten til løsningen (altså den relative toleransen) siden tilstandsverdiene uansett er veldig små (kanskje tilnærmet lik 0).

Simulink: I Simulink kan du endre på både løser og dens innstillinger via følgende sti: Modelling→Model Settings (tannhjul-icon). Alle innstillingsmuligheter er vist i figur 5.5.



Figur 5.5: Innstillingsmuligheter i Simulink.

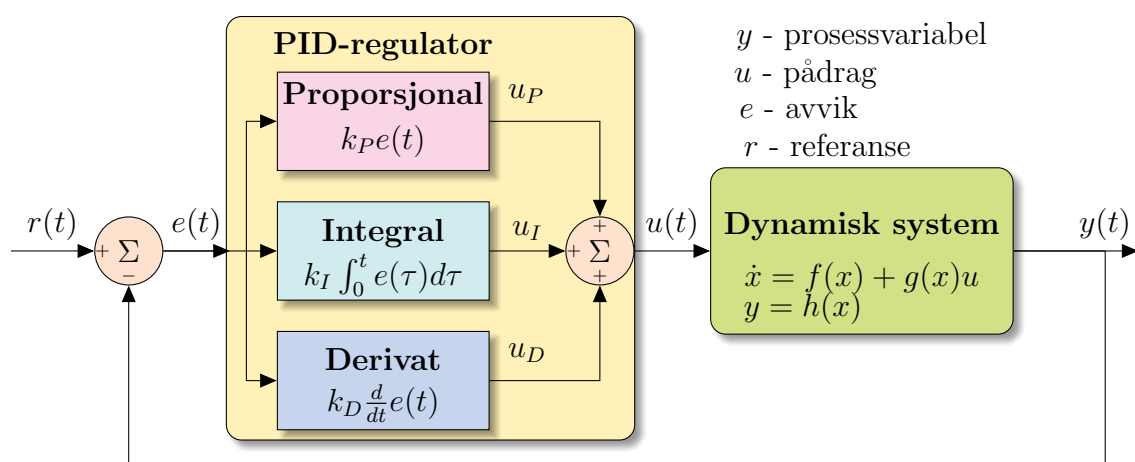
Del IV

Regulering av mono-variable systemer

6. PID-regulatoren

Alternative kilder: Kap. 8 i [Seborg et al., 2016]; Kap. 2 i [Bjørvik and Hveem, 2014]; WIKIPEDIA; Brian Douglas videoer: <https://youtu.be/wkfEZmsQqiA>.

Hvorfor et helt kapittel bare om PID? Enkelt og greit: Proporsjonal-integral-derivat- (PID)-regulatoren er, og vil trolig lenge forbli, den mest brukte reguleringsstrategien. Faktisk er trolig over 95% av regulatorene brukt i industrien PID-regulatorer, hvor de fleste er av PI-typen [Seborg et al., 2016, Desborough and Miller, 2002]. I følge spørreundersøkelsen [Samad, 2017] er PID-regulatoren den teknologien innen reguleringsteknikken med høyest (opplevd) effekt. Så populær er den, at det til og med skrives artikler om den!



Figur 6.1: Blokkdiagram av en PID-regulator på parallellform.

6.1. En PID-regulator i et nøtteskall

▶ jtpyvPdXbxg&t=14s

Den mest grunnleggende formen¹ til en Proporsjonal-Integral-Derivat (PID)-regulator er

$$u_{PID}(t) = \underbrace{u_{nom}}_{\text{nominelt pådrag}} + \underbrace{k_P e(t)}_{\text{proporsjonal}} + \underbrace{k_I \int_0^t e(\tau) d\tau}_{\text{integral}} + \underbrace{k_D \frac{de(t)}{dt}}_{\text{derivat}}. \quad (\text{PID})$$

Et blokkdiagram av en slik regulator er vist i figur 6.1. Der har vi et dynamisk system med én inngang, u , og én utgang, y , hvor målet er å få utgangen til å følge den ønskede referansen/settpunktet, $r(t)$. Vi bruker e (for “error”) for å betegne avviket, altså differansen mellom prosessvariabelen og referansen/settpunktet: $e(t) = y(t) - r(t)$.

6.1.1 P-leddet

▶ jtpyvPdXbxg&t=140s

$$\text{Tidsdomenet: } u_P(t) = k_P e(t). \quad \text{Laplacedomenet: } U_P(s) = k_P E(s)$$

Som navnet tilsier, så gir proporsjonal-leddet et bidrag som er proporsjonalt (gjennom forsterknings-parameteren k_P) med avviket. Økt proporsjonalforsterkning fører (som oftest) til følgende:

Fordeler:

- raskere respons / høyere båndbredde;
- mindre stasjonært avvik;

Ulemper:

- mindre stabilitetsmargin;
- høyere pådrag.

6.1.2 I-leddet

▶ jtpyvPdXbxg&t=268s

$$\text{Tidsdomenet: } u_I(t) = k_I \int_0^t e(\tau) d\tau \text{ eller } \dot{u}_I(t) = k_I e(t). \quad \text{Laplacedomenet: } U_I(s) = k_I \frac{E(s)}{s}$$

Et integral-ledd er som en hammer: det løser mange problemer, men skaper til gjengjeld mange problemer hvis det brukes uforsiktig og er ikke egnet til oppgaver som krever “finpresisjon”. Økt integralledd fører (som oftest) til:

Fordeler:

- at stasjonært avvik fjernes raskere;
- at man kan bruke forenklede regulerings-strukturer (se § 6.2.2).

¹I [Seborg et al., 2016] kalles dette for den “ekspanderte formen”.

Ulemper:

- tregeres respons på grunn av ekstra dynamikk, noe som kan føre til oversving;
- eventuell forverring av fenomenet windup (opptvinning) (se §9.1);
- potensiell rykk-og-napp oppførsel (og mulige grensesvingninger) på grunn av statisk friksjon/ “stiksjon” (se §3.3).

6.1.3 D-leddet

▶ jtpyvPdXbxg&t=632s

Tidsdomenet: $u_D(t) = k_D \dot{e}(t)$. Laplacedomenet: $U_D(s) = k_D s E(s)$

Man kan høre litt forskjellig når det kommer til betydningen av D -leddet; noen ganger er det hensiktsmessig å tenke på det som å legge til fiktiv demping i systemet, mens det andre ganger kan bli regnet som et “prediksjons”-ledd som forutser endringer i avviket/prosessvariabelen og korrigeres pådraget deretter.

En økning i D -leddet har som regel følgende effekt (avhenger litt av systemet og dødtiden):

Fordeler:

- raskere respons (rettere sagt, så tillater det det);
- økt stabilitet;
- mindre oversving (bidrar med en enkel prediksjon);

Ulemper:

- mater forsterket målestøy inn i systemet;
- mer oversving ved store tidsforsinkelser;
- sprang i pådrag ved endring i referansen (eng. “derivative kick”);

På grunn av den mulige forsterkningen av (hurtig-endrende) støy, implementeres D -leddet nesten alltid sammen med et derivatfilter (vi skal se på dette i §6.3).

For å unngå store sprang i pådraget grunnet endringer (sprang) i referansen, er det også vanlig at bare prosessvariabelen blir derivert (se §6.2.2).

6.1.4 Nominelt pådrag

▶ jtpyvPdXbxg&t=1136s

Nominelt pådrag er ment til å gi den pådragsverdien som i teorien vil opprettholde et ønsket arbeidspunkt. Strengt talt kan du sette dets verdi til hva enn du ønsker, og det er derfor også noen ganger kalt manuelt pådrag. Du bør allikevel prøve å sette verdien til en hensiktsmessig verdi som hjelper regulatoren din, i motsetning til at det blir virkende som en konstant forstyrrelse som må kompenseres for. Vi illustrerer dets virkemåte med et eksempel:

Eksempel: Gitt et første-ordens system $\dot{y} = u - y^2$ (kan være hastigheten til en bil, hvor y^2 da er luftmotstand), hvor vi ønsker at utgangen, $y(t)$, skal nå et settpunkt, r . I stedet for å være helt avhengig av integralvirkning, så kan vi ta $u = r^2 + k_P(r - y)$ slik at $\dot{y} \equiv 0$ når $y \equiv r$. Dynamikken til avviket $e = r - y$ er f.eks. da $\dot{e} = -(r + y + k_P)e$.

6.1.5 Direkte- vs reversvirkning

Et viktig spørsmål å spørre seg er: skal regulatorparameterne (k_P, k_I, k_D) ha positivt eller negativt fortegn? Hvis vi definerer avviket som $e = r - y$, så tilsvarer positive fortegn det vi kaller for *reversvirkning*, mens *direktevirkning* betyr negative fortegn.² Bruk derfor:

- Reversvirkning hvis et (positivt) sprang i pådraget (rettere sagt, signalet inn til pådraget) fører til en *økning* i prosessvariabelen;
- Direktevirkning hvis et (positivt) sprang i pådraget (rettere sagt, signalet inn til pådraget) fører til en *reduksjon* i prosessvariabelen.

Mer generelt, så avhenger dette av to ting: 1) hvordan vi definerer avviket/feilen; og 2) hva som er “positiv retning” for aktuatoren som gir pådraget (fører en økning i signalet som sendes til aktuatoren til en økning eller reduksjon i prosessvariabelen?).

Når det gjelder 1), så bruker vi jo i disse notatene hovedsakelig $e = r - y$ (se fig. 6.1), men, som tidligere nevnt, så er $e = y - r$ egentlig en mer egnet definisjon av et avvik.

For å illustrere 2), kan vi for eksempel se på første-ordens systemet $\dot{y} = bu$ med en konstant b . Si at vi ønsker å styre utgangen, y , til et konstant settpunkt r . Gitt avviket $e = r - y$, så har vi $\dot{e} = -\dot{y} = -bu$, slik at en proporsjonalregulator gir $\dot{e} = -bk_P e$, og dermed $e(t) = e(0) \exp(-bk_P t)$. Vi må derfor ha *positive* fortegn når b er positiv, og negative når b har negativt fortegn (se også §8.3.3 i [Seborg et al., 2016]).

Sjekkliste:

1. Brukes $e = y - r$ eller $e = r - y$ i regulatoren du bruker?
2. Er regulator-typen/-formen (mer om det senere) av typen du tror/bruker?
3. Har regulatoren mulighet for endre mellom revers- og direktevirkning eller kan du endre parameterfortegn?
4. Hvordan relateres signal fra regulatoren til pådragets respons?
5. Still inn regulatoren slik at du får negativ tilbakekobling (stabil avviksdynamikk).

6.2. Parallell-, serie-form og andre former

Det finnes mange forskjellige måter å representere en PID-regulatorer på (se, f.eks. tabell 8.1 i [Seborg et al., 2016], eller kap. 2 i [O’duyer, 2009] som også gir en oversikt over hvilke former som blir brukt av mange av de vanligste merkene).

²Noen bøker bytter om på disse definisjonene. Merk også at flere kommersielle regulator har også en egen innstilling for dette, og krever at alle parameterne har positive fortegn.

6.2.1 Parallel- og serie-form ▶ jtpyvPdXbxg&t=1807s

Foruten den ekspanderte formen (PID), så er parallell/sum og serie/produkt de to formene man vanligvis vil se på kommersielle PID-regulatorer ute i industrien.

Parallell-/sum-form

For en PID-regulator som (PID) ovenfor, er det vanlig å ta

$$k_I = k_P/T_I, \quad k_D = k_P \cdot T_D.$$

Dermed er *regulatorforsterkningen* k_p som før, mens T_I er *integral-tiden* og T_D er *derivat-tiden*. Laplace-transformasjon av en slik regulator (uten derivat-filter) har dermed følgende parallell-/sumform (denne kalles også ofte for den “ideelle” parallell-formen):

$$K_{\text{parallell}}(s) = k_P \left(1 + \frac{1}{T_I s} + T_D s \right). \quad (\text{parallellform})$$

Fun facts, bemerkninger og annet dill dall (you may skip)

Hvorfor kaller man det integraltid og derivattid?

Integraltiden stammer fra et noe hypotetisk scenario hvor avviket får et sprang fra null til en konstant verdi. Dette spranget i avviket vil forårsake at et umiddelbart hopp i proporsjonal-delen, altså $u_P = k_P e$, av regulatoren, mens integraldelen starter fra null og begynner å “integrere seg opp” via $\dot{u}_I = \frac{k_P}{T_I} e$. Integraltiden er dermed tiden det tar før integralledet har fått samme verdi som proporsjonal-delen (med antagelsen at avviket har holdt seg konstant).

Derivattiden stammer i stedet fra et scenario hvor avviket gradvis starter å øke med en konstant rate fra null. Proporsjonalledet starter derfor også fra null og øker i verdi, men derivat-leddet forblir konstant. Derivattiden er derfor tiden det tar før proporsjonal-leddet tar igjen derivat-leddet.

Serie-/produkt-form

En PID-regulator på serieform (ev. produktform) kan bli betraktet som en PI-regulator i serie med en PD-regulator:

$$K_{\text{serie}}(s) = k_P \underbrace{\left(\frac{1 + \tau_I s}{\tau_I s} \right)}_{\text{PI-del}} \underbrace{\left(\tau_D s + 1 \right)}_{\text{PD-del}}. \quad (\text{serieform})$$

Oppgave 6.1. Utled formellen til en PID-regulator på serie-form i tidsdomenet.

Omregning mellom serie of parallell:

Ser→Par: Ved å multiplisere ut parentesene i (serieform) får man

$$K_{\text{serie}}(s) = \kappa_P \left(1 + \frac{\tau_D}{\tau_I} + \frac{1}{\tau_I s} + \tau_D s \right).$$

Ved å sammenligne med (parallellform) finner vi at

$$k_P = \kappa_P(\tau_I + \tau_D)/\tau_I, \quad T_I = \tau_I + \tau_D \quad \text{og} \quad T_D = \frac{\tau_D \tau_I}{\tau_I + \tau_D}. \quad (6.1)$$

Par→Ser: For å gå fra (parallellform) til (serieform) må vi faktorisere (parallellform) for å finne dens nullpunkter. Dette kan vi bare gjøre hvis disse nullpunktene er reelle; vi vil jo ikke ha komplekse τ_I og τ_d ! Som regel er dog ikke dette et problem siden T_D/T_I normal sett er mellom 0.1 og 0.5 (formler kommer; tab 12.2 i Seborg).

Hvilken form er så best?

Når begge tilsvarer to reelle nullpunkter er det i prinsippet ingen forskjell mellom disse to formene, og man kan bruke formlene (6.1) til å gå fra den ene til den andre.

Siden (parallellform) også kan ha komplekse nullpunkt, så er jo denne noe mer generell enn (serieform). På den annen side, så er det helt tydelig for (serieform) hva de reelle nullpunktene er, slik at sum-formen har en mulig fordel i at det er lettere å forstå hva en endring i en gitt parameter fører til mtp. P-, I- og D-leddene.

Merk: Med tanke på direkte- vs reversvirkning (se §6.1.5) så er det bare fortegnet til k_P (evt. κ_P) man må ta hensyn til, mens de andre parameterne alle har positivt fortegn.

Oppgave 6.2. Du skal lage to MATLAB-funksjoner: den ene tar inn parameterene (k_p , T_i og T_d) til en PID-regulator på sum-/parallell-form og spytter ut tilsvarende parametere for en regulator på produktform; den andre funksjonen gjør det motsatte. Du kan f.eks. kalle disse `sum2prod` eller `par2ser`. Legg ved koden din, og vis hvordan du takler praktiske utfordringer som f.eks. uendelig integraltid ved en P/PD-regulator og komplekse nullpunkt.^a

^aEr du fornøyd med funksjonene dine, og ønsker å enkelt kunne bruke dem i fremtiden? Da kan du bruke `addpath` for å legge de til i MATLABS søke-sti.

6.2.2 PI-D , I-PD og beta-gamma-PID

For å unngå rykk fra proporsjonal- og derivatleddene grunnet sprang i referansen, kan man fjerne referanse fra en eller begge leddene. Hvis man fjerner referansen bare for derivat-leddet, får man en såkalt **PI-D**-regulator, som har følgende form:

$$U(s) = k_P \left[\left(1 + \frac{1}{T_I s} \right) E(s) - T_D s Y(s) \right]. \quad (\text{PI-D})$$

Ved å fjerne denne også for proporsjonal-leddet, får man en såkalt **I-PD**-regulator:

$$U(s) = k_P \left[\frac{1}{T_I s} E(s) - (1 + T_D s) Y(s) \right]. \quad (\text{I-PD})$$

En enda mer generell variant er såkalte $\beta\gamma$ -varianten³ (se [Seborg et al., 2016, 8.3.2]):

$$U(s) = k_P \left[\left(\beta + \frac{1}{T_I s} + \gamma T_d s \right) R(s) - \left(1 + \frac{1}{T_I s} + T_d s \right) Y(s) \right], \quad (\beta\gamma\text{-PID})$$

hvor $\beta, \gamma \in [0, 1]$, eventuelt

$$u = k_P (\beta r(t) - y(t)) + k_I \int_0^t (r(\tau) - y(\tau)) d\tau + k_D \frac{d}{dt} (\gamma r(t) - y(t)).$$

Merk at $(\beta, \gamma) = (0, 0)$ gir (I-PD), mens $(\beta, \gamma) = (0, 1)$ resulterer i (PI-D).

6.3. Derivat-filer



Hvorfor skal du lære dette? Et ideelt derivat-ledd har overføringsfunksjon tilsvarende en ren derivator med forsterkning, altså $T_D \cdot s$. Denne er ikke proper (høyere grad i teller enn nevner), og er dermed ikke realiserbar. Problemet er at en slik derivator vil forsterke både målestøy, referanse-endringer og til og med kvantiseringsfeil. Man må derfor alltid kombinere et slikt ledd med et filter som gjør overføringsfunksjonen realiserbar.

Et derivat-filter har som regel følgende form:

$$G_{cD}(s) = \frac{T_D s}{1 + T_f s} \quad (\text{Derivat-filter})$$

hvor $T_f > 0$ er filterkonstanten ($T_f \ll T_D$), eller eventuelt (se side 251 i [Seborg et al., 2016])

$$G_{cD}(s) = \frac{T_D s}{1 + \alpha T_D s}.$$

Tommelfingerregel: Typiske verdier er $\alpha \in [0.05, 0.2]$, hvor $\alpha = 0.1$ er vanlig.

Vi skal se mer på dette med hvilke verdier man bør velge for α i kapitlet som omhandler frekvensanalyse.

, som bare er et lavpass-filter i forbindelse med derivat-leddet, **Merk:** Det trenger ikke nødvendigvis være et første-ordens lavpass-filter. Mulig at vi faktisk ikke trenger det en gang hvis vi har et foldingfilter (mer om dette i kap. 12.1).

$$K(s) = k_p \left(1 + \frac{1}{T_I s} + \frac{T_D s}{1 + T_f s} \right). \quad (\text{Parallellform med derivat-filter})$$

$$K(s) = \kappa_p \left(\frac{1 + \tau_i s}{\tau_i s} \right) \left(\frac{\tau_d s + 1}{\tau_f s + 1} \right). \quad (\text{Serieform med derivat-filter})$$

³En variant av denne formen kalles også for en PID-regulator med to frihetsgrader (eng. 2DOF PID). Vi skal se mer på dette i § 8.1.8.

Hvordan implementere det: La $y(t)$, $y : \mathbb{R}_+ \rightarrow \mathbb{R}$, betegne en kontinuerlig og glatt (differensierbar) funksjon av tid. Vi kaller funksjonen $z(t)$ som er utgangen til det uforsterkede (ingen T_d) (**Derivat-filter**), gitt ved følgende formel, for et *estimat* av \dot{y} :⁴

$$Z(s) = \frac{as}{s+a} Y(s) \stackrel{1/a=T_f}{=} \frac{s}{T_f s + 1} Y(s). \tag{6.2}$$

Her er $Y(\cdot)$ og $Z(\cdot)$ er Laplace-transformasjonene til $y(t)$ og $z(t)$, mens $T_f = 1/a$ er positivt.

Som vist i oppgaven under, kan dette (lavpass-) filtrerte estimatet $z(t)$ av $\dot{y}(t)$ regnes ut ved hjelp av et “fiktivt” dynamisk system.

Oppgave 6.3. Vis at

$$Z(s) = \frac{bs}{s+a} Y(s)$$

tilsvar $z = w + by$ hvor $\dot{w} = -a(w + by)$.

Fra uttrykkene i oppgaven over er det lett å vise at $\dot{z} = -az + b\dot{y}$ som ønsket, og dermed

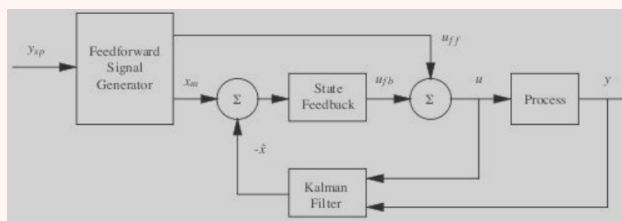
$$z(t) = e^{-at} \left[z(0) + b \int_0^t e^{a\tau} \dot{y}(\tau) d\tau \right] = \frac{b}{a} \dot{y}(t) - \frac{b}{a} e^{-at} \left[\dot{y}(0) + \int_0^t e^{a\tau} \ddot{y} d\tau \right],$$

hvor vi i det siste uttrykket antok at $z(0) = 0$. Det er derfor tydelig at hvis derivat til $y(t)$ ikke endres (altså $\dot{y} = \text{konst}$, og dermed $\ddot{y} \equiv 0$), så får vi $z(t) \rightarrow \frac{b}{a} \dot{y}$.

Fun facts, bemerkninger og annet dill dall (you may skip)

Tilstandsestimatorer og observere:

her



Ser→Par med derivat-filter: VVi ønsker nå å konvertere fra (**Serieform med derivat-filter**) til (**Parallellform med derivat-filter**) for $T_f = \tau_f$. Dette gir

$$k_P = \kappa_P(\tau_I + \tau_D - \tau_f)/\tau_I, \quad T_I = \tau_I + \tau_D - \tau_f, \quad T_D = \frac{\tau_D \tau_I + \tau_f^2 - \tau_f(\tau_I + \tau_D)}{\tau_I + \tau_D - \tau_f}, \quad T_f = \tau_f. \tag{6.3}$$

⁴Ved å bytte ut a i nevneren med et annet positivt tall b , så blir den resulterende variabelen $z(t)$ kalt et “skittent” (vektet) derivat (eng: dirty derivative) av $y(t)$ [Loría, 2015].

6.4. P(I) eller P(I)D?

▶ [jtpyvPdXbxg&t=2747s](#)

Første-ordens systemer: I teorien trenger man bare en P(I)-regulator for første-ordens systemer. En ren integrator kan tidvis brukes, men kan ofte lede til oscillerende responser (i den ideelle verden svarer jo dette til en harmonisk oscillator).

Merk: Det er et viktig unntak til regelen over, nemlig systemer med betydelige tidsforsinkelser (se §4.2).

Andre-ordens systemer: For andre-ordens systemer som ikke er veldig stabile i utgangspunktet trenger man som regel en P(I)D-regulator for å få ønsket sprangrespons.

Oppgave 6.4. Hvorfor kan det være hensiktsmessig med også et derivat-ledd for første-ordens systemer med en konstant tidsforsinkelse $\theta > 0$? Begrunn svaret. Hint: Bruk første-ordens Padé-approximasjonen til tidsforsinkelsen $e^{-\theta s}$.

Hvor kommer reglene over fra? Gitt en PD-regulator på formen

$$U_{PD}(s) = \left[\frac{k_D s + k_P}{1} \right] E(s) = \frac{\mathcal{T}_{PD}(s)}{\mathcal{N}_{PD}(s)},$$

altså som en overføringsfunksjon (fra e til u_{PD}) med ett nullpunkt $s = -k_P/k_D$. Hvis regulatoren er koblet i serie med en prosess på formen

$$P(s) = \frac{\mathcal{T}_P(s)}{\mathcal{N}_P(s)} = \frac{k_0}{n_2 s^2 + n_1 s + n_0}$$

så er jo overføringsfunksjonen til den lukkede sløyfen gitt ved

$$G_{LS}(s) = \frac{\mathcal{T}_{ls}(s)}{\mathcal{N}_{ls}(s)} = \frac{\mathcal{T}_{PID}(s)\mathcal{T}_P(s)}{\mathcal{T}_{PID}(s)\mathcal{T}_P(s) + \mathcal{N}_{PID}(s)\mathcal{N}_P(s)}.$$

Nevneren er dermed

$$\mathcal{N}_{ls}(s) = n_2 s^2 + (n_1 + k_0 k_D) s + n_0 + k_0 k_P.$$

For et første-ordens system ($n_2 = 0$), så kan vi ta

$$k_P = -\frac{(n_1 p_* + n_0)}{k_0}, \quad k_D = 0$$

for at den lukkede sløyfen skal en ønsket pol p_* . For et andre-ordens system ($n_2 \neq 0$), derimot, så er det dog tydelig at vi også må bruke et derivat-ledd for å oppnå ønsket polplassering.

6.4.1 PID for andre-ordens-dominante prosesser

Siden D-leddet fører til en mer kompleks regulator, som blant annet er mer sensitiv til målestøy, så anbefaler Skogestad [Skogestad, 2003] å bruk dette kun for andre-ordens dominante prosesser

$$P(s) = \frac{k \cdot e^{-\theta s}}{(\tau_1 s + 1)(\tau_2 s + 1)}$$

der “dominant” betyr at den nest største tidskonstanten, τ_2 ($< \tau_1$), er større enn den (effektive) tidsforsinkelsen, θ , altså:

PI vs PID: Gitt et AOPTF-system, velg en PID-regulator hvis $\tau_2 > \theta$.

6.5. PID-regulatoren i Simulink

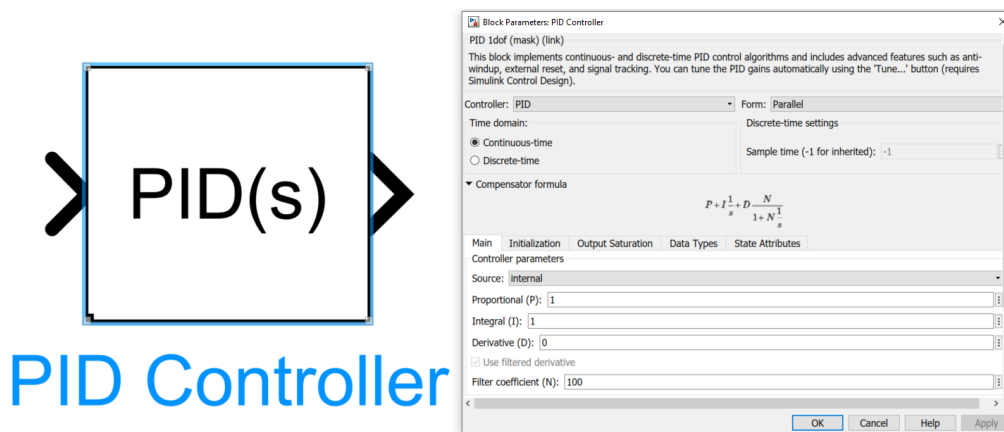
▶ [jtpyvPdXbvg&t=2977s](#)

PID-regulator-blokken vist i figur 6.2 finner du under “Continuous” i Library Browseren.

Alert! Simulink bruker ikke integral- og derivat-tid! Pass derfor på at du bruker riktig versjon. Du kan endre mellom de forskjellige versjonene (ideal og parallel) under Form oppe til høyre i figur 6.2.

Som du kan se av figur 6.3, er det også mange andre mulige innstillinger enn bare typen regulator, og dens implementasjon er relativt kompleks.

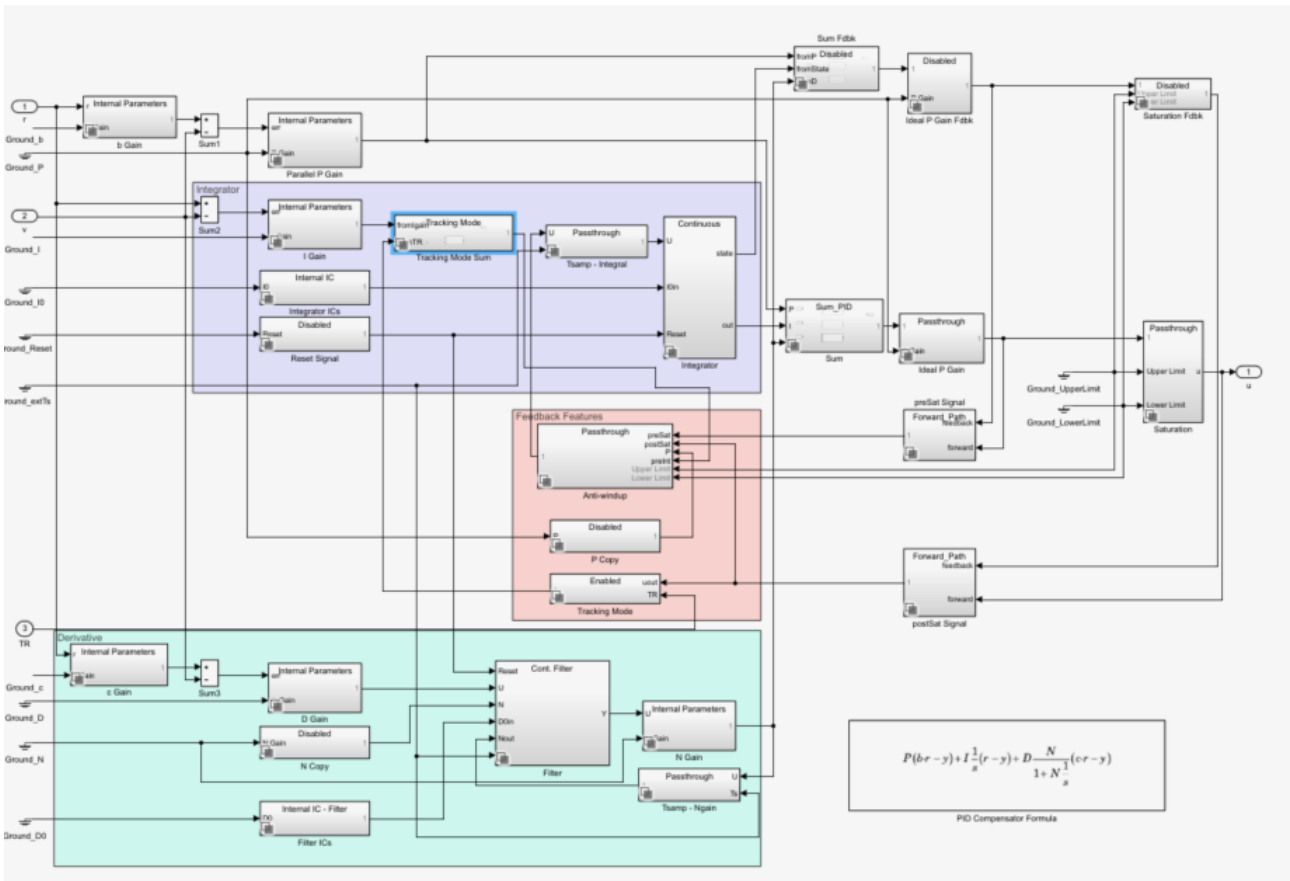
Teste Simulink control system designer for pol-plassering.



Figur 6.2: PID-regulatoren i Simulink og dens innstillinger.

6.6. Modellfri tuning av PID-regulatorer

Fra [Seborg et al., 2016, §12.5] har vi følgende punkter:



Figur 6.3: Hvordan PID-regulatoren i Simulink er implementert.

- **Regulatorinnstilling innebærer unngåelig en avveining mellom ytelse og robusthet:** Ytelses-målene fordelt på utmerket referansefølging og forstyrrelsesavvisning bør balanseres opp mot robusthetsmålet i forhold til stabil drift over et bredt spekter av forhold og arbeidsområder.
- **Regulatorinnstillinger trenger ikke å være nøyaktig bestemt:** Generelt vil en liten endring i en preglatorparameterene fra sin “beste” verdi (f.eks. ±10 %) ha liten effekt på responsen til den lukkede sløyfen.
- **For de fleste prosesser er det ikke mulig å manuelt stille inn hver regulator:** Tuning gjøres vanligvis av en kontrollspesialist (ingeniør eller tekniker) eller av en anleggsoperatør. Siden en slik spesialist i prosessindustrien kan være ansvarlig for så mange som 300 til 1000 reguleringsløyfer, er det ikke alltid mulig å justere hver regulator manuelt. I stedet fokuserer man på reguleringsløyferne som oppfattes å være de viktigste eller problematiske. De andre regulatorene kan vanligvis bare bruke forhåndsinnstillingene.

6.6.1 Etterjustering og manuell tuning



Alternative kilder: §12.7 i [Seborg et al., 2016] (se også §12.3.5); 2.9 i [Bjørvik and Hveem, 2014]

Hvorfor etterjustere? Det finnes to hovedscenarier for det vi nå skal kalle etterjustering:

- Du har stilt inn PID-regulatoren med metodene du har lært.
- Regulatoren ble stilt inn for en tid tilbake, og fungerte bra da.

Men du ikke helt fornøyd med innstillingene, selv om de på et vis fungerer. Kanskje du har brukt en litt unøyaktig matematisk modell, eller det har skjedd en endring i prosessen?

Mulige grunner til endringer:

1. En komponent i prosessen har feilet fullstendig: Ledningsbrudd, kortslutning, noe har satt seg fast, noe er tett, noe har løsnet.
2. En komponent mangler vedlikehold, og er i ferd med å svikte. Jordfeil, noe går tregt, noe går sakte, noe slarker, noe lekker.
3. En prosess i nærheten forstyrrer, eller en forstyrrelse har endret karakter: Kan skyldes det samme som over.
4. Prosessparametre har “bare forandret seg”: Kan skyldes det samme som over, men kanskje i noe mindre grad.
5. Minnet til prosessoperatørene har endret seg: “Alt var bedre før”.
6. Noen har utført reparasjon eller vedlikehold.

Sjekkliste før du justerer: Du må først forsikre deg om at alt er som det skal være:

- Ingenting kan bli bedre før problemer i kategori 1 er fikset. Start der.
- Problemer i kategori 2, 3 eller 4 kan kanskje fikses midlertidig med etterjustering. Noen av disse kommer til å bli værende til neste store vedlikehold, eller til prosessen skal avvikles. Husk ny justering når det egentlige problemet er fikset.
- Problemer i kategori 5 forekommer. Studer dokumentasjon fra tidligere innstilling av regulatoren. Dokumentér det du selv gjør.
- Problemer i kategori 6 forekommer. Egentlig er dette gode nyheter. Etterjustering burde kanskje vært planlagt, eller du burde fått beskjed. Men sånn er det.

Du har bestemt deg for å etterjustere, hva nå? Etterjustering er ingen eksakt vitenskap, men mer en blanding av intuisjon, forståelse (av både prosessen og regulator-delene), samt kunnskap og erfaring. Følgende huskeregler kan være et utgangspunkt, men er ingen fasit/guide:

Huskeregler: (for Parallellform med derivat-filter)

- Er sløyfa urolig (svingete/ oscillerende): reduser k_P og/eller $1/T_I$; øk **eller** reduser T_D .
- Er det dynamiske avviket for stort: øk $1/T_I$ og/eller k_P og/eller T_D .
- Er utgangen eller aktuatoren urolig/“vibrerende”: reduser T_D og/eller k_P .

6.6.2 Ytelses-karakteristikker og -metriker

▶ 5Tip6_DMe2A&t=9

Alternative kilder: §12.3.2 i [Seborg et al., 2016]

Som nevnt, så innebærer regulator-syntese og -tuning alltid en avveining mellom ytelse og robusthet. Robusthet skal vi se litt mer på i kapittel 11. Men når det gjelder ytelse, hvordan kan vi “måle” dette? Og ikke minst, hva er vi ute etter?

Vanlige karakteristikker:

- Dempningskarakteristikker og oversving
- Innsvingningstid og responstid

Vanlige metrikker:

- **IAE** (integral av absoluttverdien til avviket (“error”)): $M_{IAE} = \int_0^\infty |e(t)| dt$ (se fig. 6.4).
- **ISE** (integral av kvadratet (square) til avviket): $M_{ISE} = \int_0^\infty |e(t)|^2 dt$.
- **ITAE** (tids-vektet integral av absoluttverdi til avviket): $M_{ITAE} = \int_0^\infty t \cdot |e(t)| dt$.
- **IQC** (“integral quadratic cost” (brukes til LQR)): $M_{IQC} = \int_0^\infty |e(t)|^2 + |u(t)|^2 dt$.

Merk: Slike metrikker brukes både til direkte regulator-design (f.eks. finne PID-parametrene for et gitt systemet som minimerer den ønskede metrikken) og til å måle ytelse ved å kjøre eksperimenter (bytter da ∞ i integralet med eksperimenttiden T).

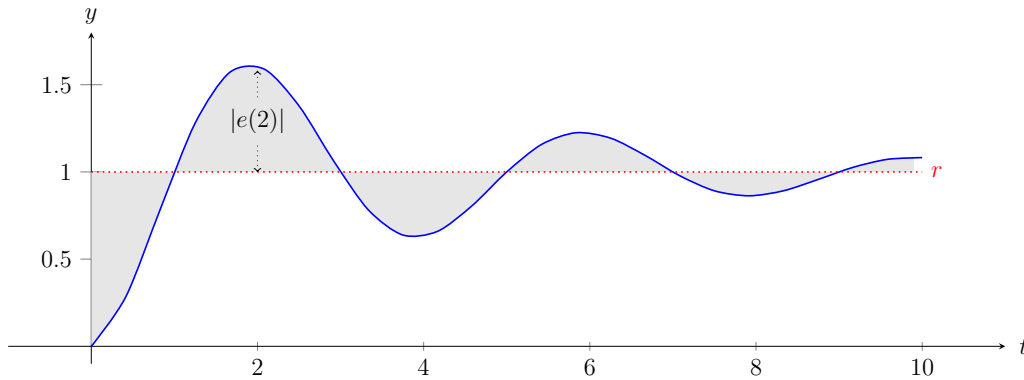
6.6.3 Ziegler-Nichols’ metoder

▶ 5Tip6_DMe2A&t=351

Alternative kilder: [Seborg et al., 2016, §12.5.1] og [Seborg et al., 2016, §12.5.3]; [Ogata et al., 2010]; Wikipedia.

Ziegler og Nichols (ZN) sine regler (som dere kjenner fra før) forble i omtrent 50 år de mest brukte tuning-reglene for PID-regulatorer. Det er dog (minst) tre problemer med ZN-reglene [Grimholt, 2018]:

1. Innstillingene er ganske aggressive for de fleste prosesser med svingninger og overskridelser.



Figur 6.4: Illustrasjon av IAE-metrikken fra en sprangrespons. M_{IAE} tilsvarer det samlede arealet til de grå områdene.

2. Regelen inneholder ingen innstillingsparameter for å justere robustheten og gjøre den mindre aggressiv.
3. For en ren tidsforsinkelsesprosess gir ZN-PID-innstillingene ustabilitet og ZN-PI-innstillingene gir svært dårlig ytelse.

Åpen-sløyfe-metoden

▶ 5Tip6_DMe2A&t=442

Ziegler-Nichols' (første) åpne-sløyfe metode for (PID) steg for steg:

For en PID-regulator på [parallellform](#):

- Steg 1:** Sett systemet i manuell modus(skurv av alle regulatorer og tilbakekobling, f.eks. ved å sette $k_P = 0$, $T_I = \infty$ (altså $k_P/T_I = 0$) og $T_D = 0$);
- Steg 2:** Sett et sprang i pådraget fra $u = u_0$ til u_{sprang} , slik at $\Delta u = u_{sprang} - u_0$;
- Steg 3:** Marker der tangenten til vendepunktet krysser de horisontale aksene for å finne θ og a som vist i figur 6.5 (merk at $a = y_\infty \cdot \theta/\tau$);
- Steg 4:** Ta $L = \Delta u/a$ og sett regulator-parameterne ut fra følgende tabell:

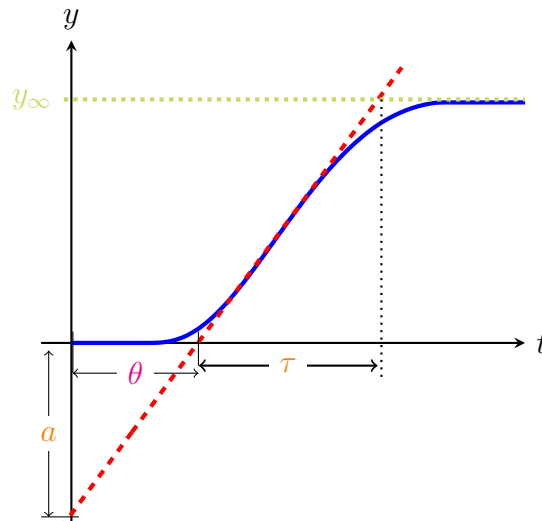
Regulatortype	k_P	T_I	T_D
P	L	∞	0
PI	$0.9L$	$\theta/0.3$	0
PID	$1.2L$	2θ	0.5θ

Merk: Man kan fra dette tilpasse en FOPTF-modell ved å ta $k = y_\infty/\Delta u$; se §4.3.2.

Fordeler: den trenger bare én test for å bestemme parameterene.

Ulemper: Følgende liste mer ulemper er noget lenger enn listen med fordeler:

1. **NB!** Krever at den uregulerte prosessen er (asymptotisk) stabil.



Figur 6.5: Illustrasjon av konstantene i Ziegler-Nichols' åpne-sløyfe metode.

2. Den eksperimentelle testen utføres i åpen-sløyfe, slik at hvis en betydelig forstyrrelse oppstår under testen, vil ingen korrigerende handling bli tatt. Testresultatene kan dermed være misvisende.
3. For en ulinear prosess kan testresultatene være følsom for størrelsen og retningen til spranget. Hvis størrelsen på spranget er for stort, kan prosess-ulineariteter påvirke resultatet. Men hvis trinnstørrelsen derimot er for liten, kan sprangresponsen være vanskelig å skille fra de vanlige svingningene på grunn av støy og forstyrrelser. Retningen til trinndringen (positiv eller negativ) bør velges slik at den regulerte variabelen ikke vil bryte noen systembegrensninger.
4. For kontinuerlige (analoge) regulatorer, har metoden en tendens til å være følsom for kalibreringsfeil.
5. Parameterne a og θ bestemmer regulatorparameterne direkt; vi har dermed ingen frihetsgrader som vi kan bruke til å "tune" responsen!

Alternativ: SIMC fra tilpasset sprangrespons.

Lukket-sløyfe-metoden ▶ 5Tip6_DMe2A&t=747

Ziegler-Nichols' andre (lukkede-sløyfe) metode for for (PID) steg for steg:

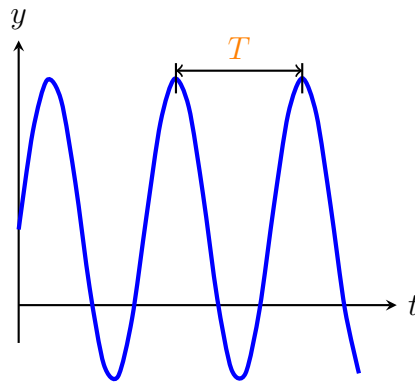
For en reversvirkende PID-regulator på [parallellform](#):

Steg 1: Sett $k_P = k_k$ hvor $k_k > 0$ er liten (≈ 0), $T_I = \infty$ (altså $k_P/T_I = 0$) og $T_D = 0$;

Steg 2: Sett et sprang i referansen fra $r = r_0$ til $r = r_{ny}$ ($\neq r_0$).

Steg 3: Øk k_k til prosessvariabelen får stående og repeterende oscillasjoner (se fig. 6.6);

Steg 4: Mål periodetiden T til oscillasjonene og sett regulator-parameterne ut fra tabellen:



Figur 6.6: Illustrasjon av stående svingninger for Ziegler-Nichols' andre metode.

Regulatorstype	k_P	T_I	T_D
P	$0.5k_k$	∞	0
PI	$0.45k_k$	$0.8T$	0
PD	$0.8k_k$	∞	$0.125T$
PID	$0.6k_k$	$0.5T$	$0.125T$

Merk: Fra T kan vi estimere den kritiske frekvens $\omega_k = \frac{2\pi}{T}$ til prosessen, altså frekvensen hvor Nyquist-kruven krysser den negative reelle tallinjen. Den kritiske forsterkningen k_k er derfor forsterkningen som bringer den lukkede sløyfen til stabilitetsgrensen (marginal stabilitet).

Fordeler:

1. Den trenger bare én test for å bestemme parameterene;
2. Testen er gjort i lukket-sløyfe, slik at den også bedre tar hensyn til dynamikk fra aktuatorer og sensorer.

Ulemper:

1. Krever at man kan oppnå stående svingninger (oscillasjoner) med kun en proporsjonalregulator (dette er dog vanlig i prosessindustrien).
2. En prøve-og-feile-strategi for å finne k_k og T kan være veldig tidkrevende, spesielt for systemer med treg dynamikk.
3. Metoden krever at man gjør den lukkede sløyfen marginalt stabilt (stående svingninger). Hvis noe uønsket skjer under innstillinger (f.eks. forstyrrelser eller endringer i prosessen), så kan dette lede til farlige eller uønskede situasjoner.
4. Regulatoren blir bestemt på grunnlag av to parametere, den kritiske-/ultimate-forsterkningen k_k og -periodetiden T . En FOPTF-modell, derimot, er gitt av tre parametere, (k, τ, θ) , noe som betyr at ZN-reglene ikke kan fungere godt på et bredt spekter av slike prosesser.

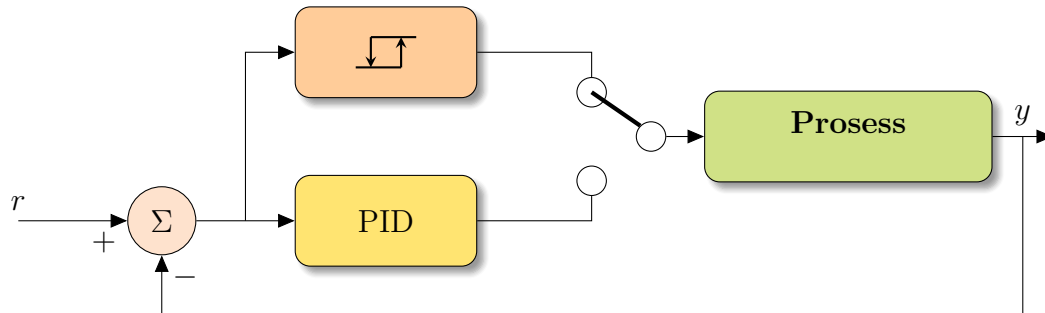
Alternativer: Skogestads lukkede-sløyfe-metode med SIMC eller Åstrøms relé-metode.

6.6.4 Auto-tuning og Åstrøms relé-metode

▶ 5Tip6_DMe2A&t=951

Åstrøms relé-metode

Alternative kilder: [Seborg et al., 2016, §12.5.2]; [Åström and Hägglund, 1984].



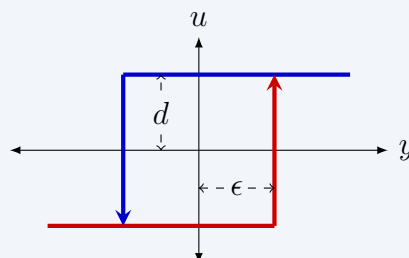
Figur 6.7: Åström (relè) auto-tuning metode.

Idé: Bruke en relé- (av/på-) regulator med hysteres (kan bare endre verdien (fra av til på eller motsatt) en viss tid etter en endring). Dette skaper grensesvinginger i prosessen (se §3.3.5) som er ca. 180° faseforskjøvet i forhold til inngangen.

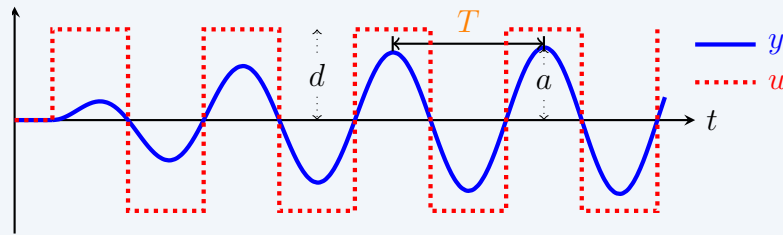
Prosedyre for Åstrøms auto-tuning relé-metode:

For PID-regulatorer på [parallellform](#):

1. *Initialisering:* Prosessen bringes til ønsket arbeidspunkt, enten av operatøren i manuell modus eller av en tidligere innstilt regulator i auto-modus. Når sløyfe-responsen har blitt stasjonær (transienter har dødd ut), så kan auto-tuningen begynne.
2. *Auto-tuning:* PID-regulatoren er midlertidig frakoblet og erstattet med et relé med hysteres, som vist i figuren under. Hysteresens bredde ϵ bestemmes automatisk fra støynivået for å unngå for hyppig endring (fra av til på og motsatt) i pådraget. Under svingningen vil amplituden til reléet, d , justeres slik at prosessvariabelen får stående (konstante) svingninger av en ønsket amplitude.



3. *Parameteravlesning:* Regn ut den ultimate forsterkningen og perioden via $k_k = \frac{4d}{\pi a}$ og T , hvor d er relé-amplituden til pådraget, a er amplituden til de stående svingningene i prosessvariabelen, og T er periodetiden, som vist i følgende figur:



4. *Parametersetting*: Gitt k_k og T , regn ut regulator-parameterne vha. (f.eks.) ZN-tabellen:

Regulatortype	k_P	T_I	T_D
P	$0.5k_k$	∞	0
PI	$0.45k_k$	$0.8T$	0
PD	$0.8k_k$	∞	$0.125T$
PID	$0.6k_k$	$0.5T$	$0.125T$

⚠ Viktig! Hvis man bruker hysteres, så bør man egentlig ta hensyn til hysteres-bredden når man regner ut den kritiske forsterkningen. Dette er fører dog til noe mer komplekse uttrykk som vi dermed utelater; se [Åström and Hägglund, 1984] for ytterligere detaljer.

Fordeler:

1. Bare en enkelt eksperimentell test kreves i stedet for en prøve-og-feile-prosedyre.
2. Amplituden til prosessutgangen kan begrenses ved å justere relé-amplituden.
3. Prosessen er ikke tvunget til en stabilitetsgrense.
4. Den eksperimentelle testen kan (relativt) lett automatiseres.

Ulemper:

1. For langsomme prosesser er det kanskje ikke akseptabelt å utsette prosessen for de to til fire syklusene med svingninger som kreves.
2. Hensyn må tas til eventuelle (konstante) forstyrrelser, som kan forårsake asymmetriske oscillasjoner.
3. Tilvarende ZN-metoden, blir regulatoren bestemt på grunnlag av to parametere, den kritiske-/ultimate-forsterkningen k_k og -periodetiden T . En FOPTF-modell, derimot, er gitt av tre parametere, (k, τ, θ) , noe som betyr at denne metoden heller ikke kan fungere godt på et bredt spekter av slike prosesser.
4. Patentbeskyttet (?).

Simulink*

Se <https://se.mathworks.com/help/slcontrol/ug/how-pid-autotuning-works.html>.

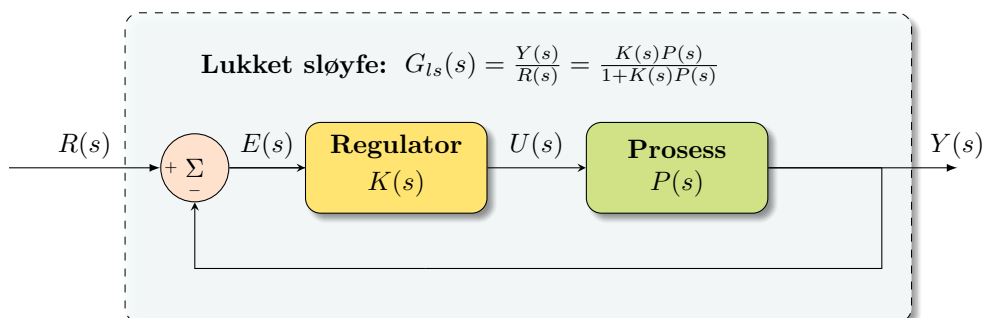
7. Modellbasert regulator-design

7.1. Direktesyntese

▶ UV0AK8Rh9pU&t=16

Alternative kilder: §12.2 i [Seborg et al., 2016].

Direktesyntese: Designe regulatoren slik at overføringsfunksjonen til den lukkede sløyfen tilsvarer en ønsket overføringsfunksjon.



Figur 7.1: Enkelt reguleringsystem i Laplace-domenet.

Motiverende eksempel: Ta reguleringsystemet i figur 7.1 uten støy og forstyrresler, sånn at overførings-funksjonen til den lukkede sløyfen tilsvarer

$$G_{LS}(s) = \frac{Y(s)}{R(s)} = \frac{K(s)P(s)}{1 + K(s)P(s)}.$$

Løser vi for $K(s)$, finner vi at

$$K(s) = \frac{1}{P(s)} \frac{Y(s)/R(s)}{(1 - Y(s)/R(s))} \quad (7.1)$$

Merk: vi kan ikke oppnå perfekt og umiddelbar regulering, altså $Y(s) = R(s)$, siden dette krever at $K(s) \rightarrow \infty$.

En mer realistisk strategi er å sikte for en lukket sløyfe hvor alle polene er plassert i venstre halvplan, såkalt **pol-plassering**. Altså, vi ønsker å finne $K(s)$ sann at f.eks.

$$\frac{Y(s)}{R(s)} = G_{\text{LS}}^{\circ}(s) = \frac{k}{(s + \lambda_1) \cdot (s + \lambda_2) \cdot \dots \cdot (s + \lambda_n)}, \quad G_{\text{LS}}^{\circ}(0) = 1,$$

hvor superscriptet “ \circ ” leses som “ønsket”, $n \geq 1$ og alle de (komplekse¹) polene $\lambda_1, \dots, \lambda_n$ har en ønsket plassering i venstre halvplan. Vi krever her $G_{\text{LS}}^{\circ}(0) = 1$ for at vi ikke skal få stasjonært avvik, slik at en mulig kandidat er

$$G_{\text{LS}}^{\circ}(s) = \frac{1}{(\tau_1^{\circ}s + 1) \cdot (\tau_2^{\circ}s + 1) \cdot \dots \cdot (\tau_n^{\circ}s + 1)}$$

for reelle, positive tall $\tau_1^{\circ}, \dots, \tau_2^{\circ}, \tau_n^{\circ}$.

Dette betyr at regulatoren må være gitt ved følgende

Direktesyntese: Gitt en ønsket overføringsfunksjon, $G_{\text{LS}}^{\circ}(s)$, ta

$$K_{ds}(s) = \frac{1}{\hat{P}(s)} \frac{G_{\text{LS}}^{\circ}(s)}{(1 - G_{\text{LS}}^{\circ}(s))}. \quad (\text{Direktesyntese})$$

Merk: Vi bruker i realiteten alltid en modell $\tilde{P}(s)$ av den ekte prosessen $P(s)$.^a

^aDu kan i disse notatene anta at $\hat{P} = P$ hvis annet ikke er direkte spesifisert.

Viktig! Krever invertering av prosessen, slik at denne må være minimum-fase (altså stabil og uten nullpunkter i høyre halvplan).

Eksempel 7.1. For prosessen $P(s) = k/(\tau s + 1)$ ønsker vi at $G_{\text{LS}}^{\circ}(s) = 1/(\tau_{\star} s + 1)$. Vi får

$$K_{ds}(s) = \frac{\tau s + 1}{k} \frac{1/(\tau_{\star} s + 1)}{1 - 1/(\tau_{\star} s + 1)} = \frac{\tau s + 1}{k \tau_{\star} s}, \quad (7.2)$$

altså en PI-regulator.

Det er dog viktig at man følger regelen under:

Acthung! En ustabil pol bør aldri (ukritisk) kanselleres vha. en regulator med et nullpunkt i høyre halvplan pga. mulig sensitivitet til usikkerhet og forstyrrelser (lav robusthet).

Neste eksempel (se også ([Seborg et al., 2016, Ex.J1]) demonstrerer viktigheten av denne huskeregelen:

Eksempel 7.2. Gitt et system på formen

$$Y(s) = P(s)(U(s) + D(s))$$

¹Komplekse poler må selvsagt komme i kompleks-konjugerte par.

med en forstyrrelse $D(s)$ og hvor prosessen

$$P(s) = \frac{0.5}{2s - 1 - \epsilon}$$

er ustabil. La oss si vi ønsker $G_{LS}^{\circ}(s) = 1/(s + 1)$ og vår modell er $\hat{P}(s) = \frac{0.5}{2s-1}$. Dermed

$$K_{Ds}(s) = \frac{1}{\hat{P}(s)} \frac{G_{LS}^{\circ}(s)}{(1 - G_{LS}^{\circ}(s))} = \frac{4s - 2}{s}.$$

Scenario 1 – feil modell: Den lukkede sløyfen er

$$G_{LS}(s) = \frac{K_{Ds}(s)P(s)}{1 + K_{Ds}(s)P(s)} = \frac{2s - 1}{s(4s - 2(1 + \epsilon)) + 4s - 2}.$$

Polene er dermed

$$p = \frac{-1 + \epsilon}{4} \pm \frac{1}{4} \sqrt{9 + \epsilon^2 - 2\epsilon}.$$

Konklusjon: Med unntak av $\epsilon = 0$, så er $G_{LS}(s)$ ustabil for alle ϵ !

Scenario 2 – korrekt modell: Anta nå at vår modell er helt korrekt ($\epsilon = 0$), slik at

$$G_{LS}(s) = \frac{K_{Ds}(s)P(s)}{1 + K_{Ds}(s)P(s)} = \frac{1}{s + 1},$$

som jo er stabil. På den annen side, så tilsvare dette $Y(s) = G_{LS}(s)R(s) + (s)D(s)$ hvor

$$G_d(s) = \frac{P(s)}{1 + K_{Ds}(s)P(s)} = \frac{0.5/(2s - 1)}{1 + 1/s} = \frac{0.5s}{(2s - 1)(s + 1)}.$$

Konklusjon: Systemet ustabil relativt til forstyrrelsen!

Oppgave 7.1. Gitt prosessen $P(s) = \frac{k}{(\tau_1 s + 1)(\tau_2 s + 1)}$, finn $K(s)$ ved hjelp av direkte syntese sånn at den lukkede sløyfen tilsvare $G_{LS}^{\circ}(s) = 1/(\tau_* s + 1)$.

7.1.1 Modifisert direkte syntese for ikke-minimum-fase systemer

▶ UV0AK8Rh9pU&t=1095

Nåværende problem: [Direktesyntese](#) bør ikke brukes direkte (pun intended) for systemer med nullpunkter i høyre halvplan og/eller tidsforsinkelser (ikke-minimum-fase systemer).

Grunnene til dette er at [Direktesyntese](#) krever invertering av prosessen $P(s)$, hvor

- Invertering av nullpunkt i høyre-halfplan gir en regulator med pol i høyre halvplan²

²Selv om ustabile poler eller nullpunkt i teorien blir kansellert når en slik regulator er i serie med en prosess, vil det alltid i virkeligheten være små forskjeller mellom den ekte modellen og våre matematiske modell, noe som potensielt kan resultere i et ubegrenset pådrag; ikke minst må vi passe oss for forstyrrelser (husk eksempel 7.2)!

- Invertering av tidsforsinkelser tilsvarer at vi kan se inn i fremtiden!

Det finnes allikevel en slags løsning på begge disse problemene: Vi kan faktorisere ut de delene av prosessen som er minimum-fase (se § 2.6.6) og bruke disse i (in-)direktesyntesen:

Modifisert (in-)direktesyntese for systemer med positive nullpunkt:

Faktorerer prosessen som følger:

$$P(s) = P_{MF}(s)P_{IMF}(s)$$

hvor

- $P_{MF}(s)$ består av delene av $P(s)$ som er minimum-fase (se § 2.6.6);
- $P_{IMF}(s)$, $P_{IMF}(0) = 1$, inneholder eventuelle tidsforsinkelser og nullpunkt i høyre halvplan.

Ved å kun kreve at $Y(s)/R(s) = G_{LS}^{\phi}(s) = P_{IMF}(s)G^{\phi}(s)$ (hvor da $G^{\phi}(0) = G_{LS}^{\phi}(0) = 1$) ser man fra (7.1) at man bare trenger å invertere $P_{MF}(s)$:

$$K_{MDS}(s) = \frac{1}{\hat{P}(s)} \frac{\hat{P}_{IMF}(s)G^{\phi}(s)}{(1 - \hat{P}_{IMF}(s)G^{\phi}(s))} = \frac{1}{\hat{P}_{MF}(s)} \frac{G^{\phi}(s)}{(1 - \hat{P}_{IMF}(s)G^{\phi}(s))} \quad (\text{Mod. direktesyntese})$$

La oss ta et eksempel hvor vi har et positivt nullpunkt:

Eksempel 7.3. Gitt

$$P(s) = \kappa \frac{(s - \alpha)}{(s + \beta)^2} \quad \text{og} \quad G^{\phi}(s) = \frac{1}{(\tau_{\phi}s + 1)^n}$$

Vi antar at $\hat{P} = P$ og faktorerer $P(s) = P_{IMF}(s)P_{MF}(s)$ som følger:

$$P_{IMF}(s) = (1 - s/\alpha) \quad \text{og} \quad P_{MF}(s) = -\kappa\alpha/(s + \beta)^2,$$

slik at $P_{IMF}(0) = 1$. Dermed får vi

$$K_{DS}(s) = -\frac{(s + \beta)^2}{\kappa\alpha} \frac{\frac{1}{(\tau_{\phi}s + 1)^n}}{1 - (1 - s/\alpha)\frac{1}{(\tau_{\phi}s + 1)^n}} = \frac{-1}{\kappa} \frac{(s^2 + 2\beta s + \beta^2)}{(\alpha(\tau_{\phi}s + 1)^n - \alpha + s)}$$

Siden graden i nevner er 2 pga. s^2 -leddet, tar vi $n = 2$, noe som gir:

$$K_{DS}(s) = \frac{-1}{\kappa} \frac{(s^2 + 2\beta s + \beta^2)}{(\alpha(\tau_{\phi}^2 s^2 + 2\tau_{\phi}s + 1) - \alpha + s)} = \frac{-1}{\kappa} \frac{(s^2 + 2\beta s + \beta^2)}{s(\alpha\tau_{\phi}^2 s + 2\alpha\tau_{\phi} + 1)}$$

7.1.2 Direktesyntese for systemer med tidsforsinkelser



Man kan som nevnt også bruke **Mod. direktesyntese** på et system med tidsforsinkelser,

$$P(s) = P_{MF}(s)e^{-\theta s}$$

Anta at vi ønsker $G_{LS}^{\theta}(s)$ på formen til et FOPTF-system, altså

$$G_{LS}^{\theta}(s) = \frac{e^{-\theta s}}{\tau_c s + 1}.$$

Vi kan da håndtere tidsforsinkelsen som dukker opp i K_{MDS} vha. to forskjellige strategier: 1. bruke den direkte i regulatoren eller 2. approksimere den (via Taylor- eller Padé-approksimasjoner).

Strategi 1: Fra (Mod. direkteyntese) har vi (se også 7.4)

$$K_{MDS}(s) = \frac{1}{\hat{P}(s)} \frac{e^{-\theta s}}{(\tau_c s + 1 - e^{-\theta s})} = \frac{1}{\hat{P}_{MF}(s)} \frac{1}{(\tau_c s + 1 - e^{-\theta s})}. \quad (7.3)$$

Tross forsinkelsen i nevner, er regulatoren (7.3) faktisk realiserbar:

Oppgave 7.2. Vis, ved hjelp av blokkdiagrammer, hvordan overføringsfunksjonen $U(s)/E(s) = K_{MDS}(s)$ gitt ved (7.3) kan realiseres hvis $P_{MF}(s)$ er minimum fase og $\frac{1}{P_{MF}(s)K_{MDS}(s)}$ er proper (høyeste grad i nevner er større eller like den høyeste graden i teller).

Importante! I vår modell $\hat{P}(s)$ av prosessen $P(s)$ vet vi ikke nødvendigvis tidsforsinkelsen θ nøyaktig. Det vil si at $\hat{P}(s) = \hat{P}_{MF}(s)e^{-\hat{\theta}s}$ hvor $\hat{\theta}$ er den *antatte* tidsforsinkelsen til systemet. Hvis $|\theta - \hat{\theta}|$ er stor, kan det skape krøll for modellbaserte metoder!

Strategi 2: En strategi som kan være noe mer robust mot unøyaktig $\hat{\theta}$ er å approksimere tidsforsinkelsen i nevneren i (7.3) som følger (se sek. 4.2.3): $e^{-\theta s} \approx 1 - \theta s$. Dette gir

$$K_{MDS,2}(s) = \frac{1}{\hat{P}_{MF}(s)} \frac{1}{(\tau_c s + 1 - (1 - \hat{\theta}s))} = \frac{1}{\hat{P}_{MF}(s)} \frac{1}{(\tau_c + \hat{\theta})s}. \quad (7.4)$$

Følgende eksempel tar i bruk dette for et FOPTF-system:

Eksempel 7.4. La prosessen være gitt av en FOPTF-modell:

$$P(s) = \frac{k e^{-\theta s}}{1 + \tau s}.$$

Regulatoren (7.4) er dermed

$$K_{MDS,2}(s) = \frac{1 + \tau s}{k} \frac{1}{(\tau_c + \theta)s},$$

altså en PI-regulator, som på serieform-form tilsvarer $k_p = \tau/(k(\tau_c + \theta))$ og $T_I = \tau$.

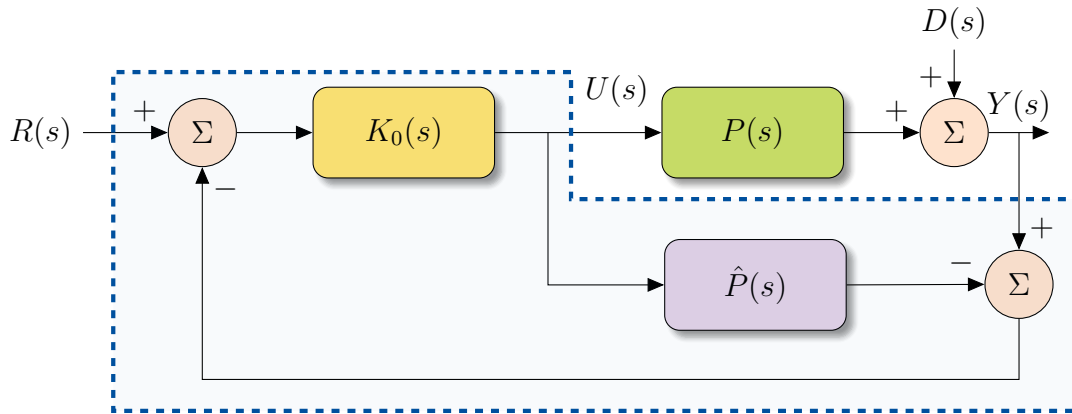
7.2. Intern-modell-kontroll



Alternative kilder: §12.2.2 i [Seborg et al., 2016].

Intern-modell-kontroll (IMK): Figur 7.2 gir en illustrasjon av metoden, hvor $\hat{P}(s)$ betegner en kjent modell^a av den virkelige prosessen $P(s)$. **Idé:** hvis $\hat{P}(s) \equiv P(s)$ og $d \equiv 0$, så oppnår man en kunstig åpen sløyfe: $Y(s) = P(s)K_0(s)R(s)$.

^aDet vil i realiteten alltid være forskjeller mellom vår modell, $\hat{P}(s)$, og den virkelige prosessen, $P(s)$.



Figur 7.2: Illustrasjon av intern-modell-kontroll. Det stiplede området tilsvarer regulatoren.

IMK vs direktesyntese: Intern-modell-kontroll (eng.: internal-model control (IMC)) metode er meget nært beslektet med direktesyntese. Metodene har mange likhetstrekk, og de produserer identiske regulatorer hvis designparametrene er spesifisert på en konsistent måte. I motsetning til direktesyntese, så har intern-modell-kontroll-metoden dog fordelen med at den tillater en å ta mer hensyn både til modellusikkerhet og avveininger mellom ytelse og robusthet på en mer systematisk måte.

Pass på! Som direktesyntese, kan ikke IMK-metoden brukes for ustabile systemer.^a

^aDet finnes dog flere modifiserte IMK-metoder for dette; se .feks. [Tan et al., 2003].

Design prosedyre: Fra figur 7.2 har vi

$$Y(s) = \frac{P(s)K_{IMK}(s)}{1 + K_0(s)(P(s) - \hat{P}(s))}R(s) + \frac{1 - \hat{P}(s)K_0(s)}{1 + K_0(s)(P(s) - \hat{P}(s))}D(s).$$

Merk også at

$$U(s) = \frac{K_0(s)}{(1 - K_0(s)\hat{P}(s))}E(s),$$

slik at intern-modell-kontroll-(IMK)-regulatoren tilsvarer en “standard” regulator fra $E(s) = R(s) - Y(s)$ til $U(s)$ som i figur 2.7 via følgende formel:

$$K_{IMK}(s) = \frac{K_0(s)}{(1 - K_0(s)\hat{P}(s))}. \tag{IMK-regulator}$$

Design-prosedyre for IMK:

Steg 1. Som med modifisert direktesyntese, så faktorer vi modellen $\hat{P}(s)$ i to deler:

$$\hat{P}(s) = \hat{P}_{IMF}(s)\hat{P}_{MF}(s),$$

hvor $\hat{P}_{MF}(s)$ består av delene som er minimum-fase (se § 2.6.6), mens $\hat{P}_{IMF}(s)$, $\hat{P}_{IMF}(0) = 1$, inneholder alt annet ubeleilig rot (tidsforsinkelser og nullpunkt i høyre halvplan).

Steg 2. For å få $G_{LS}^{\phi}(s) = G^{\phi}(s)P_{IMF}(s)$, $G_{LS}^{\phi}(0) = 1$, ta (IMK-regulator) med

$$K_0(s) = G^{\phi}(s)/\hat{P}_{MF}(s)$$

hvor $G^{\phi}(s)$ er et lavpass-filter som tilfredsstillers $G_{LS}^{\phi}(0) = 1$, f.eks

$$G^{\phi}(s) = 1/(\tau_c s + 1)^n, \quad n \in [1, 2, 3, \dots], \quad (7.5)$$

altså er n et positivt heltall, og τ_c de ønskede tidskonstantene til den lukkede sløyfen.

Eksempel 7.5. La prosessen være gitt av en første-orden-pluss-tidsforsinkelse-modell:

$$P(s) = \frac{ke^{-\theta s}}{1 + \tau s}.$$

Vi har dermed $P(s) = P_{MF}(s)P_{IMF}(s)$ hvor $P_{MF}(s) = k/(1 + \tau s)$ og $P_{IMF}(s) = e^{-\theta s}$.

Vi ønsker å designe en regulator basert på design prosedyren for intern-modell-kontroll gitt over, med $n = 1$. Vi får

$$K_0(s) = \frac{(1 + \tau s)}{k(1 + \tau_c s)}.$$

La $\hat{P}_{IMF}(s) = 1 - \theta s$ slik at $\hat{P}_{IMF}(s) \approx P_{IMF}(s)$ (se § 4.2.3). Ved å bruke (IMK-regulator) får vi

$$K_{IMK}(s) = \frac{K_0(s)}{1 - K_0(s)\hat{P}_{IMF}(s)\hat{P}_{MF}(s)} = \frac{1 + \tau s}{k(1 + \tau_c s) - k(1 - \theta s)} = \frac{1 + \tau s}{k(\tau_c + \theta)s},$$

altså en PI-regulator med $k_P = \tau/(k(\tau_c + \theta))$ og $T_I = \tau$. Legg merke til at den er identisk med den vi fant via direktesyntese i eks. 7.4.

Oppgave 7.3. Gjenta eksempel 7.5 når du bytter ut tidsforsinkelsen med en første-ordens Padé-approximasjon: $e^{-\theta s} \approx (1 - \frac{\theta}{2}s)/(1 + \frac{\theta}{2}s)$ og tar $r = 2$ i (7.5).

Merk: For såkalt lag-dominante systemer, hvor den Relative tidsforsinkelsen er liten ($\tau_r = \frac{\theta}{\theta + \tau} \ll 1$) så vil en PI-regulator som i eksempel 7.5 kunne føre til en treg forstyrrelses-respons siden integral-tiden T_I blir veldig stor. En mulig løsnings på dette er gitt av SIMC-metoden som vi skal se på i seksjon 7.3.

IMK for ustabile systemer

La oss til slutt se på et eksempel tatt fra [Tan et al., 2003] hvor man bruker en modifisert IMK-metode for et ustabil system:

Eksempel 7.6. IMK for ustabil system: Gitt et FOPTF-system:

$$P(s) = \frac{k \cdot e^{-\theta}}{\tau s - 1}.$$

Vi antar vi kjenner θ og at denne er liten. Vi bruker så følgende modell, som er baserte på første-ordens Padé-approximasjonen til tidsforsinkelsen:

$$\hat{P}(s) = \frac{k(1 - \frac{\theta}{2}s)}{(\tau s - 1)(1 + \frac{\theta}{2}s)}.$$

Vi tar så

$$K_0(s) = \frac{(\tau s - 1)(1 + \frac{\theta}{2}s)(\lambda s + 1)}{k(\tau_c s + 1)^3}.$$

I tillegg til kanselleringen av $(\tau s - 1)$, så er nullpunktet $(\lambda s + 1)$ nytt i forhold “standard” IMK-metoden; dets jobb er å sørge for at nevneren i (IMK-regulator), $1 - K_0(s)\hat{P}(s)$, kansellerer den ustabile polen til $\hat{P}(s)$. Dette holder hvis

$$\left(1 - K_0(s)\hat{P}(s)\right) \Big|_{s=1/\tau} = 0 \quad \implies \quad \lambda = \tau \left(\frac{(\frac{\tau_c}{\tau} + 1)^3}{(1 - \frac{\theta}{2\tau})} - 1 \right).$$

IMK-regulatoren $G_{IMK}(s)$ er da gitt av (IMK-regulator).

7.3. SIMC-metoden



Alternative kilder: §12.3.1 i [Seborg et al., 2016]; [Grimholt, 2018]; [Balchen et al., 2016, §9.3.3]; §2.7 i [Skogestad and Postlethwaite, 2007].

Vi skal nå se på Skogestad enkle intern-modell kontroll-metode (abbrivert SIMC) for PID-regulatorer; eller bare SIMC-metoden (eng.: simple internal model control) for enkelhets skyld.

Målet er å stille inn en PI(D)-regulator på serieform-form ved å anta en første-ordens-pluss-tidsforinskelse-prosess (FOPTF). Metoden for kun en PI-regulator er som følger:

Skogestad enkle intern-modell innstillings-regler (SIMC) for PI-regulatorer:

Gitt et FOPTF system (f.eks. funnet vha. en metode fra § 4.2.7 eller § 4.3)

$$G(s) = \frac{k e^{-\theta s}}{1 + \tau s},$$

ta parameterne til en PI-regulator,

$$K(s) = k_P \left(1 + \frac{1}{T_I s} \right),$$

som

$$k_P = \frac{1}{k} \frac{\tau}{(\tau_c + \theta)}, \quad T_I = \min(\tau, 4(\tau_c + \theta)). \quad (\text{SIMC-regler (FOPTF)})$$

der τ_c er ønsket tidskonstant for overføringsfunksjonen fra R til Y .

Men hvor har vi FOPTF-modellen fra? Metoder som SIMC krever at systemets approksimeres som en første- (eller andre-) ordens modell med tidsforsinkelse (FOPTF). Vi har tidligere gått gjennom to mulige alternativer for dette i disse notatene:

1. man kan finne/approksimere en slik modell ved å tilpasse en sprangrespons til modellen eller bruke Skogestads lukkede-sløyfe-metode (se kap. 4.3);
2. man kan forenkle en allerede kjent (høyere-ordens) modell ved å sammenslåing av tidskonstanter eller Skogestads halv-regel (se 4.2.7).

Hva med PID-regulatorer? Metoden kan også brukes til å stille inn PID-regulatorer:

Skogestad enkle intern-modell innstillings-regler (SIMC) for PID-regulatorer:

Gitt et AOPTF system

$$G(s) = \frac{ke^{-\theta s}}{(1 + \tau_1 s)(1 + \tau_2 s)}, \quad \tau_1 \geq \tau_2,$$

ta parameterne til en PID-regulator på serieform-form,^a

$$K(s) = k_P \left(1 + \frac{1}{T_I s} \right) (T_D s + 1),$$

som

$$k_P = \frac{1}{k} \frac{\tau_1}{(\tau_c + \theta)}, \quad T_I = \min(\tau_1, 4(\tau_c + \theta)), \quad T_D = \tau_2 + \frac{\theta}{3}, \quad (\text{SIMC-regler (AOPTF)})$$

hvor originalt $\alpha = 0$, mens $\alpha = 1$ er anbefalt i [Grimholt, 2018].

^aMan bør alltid bruke et Derivat-filter, hvor det anbefales at $T_f < T_D/3$; se [Grimholt and Skogestad, 2013]. Merk at et derivat-ledd gir liten effekt hvis den Relative tidsforsinkelsen, τ_r , er stor, men kan ellers gi en forbedring av både ytelse og robusthet.

Tommelfingerregler: (fra [Grimholt and Skogestad, 2013])

- $\tau_c = \frac{3}{2}\theta$ for robusthet;
- $\tau_c = \frac{1}{2}\theta$ for ytelse;
- $\tau_c = \theta$ for god balanse mellom disse.

Fordeler og ulemper med SIMC

Fordeler:

- Enkel og lett å huske.
- Det er kun en variabel som må stilles inn, nemlig τ_c ; det kan brukes til å få et ønsket kompromiss mellom ytelse og robusthet.
- Fungerer godt på mange prosesser, med tilnærmet optimal ytelse (til en PID-regulator å være) for visse relevante systemer (se [Grimholt and Skogestad, 2013]).

Ulemper: Fra [Balchen et al., 2016, §9.3.3] har vi følgende liste med eksempler på ting som kan føre til utfordringer med SIMC-metoden:

- Oscillatorisk åpen sløyfe prosess kan være vanskelig å håndtere.
- Langsomme nullpunkter (i venstre halvplan) kan gi til dels spesielle responser med stort oversving. Det kan da være utfordrende å anslå parametre i en første-ordens modell.
- Ustabile prosesser har ikke fornuftig sprangrespons, og metoden kan ikke brukes.
- Høyere ordens prosesser kan noen ganger kreve mer avansert regulering enn PI.
- Fenomener som ulineariteter kan gjøre det vanskelig å tune enkle PI-regulatorer, for både settpunkt og forstyrrelser kan påvirke en linearisert modell.
- I multivariable systemer kan interaksjoner i prosessen gjøre at innstillinger i andre regulatorsløyfer påvirker prosessmodellen som hver enkelt regulator ser.

Eksempel 7.7. SIMC vs IMK for FOPTF-system:

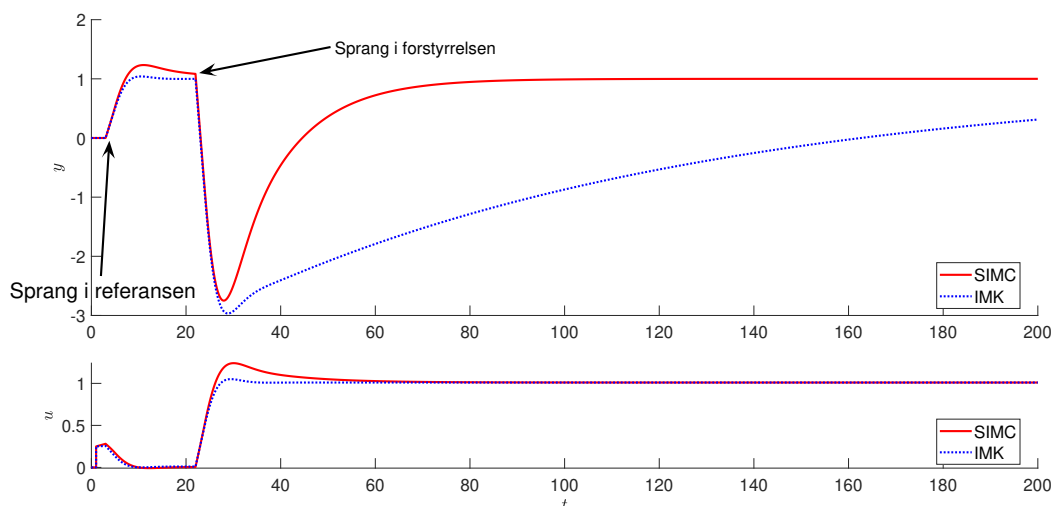
Gitt FOPTF-systemet:

$$P(s) = \frac{100}{100s + 1} e^{-2}.$$

Vi ønsker å regulere prosessen med en PI-regulator: $K(s) = k_P(1 + \frac{1}{T_I s})$. Hvis ønsket lukket sløyfe er

$$G_{LS}^{\circ}(s) = \frac{1}{\tau_c s + 1}, \quad \tau_c = 2,$$

så får vi fra IMK-/DS-metoden at $k_P = 1/4$ og $T_I = 100$ (meget stor!). SIMC-metoden med $\tau_c = 2$, derimot, gir den noe mer aggressive verdien $T_I = 16$. En sammenligning av responsene under de to metodene er vist i figur 7.3, hvor det er tydelig at IMK-regulatoren har en meget sløv respons til et sprang i forstyrrelsen.



Figur 7.3: Viser sprang- og forstyrrelses-responser for FOPTF-systemet i eks. 7.7 med en SIMC- og IMK-regulatorer.

7.3.1 Utleddning av SIMC-reglene

Det er tydelig at hvis $\tau < 4(\tau_c + \theta)$, så tilsvare SIMC-regler (FOPTF) de vi fikk fra direkte-syn-tese og IMK (se eksempel 7.4 og 7.5). Hvis dette derimot ikke holder, så er jo tidskonstanten τ stor relativ til tidsforsinkelsen θ (den Relative tidsforinskelsen er liten), noe som kan føre til treg forstyrrelses-respons. Siden τ da er stor, så kan vi anta at FOPTF-modellen kan approksimeres som $P(s) \approx k/(\tau s)$, altså som en rent integrerende prosess. Ta en PI-regulator $K(s) = k_p(1 + 1/(T_I s))$ hvor $k_p = \frac{1}{k} \frac{\tau}{(\tau_c + \theta)}$, slik at

$$K(s)P(s) = \frac{\tau(1 + 1/(T_I s))}{k(\tau_c + \theta)} \frac{k}{\tau s} = \frac{T_I s + 1}{T_I s^2(\tau_c + \theta)}.$$

Dermed blir overføringsfunksjonen til den lukkede sløyfen

$$G_{LS}(s) = \frac{K(s)P(s)}{1 + K(s)P(s)} = \frac{T_I s + 1}{T_I s^2(\tau_c + \theta) + T_I s + 1}.$$

Vi kan skrive dette på følgende form:

$$G_{LS}(s) = \frac{T_I s + 1}{T_I(\tau_c + \theta)(s^2 + 2\zeta\omega_0 s + \omega_0^2)}.$$

hvor ω_0 og ζ er henholdsvis den udedempede svingefrekvensen og den relative dempningsfaktoren (se sek. 4.1.2); disse må tilfredsstill

$$\omega_0^2 = \frac{1}{T_I(\tau_c + \theta)} \quad \text{og} \quad 2\zeta\omega_0 = \frac{1}{(\tau_c + \theta)}.$$

For at dette skal tilsvare en kritisk dempet respons, det vil si $\zeta \equiv 1$, og dermed $2\omega_0 = 1/(\tau_c + \theta)$, så må

$$\omega_0^2 = \frac{1}{2^2(\tau_c + \theta)^2} = \frac{1}{T_I(\tau_c + \theta)}.$$

Ved å løse dette med hensyn på T_I får vi uttrykket vi var ute etter, nemlig

$$T_I = 4(\tau_c + \theta).$$

7.4. Smith-prediktoren

▶ UV0AK8Rh9pU&t=3527

Alternative kilder: §16.2 i [Seborg et al., 2016]; §9.7.2 i [Balchen et al., 2016].

Hvis man jobber med en prosess som har en tidsforsinkelse (dødtid) i pådraget, så kan man noen ganger bruke en såkalt Smith-prediktor til å kompensere for denne tidsforsinkelsen:

Smith-prediktoren: La $K(s)$ betegne en *nominell* regulator (f.eks. PI) for en prosess $P(s) = P_0(s)e^{-\theta s}$ med en tidsforsinkelse θ (vi vet kun en tilnærmet verdi $\hat{\theta}$). Tilsvarende regulator (altså overføringsfunksjonen fra $E(s)$ til $U(s)$) med en Smith-prediktor er da

$$K_{SP}(s) = \frac{K(s)}{1 + K(s)\hat{P}_0(s)(1 - e^{-\hat{\theta}s})}, \quad (\text{Regulator med Smith-prediktor})$$

hvor \hat{P}_0 er en “ideell” modell av prosessen *uten* tidsforsinkelsen, altså $P(s) \approx \hat{P}_0(s)e^{-\hat{\theta}s}$.

Idé: Et blokkdiagram av en (Otto-)Smith-Prediktor er vist i figur 7.4. Fra dette diagrammet er det tydelig at hvis det ikke er noen forstyrrelser ($d \equiv 0$) og vi har en perfekt modell av systemet ($P(s) \equiv \hat{P}_0(s)e^{-\theta s}$), så er overføringfunksjonen til den indre lukkede sløyfen

$$G_{ls}^o(s) = \frac{K(s)P_0(s)}{1 + K(s)P_0(s)}$$

som er lik den indre lukkede sløyfen under regulatoren $K(s)$ *uten* noen tidsforsinkelse, mens den lukkede sløyfen

$$\frac{Y(s)}{R(s)} = G_{LS}(s) = G_{ls}^o(s)e^{-\theta s} = \frac{K(s)P_0(s)}{1 + K(s)P_0(s)}e^{-\theta s}$$

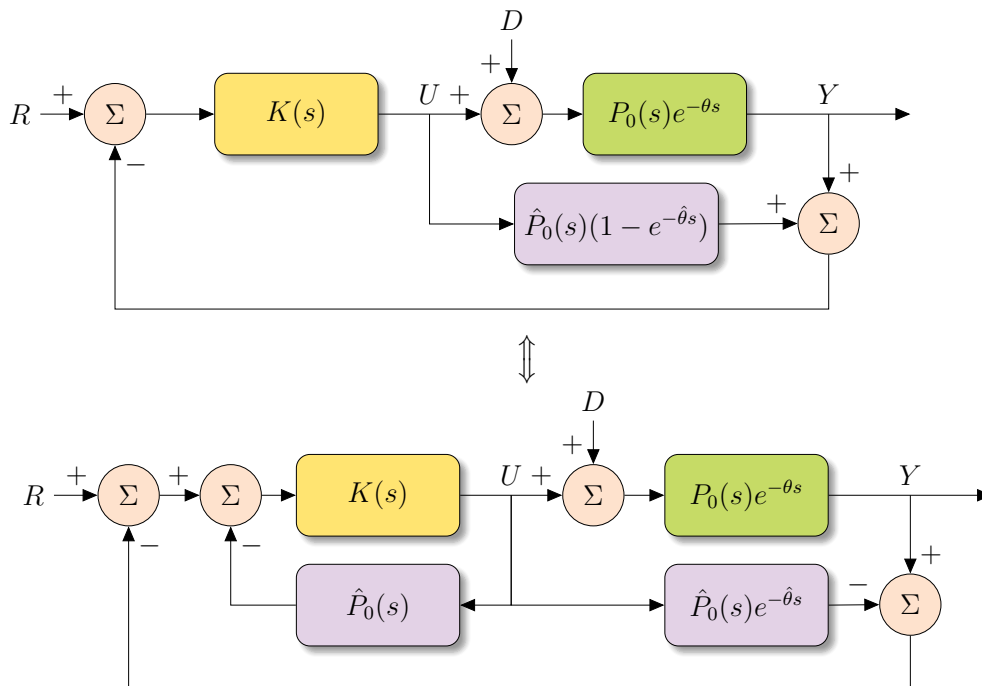
tilsvarende en respons kun forsinket med tidsforsinkelsen. Med andre ord: Vi kan (i hvert fall i den ideelle verden) designe en regulator for prosessen uten tidsforsinkelsen, og så i ettetid legge til en Smith-prediktor for å fjerne effekten av tidsforsinkelsen på regulatoren! Merk også at dette tilsvarende det vi gjorde i § 7.1.2, men da fra et noe annet perspektiv.

Er Smith-prediktoren bedre enn PID-regulatorer?

TL;DR: Ifølge [Ingimundarson and Hägglund, 2001, Skogestad, 2018, Grimholt, 2018] er svaret generelt nei.

Basert på konklusjonen vår i forrige paragraf, så virker jo en Smith-prediktor som en veldig naturlig og lovende metode for å håndtere tidsforsinkelser. Det er dog noen spørsmål vi bør tenke litt over:

- Hvor lett er den å implementere?



Figur 7.4: Blokkdiagram av en Smith-prediktor, hvor $\hat{P}_0(s)e^{-\hat{\theta}s}$ er modellen av $P(s) = P_0(s)e^{-\theta s}$.

- Hvor god bør vår modell av systemet være?
- Hva er effekten av forstyrrelser?
- Er den robust mtp. modellfeil og usikkerhet?

I forbindelse med det første spørsmålet, så krever det, hvis vi tar utgangspunkt i en førsteordens-plus-tidsforsinkelse (FOPTF) modell (se neste seksjon) og en PI-regulator, at vi må “stille inn” totalt fem parametere: PI parameterne K_p og T_I ; samt FOPTF-parameterne K_n , τ_n , og τ_d .

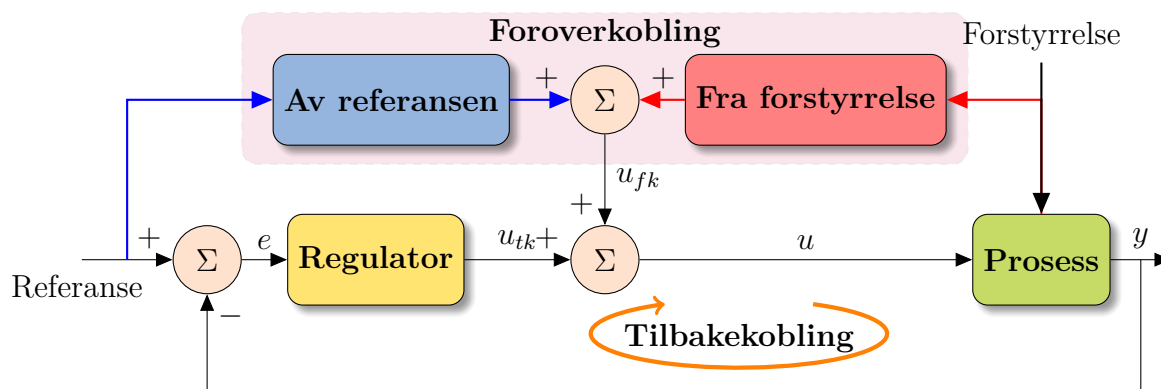
Vårt estimat av tidsforsinkelsen må også være svært godt; det må hverken være for stort eller for lite siden Smith-prediktoren er ikke veldig robust mtp. dette. Faktisk vil Smith-prediktoren kunne gi betraktelig dårligere regulering hvis vi har overestimert tidsforsinkelsen — et fenomen vi ikke får med PI(D)-regulatorer, hvor mindre tidsforsinkelse som regel fører til bedre regulering!

8. Foroverkobling og referansefølging

I dette kapitlet skal vi se på en metode som noen ganger kan brukes sammen med tilbakekobling for å forbedre en regulators ytelse, nemlig foroverkobling. Vi skal også se hvordan dette blant annet kan brukes til å kompensere for visse forstyrrelser, samt hvordan en såkalt analytisk foroverkobling kan gi bedre referansefølging.

8.1. Foroverkobling og nominelle pådrag

Alternative kilder: Kap. 15 i [Seborg et al., 2016]; Brian Douglas video.



Figur 8.1: Foroverkoblinger av både referansen (i blått) og forstyrrelsen (i rødt).

8.1.1 Hva er foroverkobling? [▶ mGP1hy57UTg&t=18](#)

Foroverkobling er en reguleringsstrategi hvor man bruker kunnskap om systemets dynamikk, samt kunnskap om ønsket referanse og/eller forstyrrelser som virker på systemet til å bestemme pådraget. For eksempel kan man i en foroverkobling bruke målinger av forstyrrelsene til å

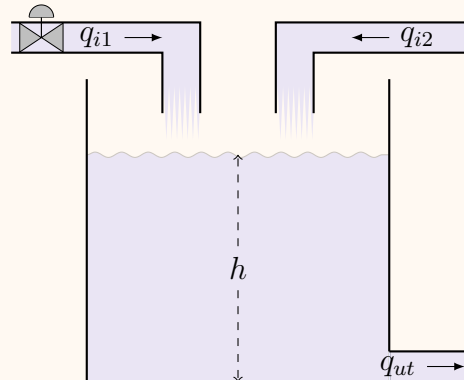
forutsi og korrigerer for deres effekt på systemet. Det brukes nesten alltid i kombinasjon med tilbakobling, slik som vist i figur 8.1.

Eksempel 8.1. (Motiverende eksempel)

Gitt et tanksystem som vist i figuren til høyre, med to strømmer, q_{i1} og q_{i2} , av samme væske inn, og én utstrøm, q_{ut} . Man kan endre q_{i1} vha. en reguleringsventil, mens q_{i2} har man ingen kontroll over.

Det er ønskelig å regulere væskehøyden h i tanken til en ønsket verdi, h_ϕ . Ved å anta at $q_{ut} = c_{ut}\sqrt{h}$, så har vi

- Utgang/tilstand: $y = h$;
- Referanse/settpunkt: $r = h_\phi$;
- Pådrag: $u = q_{i1}$;
- Forstyrrelse: $d = q_{i2}$.



Hvis vi også antar at væsken har konstant massetetthet ρ og at tanken har konstant tverrsnittsareal A , så får vi, ved å ta $a = c_{ut}/A$ og $b = 1/A$, at systemets dynamikk er beskrevet av

$$\dot{y} = -a\sqrt{y} + b(u + d). \tag{Proessedynamikk}$$

Scenario 1 – Trenger bare tilbakobling: Hvis både r og d er konstante eller endrer seg svært sakte, så vil en godt innstilt PI-regulator fungere bra nært et ønsket arbeidsområde.

Scenario 2 – Behov for foroverkobling: Hvis $r(t)$ og/eller $d(t)$ endrer seg (varierer med tiden), så vil ikke en PI-regulator fungere godt på egenhånd ved raske endringer grunnet tregheten i I-leddet. Her kan foroverkobling være en mulig løsning. For å finne ut hvordan en slik foroverkobling kan se ut, definerer vi avviket $e(t) = r(t) - y(t)$ og finner avviksdynamikken:

$$\dot{e} = \dot{r} - \dot{y} = \dot{r} + a\sqrt{y} - b(u + d). \tag{Avviksdynamikk}$$

Vi ønsker jo at avviket skal bli null, noe vi oppnår hvis vi velger pådraget u slik at $\dot{e} < 0$ for alle $e \neq 0$. En kandidat som oppnår dette er

$$u = \underbrace{\frac{a\sqrt{y}}{b}}_{\text{Kansellering av ulineær del}} + \underbrace{\frac{\dot{r}}{b}}_{\text{Foroverkobling fra referansen}} + \underbrace{\frac{-d}{b}}_{\text{Foroverkobling av forstyrrelsen}} + \underbrace{\frac{k_p e}{b}}_{\text{P-regulator}}$$

siden man da får $\dot{e} = -k_p e$, slik at $y(t) = r(t) + (y(0) - r(0)) \exp(-k_p t)$. Alternativt kan man bruke et **nominelt pådrag** fremfor å kansellere det ulineære leddet (se rød boks under):

$$u = \underbrace{\frac{a\sqrt{r}}{b}}_{\text{nominelt pådrag}} + \underbrace{\frac{\dot{r}}{b}}_{\text{Foroverkobling fra referansen}} + \underbrace{\frac{-d}{b}}_{\text{Foroverkobling av forstyrrelsen}} + \underbrace{k_p e + k_I \int_0^t e(\tau) d\tau}_{\text{PI-regulator}}$$

Foroverkobling fra referansen inkluderer i dette tilfelle også det nominelle pådraget. Legg dog også merke til at vi ikke tar hensyn til noen tidsforsinkelse her.

NB! Kansellering av ulineære ledd, såkalt tilbakekoblings-linearisering (eng. “feedback linearization”), er teoretisk sett veldig effektivt, men man skal være forsiktig med dette i praksis. Grunn: ved unøyaktig modell eller måling gjør dette leddet fort mer skade enn nytte. Alternativer: bruke ønsket settpunkt/referansen i stedet for tilstandene (tilsvarende det nominelle pådraget i regulatoren over), eller fjerne hele leddet til fordel for et stort P -ledd eller en PI-regulator.

8.1.2 Tilbakekobling vs foroverkobling

▶ [mGP1hy57UTg&t=1151](#)

Tilbakekobling: bruker målinger av utgangen/tilstandene man ønsker å regulere (væskedøden i en tank, hastigheten til en bil, etc.) til å bestemme pådraget for å korrigere for feil.

Fordeler:

- Regulatoren handler så snart prosessvariabelen avviker fra referansen, uavhengig av årsak.
- Tilbakekobling krever i utgangspunktet minimalt med kunnskap om prosessen som skal reguleres; f.eks. er ikke en matematisk modell av prosessen nødvendig, selv om det kan være veldig nyttig for både regulator-design og -tuning.
- PID-regulatorer fungerer tilfredsstillende på de fleste systemer; de er både allsidig og robuste, og kan lett stilles inn på nytt hvis prosessen endrer seg.

Ulemper:

- Regulatoren reagerer først når et avvik har oppstått. Dermed er perfekt regulering under forstyrrelses- eller settpunkt-endringer teoretisk umulig.
- Alltid kompromiss mellom ytelse og robusthet.
- Ikke proaktiv/prediktiv i forhold til kjente eller målbare endringer forstyrrelser og referansen.
- Kan gi utilfredsstillende regulering for trege prosesser (store tidskonstanter og/eller lange tidsforsinkelser). Ved varierende forstyrrelser er det dermed ikke alltid mulig å få prosessen til ønsket referanse.
- I noen situasjoner kan man ikke direkte måle prosessvariabelen, slik at tilbakekobling ikke er direkte gjennomførbart.

Foroverkobling: bruker målinger/informasjon om variabler/signaler (forstyrrelser eller ønsket referanse) man *ikke* kan/er ute etter å regulere for å bestemme pådraget (i kombinasjon med en tilbakekobling), slik at man ideelt sett oppnår forbedret regulering.

Fordeler:

- Utnytter ekstra, tilgjengelig informasjon til å forbedre regulatorens ytelse og/eller robusthet.
- Reagerer umiddelbart på endringer i forstyrrelser og/eller referansen.
- For lineære systemer påvirker ikke foroverkoblingen stabiliteten til systemet.

NB! Holder generelt sett ikke for ulineære systemer, altså for de fleste ekte systemer!

Ulemper:

- Forstyrrelser må måles, noe som krever ekstra sensorer, og som for mange applikasjoner ikke er gjennomførbart.
- Effektiv bruk av foroverkobling krever en god matematisk modell av systemets dynamikk.
- Ideelle foroverkoblinger er ofte ikke realiserbare (praktiske tilnærminger er dog mulig).
- Foroverkobling av forstyrrelsen kan noen ganger komme i “konflikt” med tilbakekoblingen, ved at de ved et sprang i forstyrrelsen begge prøver å løse samme problem.

Fun facts, bemerkninger og annet dill dall (you may skip)

Ofte har man ikke direkte målinger av forstyrrelser tilgjengelig. Har man derimot en god modell av det dynamiske systemet, samt måling av tilstandene, så kan man i foroverkoblingen i stedet bruke et estimat av en forstyrrelse som er generert vha. en matematisk modell, såkalte **forstyrrelses-estimatorer** (eller “disturbance observers” på engelsk).

Hvorfor da bruke foroverkobling sammen med tilbakekobling? Som nevnt, så går foroverkobling ut på å utnytte annen tilgjengelig informasjon/målinger enn bare tilstandsmålinger til å beregne pådraget. Dette er typisk informasjon relatert til det ønskede arbeidspunktet/referansen eller en eller flere forstyrrelser.

Design prosedyre for regulator med både tilbakekobling og foroverkobling (regulator med to frihetsgrader):

1. Designe *tilbakekoblingen* for å:

- minimere sensitiviteten til forstyrrelser;
- minimere/attenuere effekten av målestøy;
- øke robustheten mtp. usikkerhet og variasjoner i prosessen.

2. Deretter designe *foroverkobling* for å oppnå ønsket respons gitt referansen $r(t)$, samt raskt eliminere effekten av forstyrrelser.

Merk:¹ For mange problemer relatert til prosessregulering er lastforstyrrelsesresponsen mye viktigere enn settpunktresponsen. Settpunktresponsen er viktigere i bevegelseskontroll. Få lærebøker og vitenskapelig artikler viser dog dessverre mer enn settpunkt betraktninger.

8.1.3 Nominelt pådrag

▶ L9R5cSvFKqA&t=866s

Et nominelt pådrag kan ses på som en type foroverkobling. La oss minne oss på definisjonen:

Nominelt pådrag: Pådraget som opprettholder et ønsket arbeidspunkt selv uten et integral-ledd. For et dynamisk system $\dot{x} = f(x, u)$ og et arbeidspunkt, x_a , så må det

¹Sitat av prof. Anders Robertsson, Univ. i Lund; se også Shinskeys bemerkning (f.eks [Shinskey, 2002]).

nominelle pådraget, u_{nom} , tilfredsstillende $f(x_a, u_{nom}) = 0$.

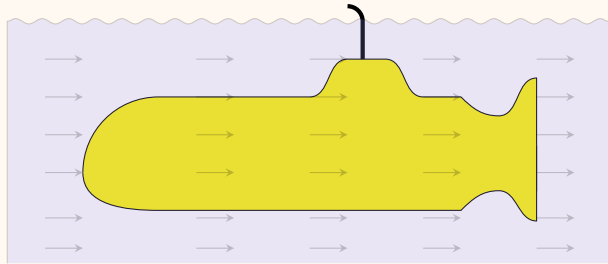
Tross definisjonen over, så ser man at i praksis (spesielt i prosessindustrien) at det nominelle pådraget ofte brukes kun som et «manuelt» pådrag man setter til en ønsket verdi. Det er ofte også skrudd av til fordel for et aktivt integral-ledd. Dette er helt OK, og på ingen måte direkte «feil». Jeg vil dog at dere alltid skal tenke på det i form av det som står i boksen over, siden dette lar seg generalisere både langt og bredt.

Det vil si: tenk på det nominelle pådraget som det som utfører den ønskede oppgaven i en ideell (perfekt) verden; det trenger ikke være konstant, det trenger ikke være basert på målte verdier, det trenger ikke basere seg kun på fortid og nåtid. Akkurat hva jeg mener her blir (forhåpentligvis) mer klart når vi skal se på analytiske foroverkoblinger fra referansen.

Eksempelet under gir også innsikt i hvordan nominelt pådrag alternativt kan brukes. Spesifikt er hensikten med eksempelet under at man f.eks. kan bruke det nominelle pådraget som et erfarings-basert pådrag, hvor man inkorporerer kunnskapen man har opparbeidet seg som systemet og eventuelle forstyrrelser, etc.

Eksempel 8.2. Spionubåt i sterk strøm: I forbindelse med sitt meget suksessfulle få-tak-i-rikinger-fra-norge-program, har Sveits bedt sin marine sende deres beste (og eneste) ubåt til de norske fjorder for å autonomt innhente informasjon om norske laksebaroner.

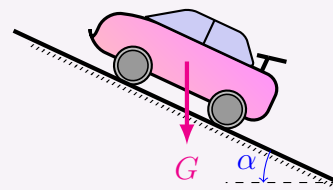
Selv om den autonome ubåten er utstyrt med mekaniske tilbakekoblingsløyfer basert på sveitsisk uverk av ypperste kvalitet, så har de sterke havstrømmene i fjordene gjort det vanskelig å opprettholde en ønsket posisjon.



Heldigvis viser strømmene seg å variere sakte, samt at de svært forutsigbare i forhold til tidevannet (flo og fjære), slik at de over tid kan kartlegges for de forskjellige fjordene. En effektiv løsning er derfor å kompensere for dem vha. et nominelt pådrag som tidsvis blir (automatisk) oppdatert i forhold til lokasjon og tid på døgnet.

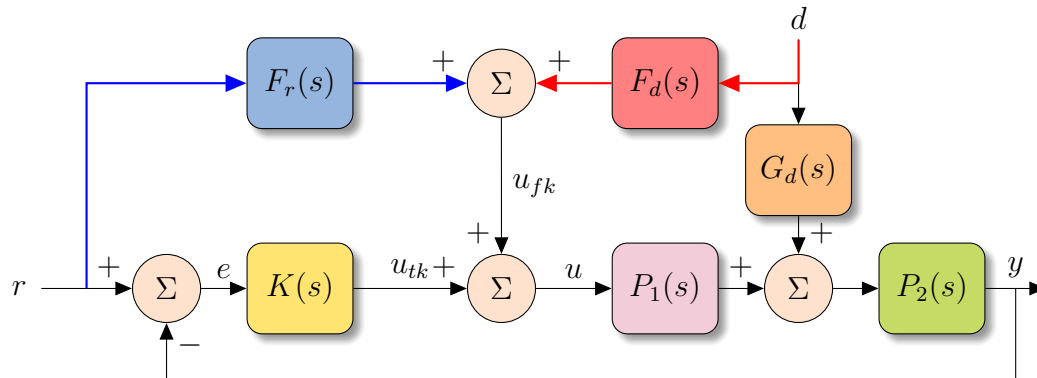
Oppgave 8.1. Bil opp en bakke:

En bil med masse $m = 1500$ kg kjører opp en bakke med helningsgrad $\alpha = 25^\circ$. Bilen skal holde en konstant hastighet på 50 km/t. Hjulradiusen er $r = 30$ cm. Det er ønskelig å utvikle en cruise-control for dette systemet, i form av en P-regulator med nominelt pådrag.



Spørsmål: Gitt at bilen har firhjulstrekk, hva skal det nominelle pådraget være, i form av et dreiemoment som virker likt på hvert av hjulene, for situasjonen over?

8.1.4 Ideelle foroverkoblinger for lineære systemer



Figur 8.2: Fremoverkoblinger av både referansen (i blått) og forstyrrelsen (i rødt).

Vi vil nå se på hvordan man kan designe foroverkobling for et system som vist i figur 8.2. Vil vil både se på foroverkobling fra referansen (se den blå delen i figur 8.2) og fra en forstyrrelse (se den røde delen i figur 8.2). Fra figuren kan man se at

$$Y(s) = P_2(s) [(s)D(s) + P_1(s)U(s)]$$

hvor

$$U(s) = F_d(s)D(s) + F_r(s)R(s) + K(s)E(s),$$

slik at når $E(s) = 0$ så har man

$$Y(s) = P_2(s) [G_d(s) + P_1(s)F_d] D(s) + P_2(s)P_1(s)F_r(s)R(s).$$

Ideelle foroverkoblinger: Effekten av forstyrrelsen, $d(t)$, elimineres (i teorien) ved å ta

$$F_d^{id}(s) = -\frac{G_d(s)}{P_1(s)}. \quad \text{(ideell foroverkobling av forstyrrelsen)}$$

Mens følgende ideele foroverkobling fører (teoretisk sett) til perfekt følging av referansen:

$$F_r^{id}(s) = \frac{1}{P_1(s)P_2(s)} \quad \text{(ideell foroverkobling fra referansen)}$$

NB! Dessverre er disse *ideelle* foroverkoblingene **sjeldent realiserbare** (se § 2.6.7):

- For at $F_d^{id}(s)$ skal være realiserbar så må:
 1. $G_d(s)/P_1(s)$ være proper (se § 2.6.4) og stabil;
 2. $G_d(s)$ må være kjent;
 3. tidsforsinkelsen i $P_1(s)$ må være kortere enn en eventuell tidsforsinkelse i $G_d(s)$;

Matchet forstyrrelse: Vi sier at en forstyrrelse $d(t)$ er *matchet* (i forhold til $u(t)$) hvis $G_d(s) = \alpha \cdot P_1(s)$ for et reelt tall α , slik at $F_d^{lp}(s) = -\alpha$.

- For at $F_r^{lp}(s)$ skal være realiserbar så må:
 1. $1/(P_1(s)P_2(s))$ må være proper (se § 2.6.4) og stabil;
 2. $P_1(s)$ og $P_2(s)$ kan ikke inneholde tidsforsinkelser.

Eksempel 8.3. (Ikke-realiserbare foroverkoblinger fra forstyrrelsen)

- Gitt $P_1(s) = \frac{1}{(2s+1)(s+3)}$ og $G_d(s) = \frac{1}{s+2}$, så er ikke $F_d(s) = -\frac{(2s+1)(s+3)}{s+2}$ realiserbar.
- Gitt $P_1(s) = \frac{1}{(s+1)}e^{-2s}$ og $G_d(s) = \frac{1}{s+1}e^{-s}$, så er ikke $F_d(s) = -\frac{(s+1)e^{2s}}{(s+1)e^s} = -e^s$ realiserbar.
- Gitt $P_1(s) = \frac{1}{(s+1)}$ og $G_d(s) = \frac{1}{s-1}$, så er ikke $F_d(s) = -\frac{(s+1)}{(s-1)}$ realiserbar (den er ustabil).

Men: Selv om disse overføringsfunksjonene ikke skulle være realiserbare, så kan vi fote bruke disse til å utvikle alternative foroverkoblinger. For dette skal vi se på neo alternativer:

- 1) Tilnærme en stasjonære eller dynamisk foroverkobling basert på den ideelle;
- 2) Utlede en analytisk-basert foroverkobling i tidsdomenet (hovedsaklig kun fra referansen).

Vi skal se på alternativ 1) i det som følger, mens 2) skal vi se nærmere på i seksjon 8.2.

8.1.5 Utlede realiserbare tilnærmede foroverkoblinger

mGPIhy57UTg&t=2804

Anta at vi kjenner en *ideell foroverkobling av forstyrrelsen*, $F_d(s)$, som ikke er realiserbar. Vårt mål nå er å se på forskjellige tilnærmingsmåter som lar oss bruke kunnskapen om $F_d(s)$ til å finne en alternativ foroverkobling, $\hat{F}_d(s)$, som faktisk er realiserbar.

Stasjonær foroverkoblinger

mGPIhy57UTg&t=2816

Den enkleste alternative foroverkoblingen er en stasjonær foroverkobling:

$$\hat{F}_d^{stat} = F_d(0). \quad (\text{stasjonær foroverkobling})$$

En slik foroverkobling vil kunne fungere greit for forstyrrelser som endres sakte, men dårlig for hurtigendrende forstyrrelser. Merk dog at hvis vi har en matchet forstyrrelse (se forrige avsnitt), så er den stasjonære foroverkoblingen den ideelle.

Eksempel 8.4. (Stasjonære foroverkoblinger)

- Gitt $F_d(s) = -\frac{(2s+1)(s+3)}{s+2}$, så har vi $\hat{F}_d^{stat} = -\frac{3}{2}$.

- Gitt $F_d(s) = -\frac{(s+1)e^{2s}}{(s+1)e^s} = -e^s$, så har vi $\hat{F}_d^{stat} = -1$.
- Gitt $F_d(s) = -\frac{(s+1)}{(s-1)}$, så har vi $\hat{F}_d^{stat} = 1$.

Dynamisk tilnæringer av ikke-proper foroverkoblinger



Nåværende problem: En kjent **ideell foroverkobling av forstyrrelsen**, $F_d(s)$, er ikke proper (se § 2.6.4), men stabil og uten ledd på formen $e^{\theta s}$. Vi ønsker å finne en god dynamisk tilnærming, $\hat{F}_d^{dyn}(s)$, til den ideelle foroverkoblingen.

Problemet er altså at vi har høyere-orden i teller enn i nevner, noe som motiverer følgende: vi kan legge til et lavpass-filter av ønsket i orden i serie med det ideelle filteret. Som regel er det alltid både ønskelig og lurt å filtrere målte signaler for å fjerne målestøy, så slik sett er dette naturlig i seg selv. La oss ta et eksempel:

Eksempel 8.5. Gitt $G_d(s) = \frac{k}{(\tau_2 s + 1)}$ og $P(s) = \frac{k}{(\tau s + 1)(\tau_2 s + 1)}$ slik at $F_d(s) = -(\tau s + 1)$.

Alternativ 1: For $0 < \alpha \ll 1$, ta

$$\hat{F}_d^1(s) = -\frac{(\tau s + 1)}{(\alpha \tau s + 1)}.$$

Altså et lavpass-filter med liten tidskonstant; det er ønskelig å ta denne så liten som mulig uten å forsterke unødvendig mye støy.

Alternativ 2: For $0 < \alpha \ll 1$, legg merke til at den ideelle foroverkobling kan skrives som

$$F_d(s) = -(\tau s + 1) \frac{(\alpha \tau s + 1)}{(\alpha \tau s + 1)} = -\frac{(\alpha \tau^2 s^2 + (1 + \alpha)\tau s + 1)}{(\alpha \tau s + 1)}.$$

Ved lave frekvenser er leddet $\alpha \tau^2 s^2$ lite, slik at vi kan tilnærme den ideelle som følger:

$$\hat{F}_d^2(s) = -\frac{((1 + \alpha)\tau s + 1)}{(\alpha \tau s + 1)}.$$

Man kan kombinere disse metodene på en rekke måter ved å slå sammen tidskonstante slik vi gjorde i § 4.2.7. For enkelhets skyld, ser vi bare nærmere på to metoder:

Tilnærming av ikke-propre foroverkoblinger: Gitt en ikke-proper, stabil, ideell foroverkobling på følgende form

$$F(s) = \frac{k(1 + \ell_1 s) \cdots (1 + \ell_m s)}{(1 + \tau_1 s)(1 + \tau_2 s) \cdots (1 + \tau_n s)} = k \frac{1 + (\ell_1 + \cdots + \ell_m)s + \mathcal{O}(s^2)}{1 + (\tau_1 + \tau_2 + \cdots + \tau_n)s + \mathcal{O}(s^2)},$$

hvor da $m > n$ og $\tau_1 \geq \tau_2 \geq \cdots \geq \tau_n > 0$.

Metode 1 (lavpass-filter): Ta

$$\hat{F}(s) = F(s)G_{LPF}(s)$$

hvor $G_{LPF}(s)$ er et lavpass-filter (LPF) av orden $m - n$, hvor den største tidskonstanten er mindre enn τ_n , f.eks.

$$G_{LPF}(s) = \frac{1}{(\alpha\tau_n s + 1)^{m-n}}, \quad 0 < \alpha \ll 1.$$

Metode 2 (lead-lag): For $0 < \alpha \ll 1$, ta

$$\hat{F}(s) = k \frac{1 + (\ell_1 + \dots + \ell_m + (m - n)\alpha\tau_n)s}{1 + (\tau_1 + \tau_2 + \dots + (1 + (m - n)\alpha)\tau_n)s}.$$

Tilnærmede foroverkoblinger i systemer med tidsforsinkelser

▶ mGPIhy57UTg&t=3930

Nåværende problem: En kjent **ideell foroverkobling** av **forstyrrelsen**, $F_d(s)$, er proper (se § 2.6.4) og stabil, men har et ledd på formen $e^{\theta s}$ for $\theta > 0$. Vi ønsker å finne en god dynamisk tilnærming, $\hat{F}_d^{dyn}(s)$, til den ideelle foroverkoblingen.

Vi vil her ta i bruk noen av metodene vi så på i § 4.2.3 og § 4.2.7. Spesielt tilnærmingen

$$e^{\theta s} \approx 1 + \theta s$$

vil være nyttig. Mer dog at f.eks. Padé-approksimasjonen $e^{\theta s} = \frac{1 + \frac{\theta}{2}s}{1 - \frac{\theta}{2}s}$ her ikke kan brukes siden den introduserer en pol i høyre halvplan.

Eksempel 8.6. (tilnærming ved positivt nullpunkt) Den ideelle foroverkoblingen

$$F(s) = \frac{1 + 2s}{(3s + 1)} e^s$$

er ikke realiserbar pga. e^{1s} -leddet. Vi kan dog bruke at $e^s \approx 1 + s$, samt lavpass-filtermetoden fra forrige avsnitt til å motivere følgende realiserbare tilnærming:

$$\hat{F}(s) = \frac{(1 + 2s)(1 + s)}{(3s + 1)(1 + \alpha s)}, \quad 0 < \alpha \ll 1.$$

Det er også flere andre scenarier hvor metoden fra § 4.2.3 og § 4.2.7 her kan brukes. Vi tar to eksempler:

Eksempel 8.7. (negativt nullpunkt som tidsforsinkelse) Den ideelle foroverkoblingen

$$F(s) = \frac{1 - 2s}{(3s + 1)} e^{2s}$$

er ikke realiserbar pga. e^{2s} -leddet. Vi kan dog bruke at $e^{-2s} \approx 1 - 2s$ til å motivere følgende realiserbare tilnærming:

$$\hat{F}(s) = \frac{1}{3s + 1}.$$

Eksempel 8.8. (Skogestads halv-regel) Den ideelle foroverkoblingen

$$F(s) = \frac{1 + 5s}{(4s + 1)(5s + 1)} e^{2s}$$

er ikke realiserbar pga. e^{2s} -leddet. Vi kan dog bruke Skogestads halv-regel (se § 4.2.7) til å motivere følgende realiserbare tilnærming:

$$\hat{F}(s) = \frac{1 + 5s}{((5 + \frac{4}{2})s + 1)} e^{(2 - \frac{4}{2})s} = \frac{1 + 5s}{7s + 1}.$$

8.1.6 Analytisk foroverkobling fra referansen

▶ mGPIhy57UTg&t=4312

I § 8.1.5 så vi på måter å tilnærme ideelle foroverkoblinger som ikke var realiserbare, ved at de f.eks. ikke var proper eller krevde invertering av en tidsforsinkelse. Når det kommer til referansen, så er det dog ofte sånn (spesielt i bevegelsesstyring) at vi kjenner dette signalet eksakt, og ofte vet vi også hva det skal være en stund frem i tid. Følgende strategi er derfor noen ganger mulig:

Idé: Regne ut (hvis mulig) $u_{fk}(t) = \mathcal{L}^{-1} \{F_r^{inv}(s)R(s)\}$.

Vi nøyer oss med et eksempel for å illustrere denne strategien (som er mye brukt i bevegelsesstyring/servokontroll):

Eksempel 8.9. (Analytisk foroverkobling fra referansen) Gitt en andre-ordens prosess med tidsforsinkelse:

$$P(s) = P_1(s)P_2(s) = \frac{e^{-3s}}{s^2 + 2s + 1}.$$

Den **ideell foroverkobling fra referansen**, $F_r^{inv}(s) = 1/(P_1(s)P_2(s)) = (s^2 + 2s + 1)e^{3s}$, er utvilsomt ikke realiserbar.

Vi har derimot at

$$\mathcal{L}^{-1} \{(s^2 + 2s + 1)R(s)e^{\theta s}\} = \ddot{r}(t + \theta) + 2\dot{r}(t + \theta) + r(t + \theta).$$

Ved å anta at målet er å følge referansen $r(t) = \sin(2t)$, så har vi jo $\ddot{r}(t) + 2\dot{r}(t) + r(t) = -4\sin(2t) + 4\cos(2t) + \sin(2t)$, slik at vi derfor kan bruke følgende analytisk-baserte foroverkobling fra referansen:

$$u_{fk}(t) = 4\cos(2(t + 3)) - 3\sin(2(t + 3)).$$

Dette krever selvsagt at **1)** vi kjenner referansen som en glatt (kontinuerlig differensierbar) funksjon, og **2)** vi vet hva dens verdi skal være minst 3 sekunder frem i tid.

8.1.7 Foroverkobling basert på lead-lag-element



Alternative kilder: §6.1 og 15.4 i [Seborg et al., 2016].

Nåværende problem: Eksperimentelt innstille en lead-lag-basert foroverkobling:

$$F_d(s) = k_f \frac{(\tau_{lead}s + 1)}{(\tau_{lag}s + 1)}. \quad (\text{Lead-lag fremoverkobling})$$

Antagelse: Vi trenger ikke å ta høyde for noen tidsforsinkelser.

Merk: Hvis $U_d(s) = F_d(s)D(s)$, så tilsvarende dette f.eks. følgende differensialligning:

$$\begin{aligned} \dot{z} &= -\frac{1}{\tau_{lag}}z + \left(\frac{\tau_{lag} - \tau_{lead}}{\tau_{lag}}\right)d \\ u_d &= \frac{k_f}{\tau_{lag}}z + \tau_{lead}\frac{k_f}{\tau_{lag}}d \end{aligned}$$

hvor z er en intern tilstand for dette lead-lag-elementet, som da fremhever at dette er et dynamisk element.

Fra [Seborg et al., 2016, §15.7] har vi følgende prosedyre for å finne parameterne $(k_f, \tau_{lead}, \tau_{lag})$:

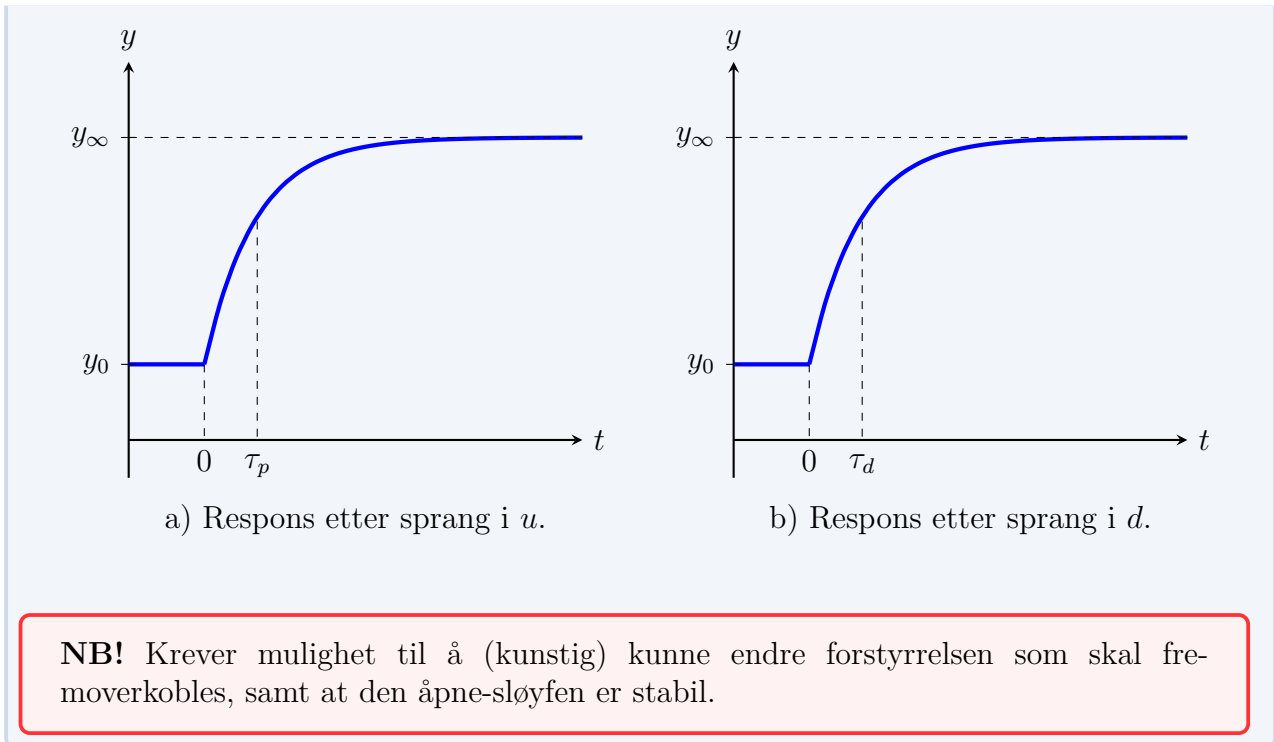
Eksperimentell tuning av lead-lag-basert foroverkobling:

Steg 0: Styr prosessen til ønsket arbeidspunkt, og skru av/begrens tilbakekoblingen om mulig.

Steg 1: Sett $\tau_{lead} = \tau_{lag} = 0$ og k_f til en liten verdi;

Steg 2: (*Innstilling av k_f .*) Sett en liten (kunstig) endring (3-5%) i forstyrrelsen som skal fremoverkobles. Hvis dette fører til et statisk avvik, øk k_f til avviket forsvinner;

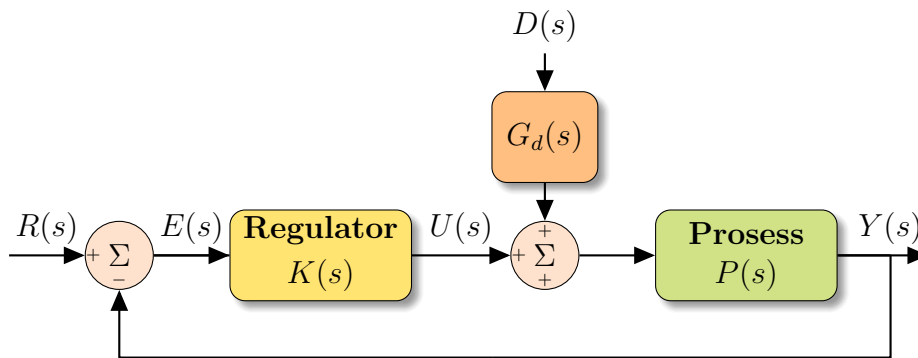
Steg 3: (*Innstilling av τ_{lead} og τ_{lag} .*) Sett $k_f = 0$. Innfør først et sprang i pådraget, u , og sett τ_{lead} lik tidskonstanten τ_p (se fig. a). Innfør så et sprang i forstyrrelsen, d , og sett τ_{lag} lik tidskonstanten τ_d (se fig b). For videre fininnstilling av τ_{lead} og τ_{lag} , se [Seborg et al., 2016].



Så hva er egentlig et lead-lag-element?

TODO

8.1.8 Regulator med to frihetsgrader



Figur 8.3: Enkel lukket sløyfe med forstyrrelse.

Gitt et systemet vist i figur 8.3. Vi har

$$Y(s) = \frac{K(s)P(s)}{1 + K(s)P(s)}R(s) + \frac{1}{1 + K(s)P(s)}G_d(s)P(s)D(s) = T(s)R(s) + S(s)G_d(s)P(s)D(s),$$

hvor $T(s)$ er *følgeforholdet* og $S(s)$ er *avviksforholdet*. Legg her merke til at

$$T(s) + S(s) = 1.$$

Dette gjør at man ofte møter på følgende problem når man kun har tilbakekobling:

Kompromiss: Med ren tilbakekobling, ved å optimalisere responsen mtp. sprang i referansen reduserer man ytelsen mtp. forstyrrelses-responsen, og vice versa.

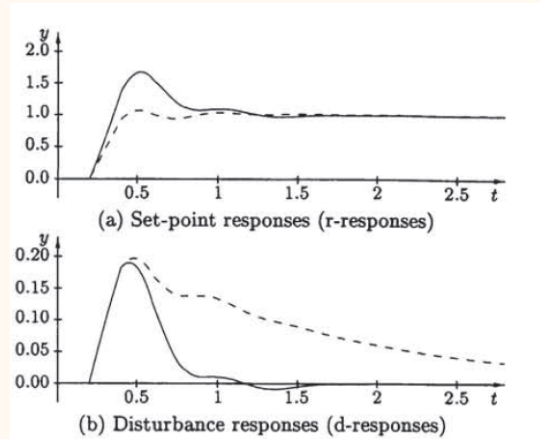
Eksempel 8.10. (Eksempel og figur fra [Taguchi and Araki, 2000]; se også fig. 7.3)

Gitt en FOPTF-prosess: $P(s) = e^{-0.2s}/(s + 1)$. For en PID-regulator på **parallellform** er følgende parametere optimale mhp. responsen etter et sprang i referansen (stiplede linjer):

$$k_p = 4.75, \quad T_I = 1.35, \text{ og } T_D = 0.094;$$

mens følgende er optimale mhp. forstyrrelses-responsen (hele linjer i fig.):

$$k_p = 6.0, \quad T_I = 0.4, \text{ og } T_D = 0.084$$

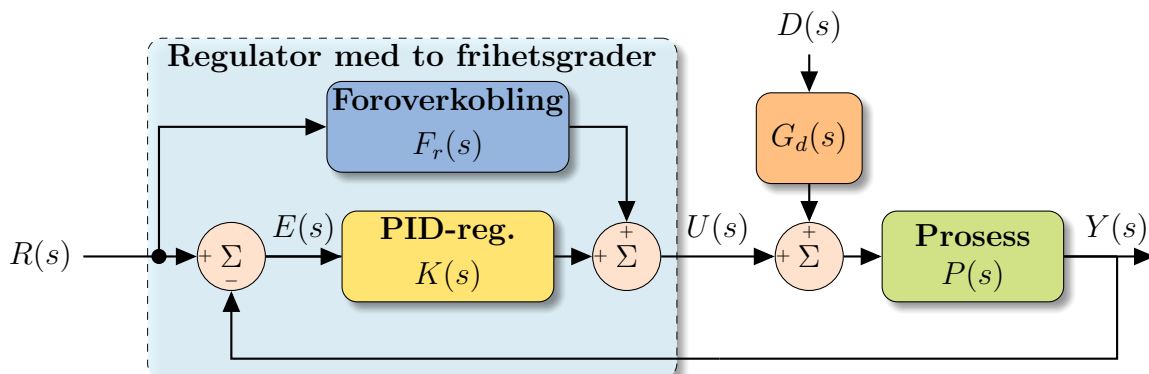


Hvis vi ikke kan legge til en foroverkobling fra forstyrrelsen, så er en annen mulig løsning på dette problemet har vi jo allerede sett på: vi kan legge på en foroverkobling fra referansen for å øke ytelsen, så kan vi tune regulatoren $K(s)$ slik at man får god forstyrrelses-respons. Men hva hvis vi ikke har en god modell av prosessen, hvordan finner vi da foroverkobling? Vi oppsummerer problemstilling;

Nåværende problem: Gitt et system som i figur 8.3. Vi ønsker både god referanse- og forstyrrelses-respons. Men

- vi kan ikke måle forstyrrelsen(e);
- vi har muligens ingen god modell av prosessen.

Mulig løsning: vi kan «tune» er foroverkobling fra referansen (tilsvarende det vi gjorde med lead-lag-enheten for en foroverkobling av forstyrrelsen; se § 8.1.7).



Figur 8.4: Illustrasjon av regulator med to-frihetsgrader.

En slik regulator, som er illustrert i figur 8.4, sies ofte å ha to frihetsgrader²:

²I [Seborg et al., 2016] brukes en relatert, men litt annerledes definisjon av en regulator med to frihetsgrader.

- Tilbakekobling for forstyrrelses-fjerning/robushet;
- Foroverkobling for ytelse.

Innstilling av regulator med to frihetsgrader (RTFG): Vi antar at vi har en PID-regulator på [parallellform](#),

$$K(s) = k_P \left(1 + \frac{1}{T_I s} + T_D s \right),$$

samt at foroverkoblingen fra referansen har følgende form (evt. et lead-lag-element):

$$F_r(s) = \alpha + \frac{\beta s}{\gamma s + 1}. \quad (\text{Foroverkobling i RTFG})$$

Prosedyre:

- Velg $K(s)$ (altså k_P , T_I og T_D) for å få ønsket forstyrrelses-respons (ideelt sett tunet vha. kunstige sprang i forstyrrelsen);
- Velg $F_r(s)$ (altså α og β , og evt. γ) for ønsket respons fra sprang i referansen.

8.2. Referanse-glatting og -følging

▶ mGPIhy57UTg&t=5861

Nåværende problem: Tenk deg følgende typiske scenario, hvor man ønsker å endre fra en referanseverdi (settpunkt), r_0 , til en annen ny, konstant referanse, r_1 . Dette fører naturlig nok til et sprang i referansen, som kan resultere i flere uønskede fenomener:

1. Spranget i referansen fører til et sprang i avviket, som igjen fører til et sprang i pådraget;
2. Regulatoren vil ikke kunne henge med på spranget (ville krevd uendelig båndbredde);
3. En for aggressiv regulator vil kunne føre til at (§ 3.3.1)
 - vi når pådragsorganenes/aktuatorenes ratebegrensninger (hvor fort de kan endre seg);
 - aktuatorene går i metning, slik at fenomenet wind-up potensielt kan oppstå.

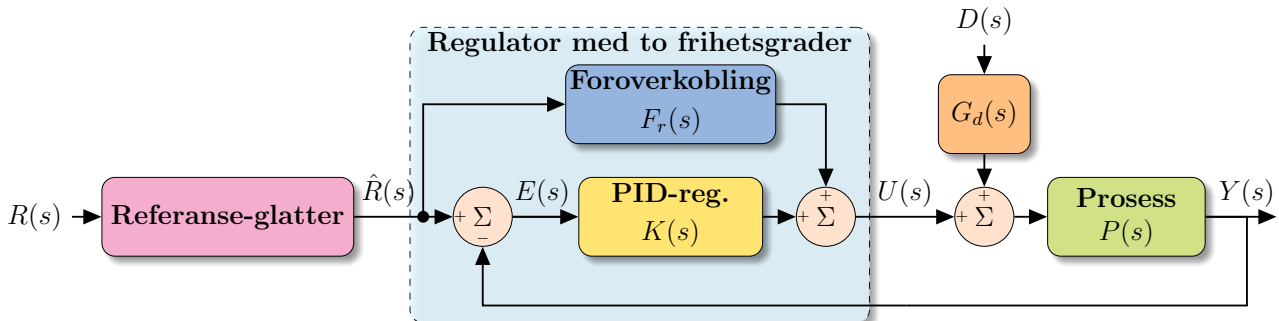
En mulig strategi for å unngå disse problemene er **referanse-glatting**:

Referanse-glatting: I stedet for å endre fra et konstant settpunkt, r_0 , til et annet, si r_1 , i et sprang (hopp), så generer man et kontinuerlig og tidsvarierende referansesignal som kobler sammen r_0 og r_1 ; altså, man glatter ut det diskrete hoppet til en kontinuerlig funksjon som regulatoren klarer å følge (**referanse-følging**).^a

^aDenne strategien egner seg også til såkalt **myk-starting**, hvor man prøver å gradvis øke pådraget for å

oppnå en ønsket verdi, fremover å gjøre dette som hopp i en referanse.

En illustrasjon av denne strategien er vist i figur 8.5, hvor et referanse-glatte modifierer (den potensielt diskontinuerlige) referansen inn, r , til en kontinuerlig referanse, $\hat{r}(t)$.



Figur 8.5: Illustrasjon av referanse-glatte før regulator med to frihetsgrader.

Alternativer til referanse-glatte: Vi skal se på to typer referanse-glattings-strategier:

- **Lavpassfiltere:** Sende r gjennom et lavpassfilter av ønsket orden;

Fordeler: Relativt enkle å bruke, ikke veldig numerisk kostbare å implementere, samt lite sensitive til referanse-signalet;

Ulemper: Vanskelig å forme responsen akkurat som man ønsker, kan ikke (lett) brukes med analytisk-foroverkobling, samt kun asymptotisk konvergering til ny referanse (dette er dog ikke et stort problem i praksis).

- **Interpolasjon:** [interpolere](#) mellom den gamle- og nye referansen.

Fordeler: Stor frihet i å kunne forme referansen som ønsket, kan brukes i kombinasjon med analytisk foroverkobling;

Ulemper: Kan være numerisk kostbare og krevende å implementere.

Noen eksempler på slike referanse-glattingsmetoder er illustrert i figur 8.6, hvor man har et sprang ved tiden t_0 fra r_0 til r_1 . De forskjellige metodene er som følger:

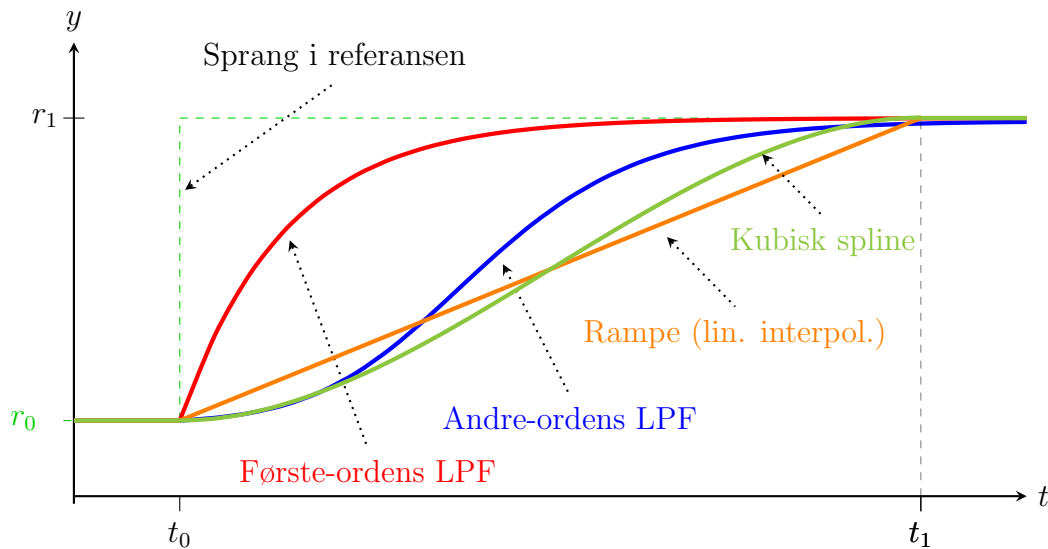
Første-ordens lavpassfilter (LPF): For en ønsket $a > 0$ (større a = raskere respons), ta

$$\frac{d\hat{r}}{dt} = a(r_1 - \hat{r}), \quad \hat{r}(t_0) = r_0.$$

Siden dette fører til en knekk ved tiden t_0 , egner denne seg best sammen med P(I)-regulator og statisk foroverkobling, selv om både en PID med stort derivat-filter og lead-lag-foroverkobling selvsagt også kan brukes.

Andre-ordens LPF: For $a, b > 0$, ta $\hat{r}(t) = x_1(t)$, hvor

$$\begin{aligned} \dot{x}_1 &= x_2, & x_1(t_0) &= r_0, & x_2(0) &= 0, \\ \dot{x}_2 &= a(r_1 - x_1) - bx_2. \end{aligned}$$



Figur 8.6: Illustrasjon av forskjellige referanse-glattings-strategier for et sprang i referansen.

Her er $\frac{d\hat{r}}{dt}$ kontinuerlig, så denne metoden er egnet i kombinasjon med PID-regulatorer. Merk at man kan få underdempet respons med oversving ved dårlig valg av a og b , slik at $b = 2\sqrt{a}$ (kritisk demping) anbefales.

Rampe/ lineær interpolasjon: Lineære interpolering mellom r_0 og r_1 :

$$\hat{r}(t) = (r_1 - r_0) \frac{(t - t_0)}{(t_1 - t_0)} + r_0 \text{ for } t \in [t_0, t_1]$$

$$\hat{r}(t) = r_1 \text{ for } t \geq t_1.$$

Ulempe at $\frac{d\hat{r}}{dt}$ ikke er veldefinert ved t_0 og t_1 , men metoden kan uansett brukes med analytisk foroverkobling tross nevnte sprang i $\frac{d\hat{r}}{dt}$ siden den er stykkvis konstant ellers.

Kubisk-spline-interpolasjon: Ved å definere $\hat{t}(t) = \frac{(t-t_0)}{(t_1-t_0)}$, så har man

$$\hat{r}(t) = a\hat{t}^3 + b\hat{t}^2 + c\hat{t} + d \text{ for } \hat{t} \in [0, 1] \text{ (samme som } t \in [t_0, t_1]),$$

$$\hat{r}(t) = r_1 \text{ for } t \geq t_1.$$

Man må her velge (a, b, c, d) slik at $\hat{r}(t_0) = r_0$ og $\hat{r}(t_1) = r_1$, samt tar man som regel $\frac{d\hat{r}}{dt}(t_0) = \frac{d\hat{r}}{dt}(t_1) = 0$. Dette krever som oftest at man må løse en matriseligning på formen $\mathbf{Ax} = \mathbf{b}$ mhp. \mathbf{x} for å finne koeffisientene (a, b, c, d) .

TODO

9. Anti-windup og rykkfri overføring

9.1. Anti-windup for PID-regulatorer

Alternative kilder: Kap. 8.2.2 i [Seborg et al., 2016]; Wikipedia; §12.5 i [Balchen et al., 2016]; [Tarbouriech and Turner, 2009].

9.1.1 Hva er windup?

▶ eF2uHaFNHms&t=17

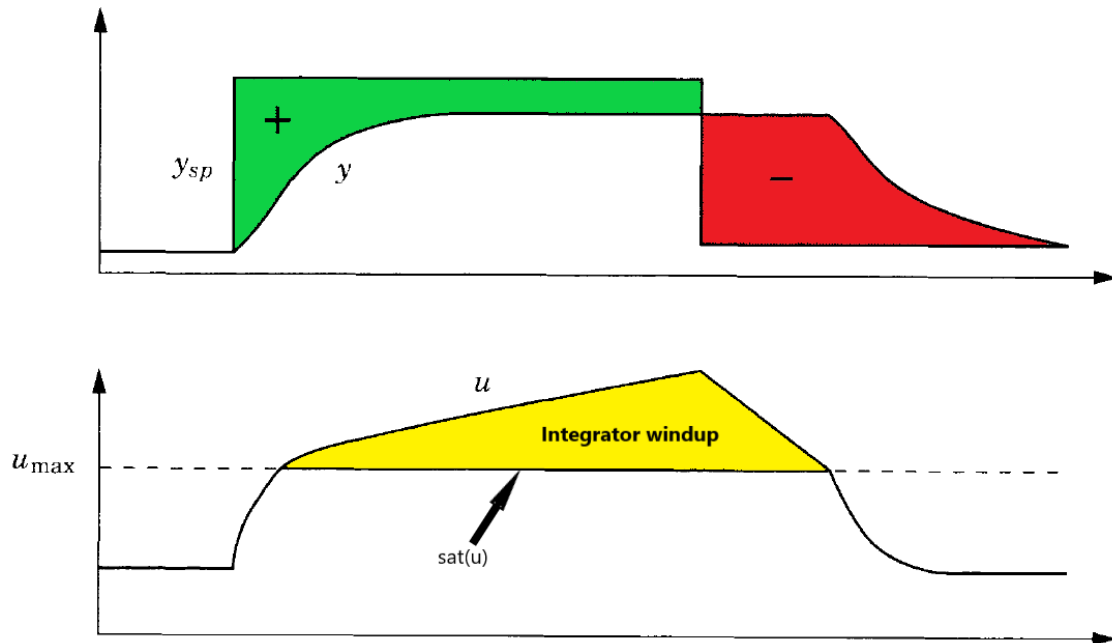
“Wind-up” (evt. overlading) er et fenomen som kan oppstå gitt følgende «ingredienser»:

1. En regulator med et integral-ledd (eller mer generelt, en dynamisk regulator av noe slag);
2. Pådragsorgan (aktuator) med ratebegrensninger og/eller som kan gå i metning.

Integrator-leddet i en regulator kan man tenke på som en trek-opp-bil som kan trekkes/lades opp eller ned (opptvinning=windup). Når man da bruker en aktuator med metnings- eller ratebegrensninger (en ventil kan f.eks. bare være et sted mellom helt åpen og helt lukket), så vil integrator-leddet kunne lades opp (wind up) selv om man har gått i metning. I-leddet vil da fortsette å øke i magnituden uten at det endrer aktuatoren og vil dermed ikke ha noen umiddelbar effekt på systemet. Når man så treffer ønsket settpunkt (hvis man er så heldig) eller det endres, da må det “overflødig” integral-leddet igjen “tømmes ut”, noe som kan ta tid. Vi kaller dette fenomenet for integrator windup. **Mulige konsekvenser av windup inkluderer:**

- overskyting av settpunkt,
- treg respons (forsinkelser og lange transienter),
- kanskje til og med ustabilitet.

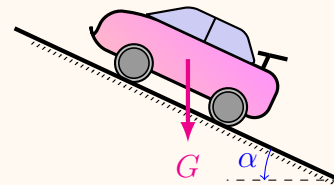
En illustrasjon av dette fenomenet er vist i figur 9.1. Der ser man at systemet går i metning før man klarer å nå ønsket settpunkt, y_{sp} . Integral-leddet forsetter dog uansett å øke (se gul region) på grunn av det vedvarende avviket (se grønn region). Når man så endrer endrer settpunktet, så tar det lang tid før regulatoren reagerer (se den røde regionen).



Figur 9.1: Illustrasjon av fenomenet windup; figur fra [Åström and Hägglund, 2006].

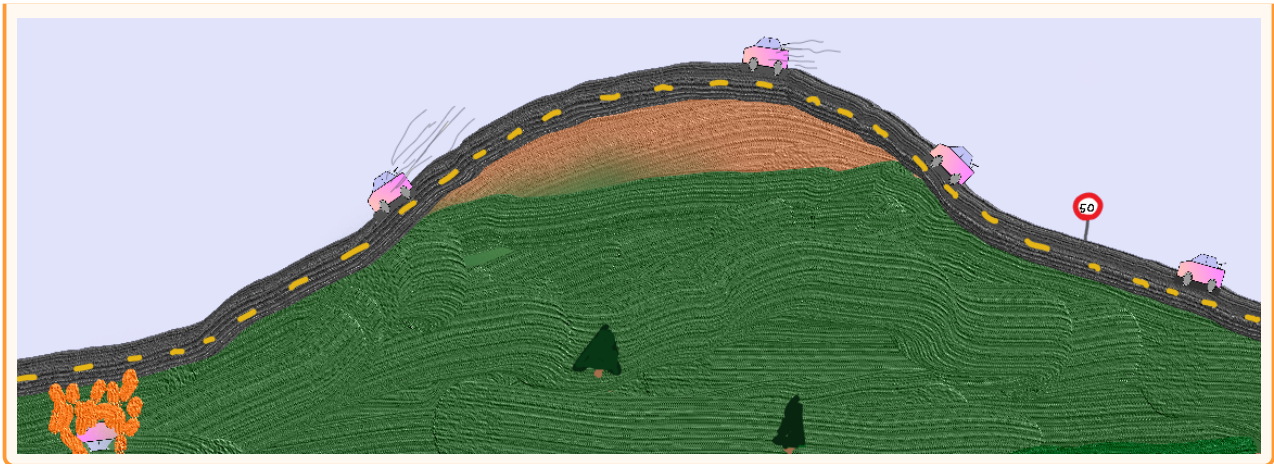
Eksempel 9.1. (Tafatt bil med cruise-kontroll som skal over en ås)

En meget snasen bil skal kjøre over en ås. Bilen har implementert cruise-kontroll ved hjelp av en PI-regulator. Dessverre står ikke bilens motor i stil med dens lekre design, noe som gjør at den ikke klarer å opprettholde den ønskede hastigheten på 50 km/t i de bratteste delene av bakken (se også oppgave 8.1 for hva som kreves av dreiemoment på hjulene her).



Konsekvensene av dette er som følger:

1. Bilens motor klarer ikke å gi nødvendig dreiemoment—den går i metning;
2. Et (stasjonært) avvik oppstår;
3. Avviket vil få regulatorens integral-ledd til å øke tross i at metningsgrensen er nådd;
4. I-leddet vil fortsette å øke til bakken flater ut og man da endelig når settpunktet;
5. Man vil måtte ligge en stund over settpunktet for at I-leddet skal “lades ut”;
6. Kraftig oversving er mulig, spesielt hvis nedoverbakken begynner før full nedlading;
7. Dette kan potensielt ha katastrofale følger.



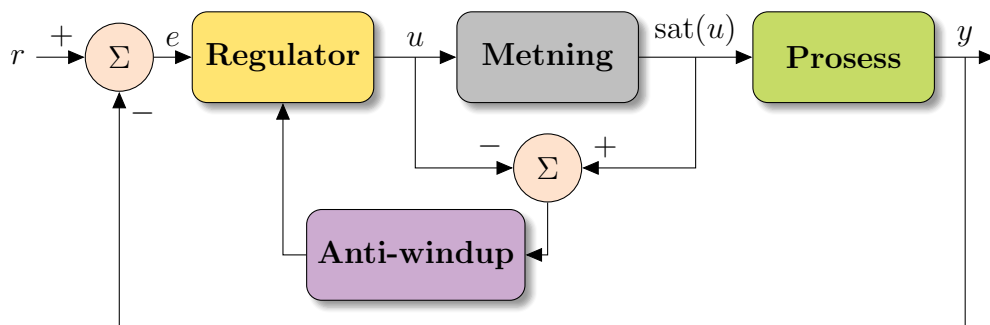
MATLAB/Simulink: Det finnes et ferdig eksempel som illustrerer dette fenomenet. Du finner det [her](#) eller ved å skrive følgende kommandovinduet i MATLAB:
`openExample('simulink_industrial/AntiWindupControlUsingAPIDControllerExample')`

9.1.2 Anti-windup-metoder ▶ eF2uHaFNHms&t=938

Metning er “saturation” på engelsk. Vil bruker derfor `sat(·)` til å betegne metningsfunksjonen:

$$\text{sat}_{\underline{u}}^{\bar{u}}(u) = \min(\bar{u}, \max(u, \underline{u})) = \begin{cases} \bar{u} & \text{når } u \geq \bar{u}, \\ u & \text{når } \underline{u} \leq u \leq \bar{u}, \\ \underline{u} & \text{når } u \leq \underline{u}. \end{cases} \quad (\text{Metningsfunksjonen})$$

For enkelthetskyld vil vi som regel bare skrive `sat(u)`, altså droppe metningsverdiene \bar{u} og \underline{u} .
 En god måte å unngå wind-up som oppstår pga. sprang i referansen, er å implementere referanse-glatting og -følging (se § 8.2). Er dette gjort på en god måte, så vil pådraget aldri gå i metning eller nå sine ratebegrensninger. På den annen side vil ikke dette ha noen effekt på wind-up som oppstår pga. av en forstyrrelse eller et stasjonært avvik vi ikke kan gjøre noe med.



Figur 9.2: Vanlig arkitektur for anti-windup-metoder.

For systemer med metning¹ som er utsatt for store forstyrrelser, og som bruker integralvirkning i regulatoren, bør man derfor bruke en **anti-windup metode**/-kompensator. Det finnes

¹Som nevnt kan wind-up også oppstå pga. ratebegrensninger, men dette er et mer sjeldent problem enn problemet fra metninger, samt mer krevende å håndtere, så vi skal ikke se på hvordan man håndterer det.

to hovedkategorier av slike metoder:

- **En-stegsmetodene:** pådragsmetning blir tatt hensyn til direkte i regulatorsyntesen (når man designer regulatoren); og
- **To-steg-metoder** (også kalt anti-windup-kompensering): man designer først regulatoren uavhengig av metning, og så velger man en strategi for kompensere for/forhindre windup.

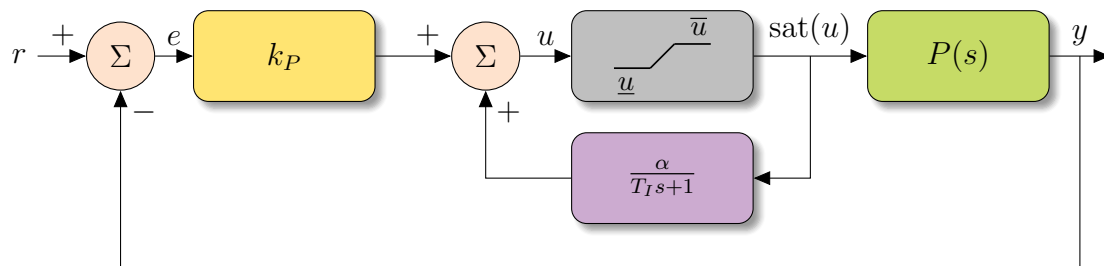
En typisk anti-windup-struktur er vist i figur 9.2, hvor man ser på differansen til pådraget før og etter en metnings-blokk. Hvis differansen ikke er null, så påvirker man regulatoren på en eller annen måte.

Merk: som oftest måler man ikke direkte om pådragsorganet har gått i metning eller ikke, slik at den grå-blokken i figur 9.2 som oftest er en «fiktiv» metning implementert i koden til regulatoren man har laget.

Integral-leddet som tilbakekobling



En enkel anti-windup strategi for PI-regulatorer er vist i figur 9.3. Man kan vise (se oppgaven under) at når pådraget ikke er i metning, så tilsvarer dette (for $\alpha = 1$) en standard PI-regulator.



Figur 9.3: PI-regulator med anti-windup, hvor integral-delen er implementert som en tilbakekobling over metnings-blokken.

Hvis derimot regulatoren går i metning, slik at $\text{sat}(u) = u_{\text{met}}$ får en konstant verdi, så vil også (så lenge pådraget forblir i metning) utgangen av den lille blokken gå mot en konstant verdi lik αu_{met} (det er også en oppgave hvor du skal vise dette). Noen fordeler og ulemper med denne metoden inkluderer:

Fordeler: Velding enkel. **Ulemper:** Regulator-spesifikk; påvirker også P-leddet².

Oppgave 9.1. Vis at regulatoren i figur 9.3 tilsvarer en vanlig PI-regulator hvis man ikke har metning (altså $u = \text{sat}(u)$ er alltid sant).

²Ulempen med at P-leddet også blir påvirket av metoden i figur 9.3 kan fikses ved flytte den grå metnings-blokken ned i avgrensingen før den lille blokken. Dette gjør det også mulig å bruke alternativer (bl.a. såkalt “batch unit”) til en ren metningsfunksjon; se [Åström and Hägglund, 2006] for ytterligere detaljer rundt dette.

Oppgave 9.2. Vi at hvis systemet i figur 9.3 går i metning, slik at $\text{sat}(u) = u_{met} = \text{konstant}$ får en konstant verdi, så vil også (så lenge pådraget forblir i metning) utgangen av den lille blokken gå mot en konstant verdi lik αu_{met} .

Hint: sluttverdi-teoremet.

Oppgave 9.3. Vis hvordan du kan legge til derivat-virking til PI-regulatoren i figur 9.3 slik at du får en PID-regulator, både på serie- og parallellform.

Begrense absoluttverdien av I-leddet ▶ eF2uHaFNHms&t=1292

En annen enkel anti-windup-metode er å begrense absoluttverdien integralet-leddet kan ha. Altså at man stopper integrasjonen hvis absoluttverdien er over et spesifisert nivå, gitt at integrator-inngangen ikke har motsatt fortegn.

Hvis tilsvarende ideelle integralledd med $k_I > 0$ har formen

$$u_I = k_I \int_0^t e(\tau) dt \quad \iff \quad \dot{u}_I = k_I e(t),$$

så tilsvarende denne metoden følgende logikk:

$$\dot{u}_I = \begin{cases} 0 & \text{når } u_I \geq \bar{u}_I \text{ og } e(t) \geq 0, \\ k_I e(t) & \text{når } u_I \geq \bar{u}_I \text{ og } e(t) < 0, \\ k_I e(t) & \text{når } \underline{u}_I \leq u_I \leq \bar{u}_I, \\ k_I e(t) & \text{når } u_I \leq \underline{u}_I \text{ og } e(t) > 0, \\ 0 & \text{når } u_I \leq \underline{u}_I \text{ og } e(t) \leq 0. \end{cases}$$

Fordeler: Velding enkel. **Ulemper:** Veldig konservativ.

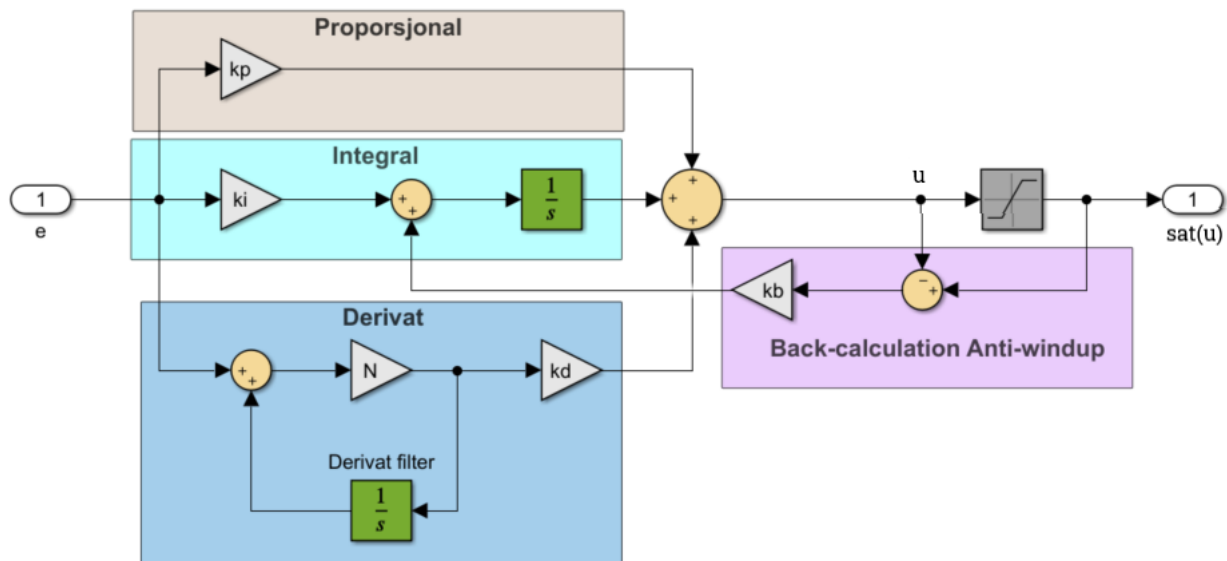
Clamping ▶ eF2uHaFNHms&t=1470

Clamping-strategien er lik den forrige, men i stedet for å kun se på I-leddet baserer denne metoden seg på å detektere at systemet har gått i metning: hvis $\text{sat}(u) - u \neq 0$, så “skru av” integratoren i integral-leddet, gitt at inngangen og utgangen til integrator-delen har samme fortegn. Med andre ord, hvis pådragsorganet har gått i metning så får ikke I-leddet endre verdi hvis det forverrer problemet. Integrasjonen gjenopptas hvis regulatorutgangen er mindre enn metningsgrensen, eller hvis utgangen av I-leddet har motsatt fortegn i forhold til dets inngang.

Fordeler: Umiddelbar effekt. **Ulemper:** Krever implementasjon i kode, med potensiale for “bugs” for mer komplekse reguleringsstrategier; kan gi dårlig transient-respons for systemer med store tidsforsinkelser.

Back calculation ▶ eF2uHaFNHms&t=1671

En Simuink-implementasjon av back-calculation anti-windup er vist i figur 9.4. Back calculation tilsvarende derfor at den lille anti-windup-blokken i figur 9.2 er lik $\frac{-k_b}{s}(U(s) - U_n(s))$. I-leddet



Figur 9.4: Simulink-diagram av Back-calculation anti-windup.

vil derfor prøve begynne å “bevege seg “motsatt vei” når pådraget har gått i metning, med hastighet gitt av størrelsen på k_b (større = raskere). Man skulle kanskje tro at man burde ta k_b så stor som mulig, men har man et derivat-ledd skal man være forsiktig med å ta k_b for stor. Grunnen er at derivat-leddet kan føre til raske, kort-levde økninger i pådraget som fører det i metning, noe som da kan “nulle-ut” I-leddet.

Tommelfingerregel (fra [Åström and Hägglund, 2006]): Ta T_b større enn $T_D = k_d/k_p$ og mindre enn $T_I = k_p/k_I$, f.eks. $T_b = \sqrt{T_I T_D}$.

Fordeler: Enkel og effektiv, egner seg godt for systemer med store tidsforsinkelser. **Ulemper:** Ikke en umiddelbar effekt; ikke alltid lett å “tune” k_b .

9.2. Tracking

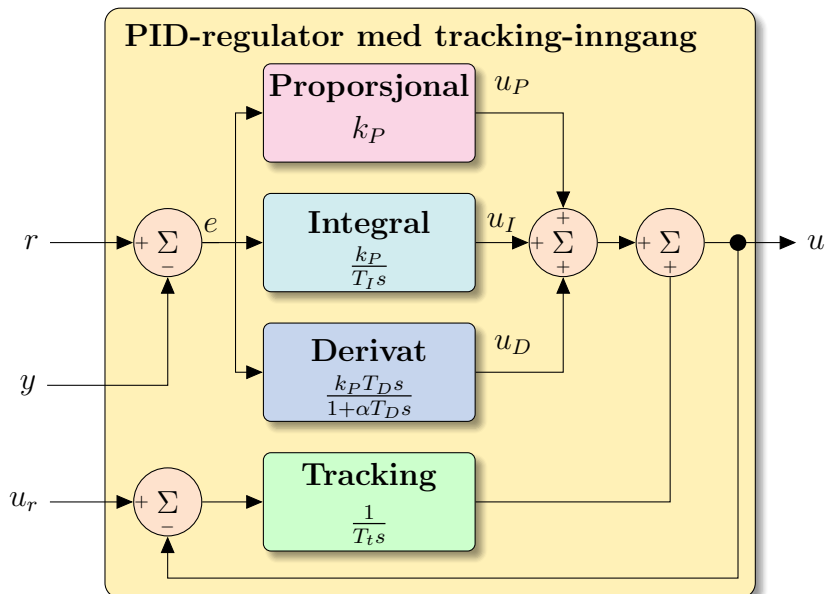


En regulator med back-calculation kan sies å ha to moduser:

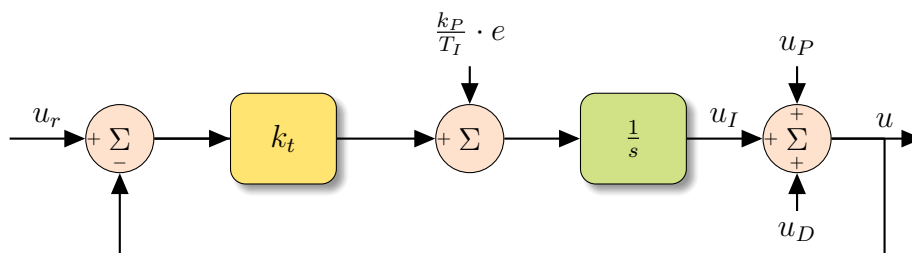
- normal-modus (når systemet ikke er i metning); og
- **tracking**-modus (når regulatoren prøver å følge (“tracke”) utgangen av metnings-blokken).

Som vist i figur 9.5 kan vi generalisere denne ideene ved å se på differansen mellom utgangen til regulatoren, u , og et annet «referansesignal» u_r , og så «mate» dette inn i integral-leddet. Dette kalles for **tracking**. Som vist i figur 9.6 kan tracking derfor ses på som en slags tilbakekoblingsløype med P-regulator for regulator-utgangen.

Gitt en PID-regulator med tracking-inngang, kan f.eks. back-calculation implementeres ved å ta $u_r = \text{sat}(u)$. Dette er jo for så vidt ikke noe nytt i seg selv, men, som vist i figur 9.7,

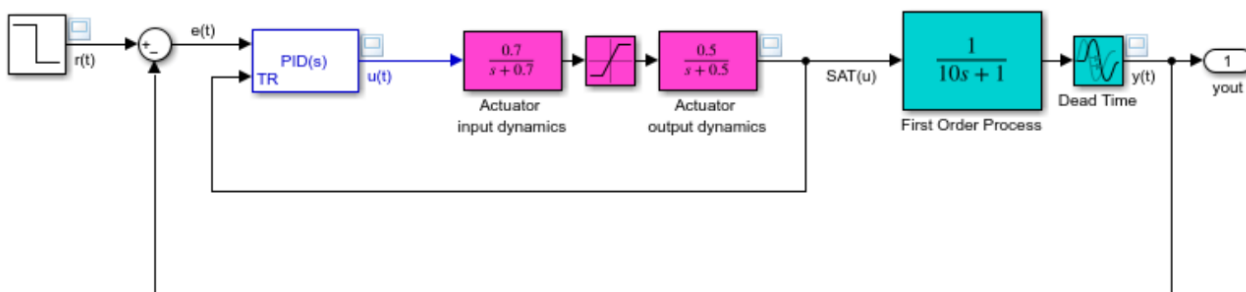


Figur 9.5: PID-regulator med tracking-inngang.



Figur 9.6: Tracking illustrert som en tilbakekoblingsløyfe for regulatorutgangen ($k_t = \frac{1}{T_t}$).

tillater tracking-inngangen at man også kan ta høyde for mer kompleks aktuatordynamikk. Du kan se mer på dette MATLAB-eksempelet via: <https://se.mathworks.com/help/simulink/sref/anti-windup-control-using-a-pid-controller.html>. Der kan du også se hvordan man kan ta høyde for eventuelle foroverkoblinger.



Figur 9.7: Bruk av tracking-inngangen til PID-regulatoren i Simulink for å ta høyde for komplisert aktuator dynamikk i tillegg til metning; skjermbilde fra [lenke](#).

Tracking i Simulink: I PID-blokken i Simulink har du mulighet til å skru på «Tracking mode». Som vist i figur 9.7 gir dette PID-blokken en ekstra inngang. Når tracking er aktivt, føres forskjellen mellom tracking-innganen og regulatorutgangen tilbake til integral-delen (se fig. 9.5).

Merk at tracking-innganen også kan brukes til andre formål enn bare anti-windup, deriblant rykkfri overføring som vi skal se på i neste avsnitt.

9.3. Rykkfri overføring for PID-regulatorer

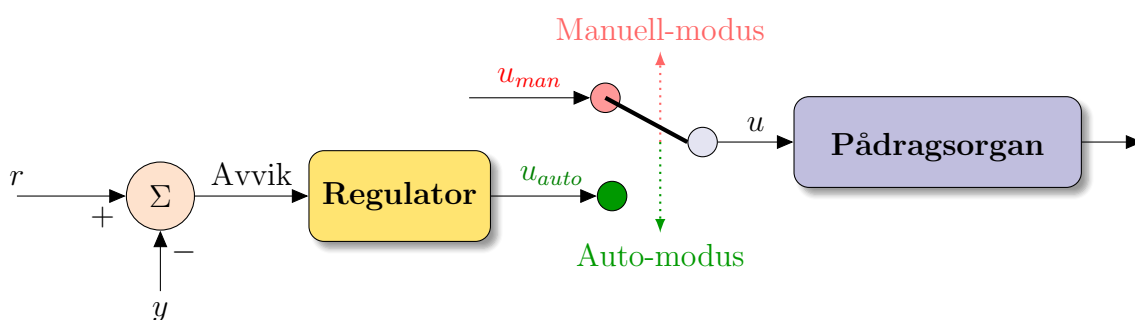
▶ eF2uHaFNHms&t=2398

Hvorfor skal du lære dette? Ved et skifte mellom forskjellige reguleringsmodi eller bytte til forskjellige av parametre, er det sjeldent ønskelig at man forstyrrer prosessen mer enn nødvendig eller at man får unødvendige sprang i pådraget—man ønsker en *rykkfri* overføring.

Et typisk scenario hvor rykkfri overføring er relevant er et bytte mellom manuell- og auto-modus, som illustrert i figure 9.8. For eksempel, si du har regulatoren i auto-modus og det ved et tidspunkt gir ut et pådrag på 20%. Så endrer man plutselig til manuell-modus med et pådrag på 50%. Dette fører jo til et kraftig sprang i pådraget, noe som sjeldent er ønskelig i praksis.

Rykkfri overføring mellom manuell og auto-modus: Hvis man selv får lov å velge verdiene ved et bytte, kan følgende trivielle strategier være et alternativ (vær forsiktig og tenk grundig igjennom denne strategien hvis du f.eks. har anti-windup eller lignende implementert):

- **Fra manuell til auto:** Initialiser integral-leddet (evt. et nominelt pådrag) slik at $u_{auto} = u_{man}$ ved bytte;
- **Fra auto til manuell:** Sett $u_{man} = u_{auto}$.



Figur 9.8: Man ønsker med rykkfri overføring å unngå sprang i pådraget når man f.eks. bytter mellom manuell- og auto-modus.

Generelle strategier: Det er mange mer generelle måter å løse disse utfordringene på, men de fleste baserer seg på bruk av en integrator på en eller annen måte. Grunnen til dette er at vi ønsker å gå fra en gammel verdi til en annen, ny verdi uten å få et diskontinuerlig hopp

(i den digitale verden er jo ikke hopp til å unngå, så da ønsker vi bare at hoppene skal være små), og en integrator har den fine egenskapen at den glatter ut selv diskontinuerlige signaler.³

Bruk av tracking

I § 9.1.2 så vi på konseptet **tracking**. Der ble det også så vidt nevnt at tracking kan brukes til å implementere rykkfri overføring mellom to regulatorer, eller fra manuell- til auto-modus. Merk dog at den ikke tar hensyn til et bytte fra auto til manuell.

Et Simulink eksempel er tilgjengelig via [denne lenken](#).

³Metodene vi brukte for referanseglatting i § 8.2 kan for så vidt brukes her, men det kan fort bli litt unødvendig tregt og/eller komplekst å gjøre det på den måten, selv om det er mange likhetstrekk med både tracking- og ratebegrensnings-metodene.

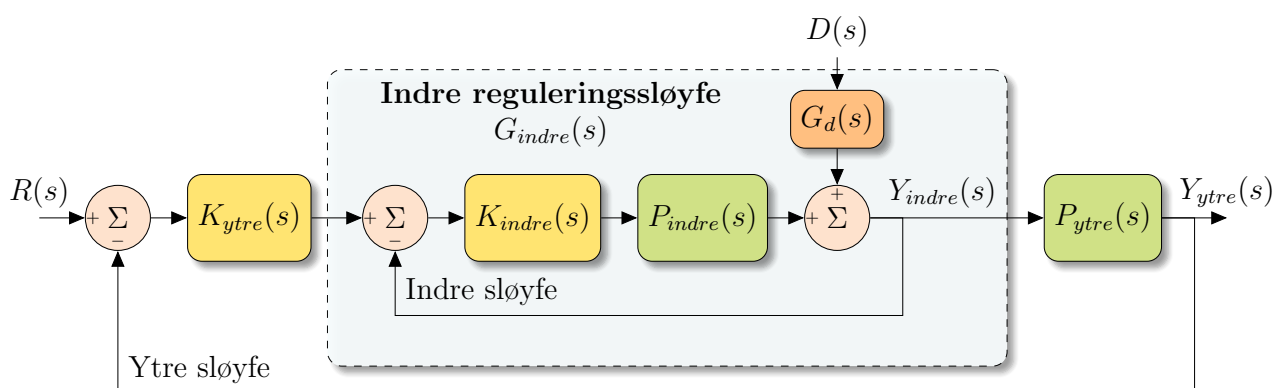
10. Alternative reguleringsstrukturer

I dette kapitlet skal vi se på viktig reguleringsstrukturer som kaskade- og forholdsregulering, samt se på hvordan man kan lage regulatorer for ulinære systemer ved hjelp av parameterstyring.

10.1. Kaskade-regulering

ngkRAsYEatK

Alternative kilder: §16.1 i [Seborg et al., 2016]; §9.6.6 i [Bjørvik and Hveem, 2014]; §10.5.2 i [Skogestad and Postlethwaite, 2007].



Figur 10.1: Illustrasjon av en kaskaderegulerings-sløyfe, bestående av en indre reguleringsløyfe, $G_{indre}(s)$, som inngår i den ytre sløyfen.

En ulempe med standard tilbakekoblinger (f.eks. en PID-regulator) er at regulatoren først reagerer på en forstyrrelse når den har endret avviket. Vi har dog allerede sett på en metode som lot regulatoren reagere mye raskere på forstyrrelser, nemlig foroverkobling (se § 8.1). Men foroverkoblinger krever i utgangspunktet at vi kan måle forstyrrelsen vi ønsker å foroverkoble.

Hvis vi derimot ikke har en slik måling, men i stedet har en ekstra måling relatert til prosessvariabelen som ligger «nærmere» eventuelle forstyrrelser som virker på systemet, så kan vi potensielt få en mye bedre forstyrrelses-respons *vhva. kaskade-regulering*.

Kaskaderegulering betyr at man har to (eller flere) regulatorer i serie (kaskade), hvor den første (ytre) regulatoren setter referansen til den andre (indre) regulatoren; se figur 10.1.

Målet er at den lukkede, indre sløyfen, som sett fra den ytre, kan regnes som en triviell enhetsforsterkning, altså $G_{indre}(s) \approx 1$ (se figur 10.1). Dermed kan man ignorere både den indre sløyfen og (!) forstyrrelsen $d(t)$ når man designer den ytre regulatoren.

Dette krever dog at 1) den indre reguleringsløyfen er betydelige raskere enn den ytre, og 2) at man har en tilsvarende rask måling av den indre prosessvariabelen, $y_{indre}(t)$, i tillegg til en (tregere) måling av den ytre variabelen, $y_{ytre}(t)$.

Innstillings-prosedyre: Still først inn den indre sløyfen med den ytre skrudd av, så still inn den ytre sløyfen med den indre på.

⚠ Pass på! $G_{indre}(s) \approx 1$ er ikke alltid mulig i praksis. Man bør derfor også ta høyde for $G_{indre}(s)$ når man designer den ytre sløyfen (se f.eks. eksempel 10.2).

Kaskade kan derfor (og kanskje også bør) brukes når man bare har et pådragsorgan, men flere målinger relatert til det man ønsker å styre.

Det er typisk at den indre sløyfen tilsvarende styring av aktuatordynamikken (pådragsorganet). Noen vanlige eksempler følger: posisjonen til en reguleringsventil er den indre, mens væskestrømmen gjennom ventilen er den ytre; væskestrømmen gjennom en reguleringsventil er den indre, mens væskehøyden i en tank er den ytre; strømsløyfen til en elektromotor er den indre, mens rotorhastigheten er den ytre; rotorhastigheten til en elektromotor er den indre, mens rotor posisjonen (etter et gir) er den ytre. Legg her merke til at ved å kombinere de to første eller to siste eksemplene, så får vi et reguleringsystemet bestående av tre tilbakekoblinger i kaskade.

Fordeler:

- Potensielt høyere båndbredde;
- Bedre stabilitetsmargin;
- Forbedret reduksjon av forstyrrelsen $d(t)$ (spesielt ved matchede-/last-forstyrrelser);
- Bedre håndtering av ulineariteter;
- Generelt bedre robusthet mtp. usikkerhet og forstyrrelser.

Ulemper: Trenger en ekstra måling (og dermed sensor).

Eksempler

Relevant MATLAB-eksempel: Skriv `openExample('control/cascadepiddemo')` i kommandovinduet eller bruk følgende lenke:

<https://se.mathworks.com/help/control/ug/designing-cascade-control-system-with-pi-controllers.html>.

Eksempel 10.1. (EL-motor) TODO

Eksempel 10.2. (SIMC i kaskade, eks. 10.12 fra [Skogestad and Postlethwaite, 2007].)
Gitt følgende system tilsvarende strukturen i figur 10.1 (i=indre, y=ytre):

$$P_i(s) = \frac{1}{(6s+1)(0.4s+1)}, \quad P_y(s) = \frac{(-0.6s+1)}{(6s+1)}e^{-1s}.$$

1. Uten kaskade: Anta at vi bare kan måle y_y . Vi ønsker å bruke SIMC-reglene til å stille inn en PI-regulator. Vi starter derfor med å finne en effektive tidsforsinkelsen vha. Skogestads halv-regel:

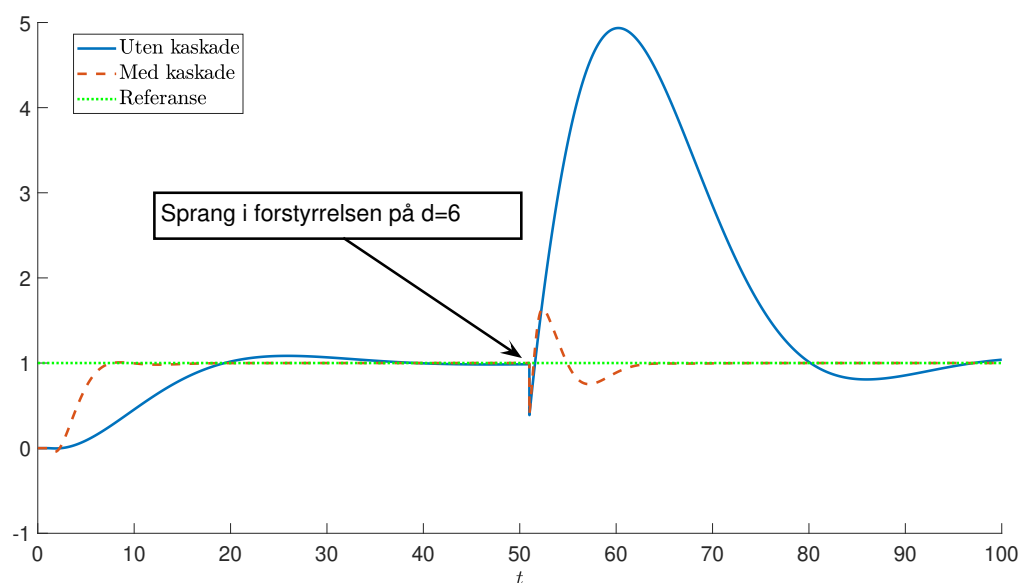
$$\hat{\theta} = 1 + 0.6 + 0.4 + 3 = 5.$$

Vi har dermed en tilpasset FOPTF-modell: $\hat{P}(s) = \frac{1}{9s+1}e^{-5s}$. Ved å ta $\tau_c = \hat{\theta}$, får vi følgende fra SIMC-regler (FOPTF): $k_P = \frac{\tau}{k(\hat{\theta} + \tau_c)} = 9/10 = 0.9$ og $T_I = \min(\tau_1, 4(\tau_c + \hat{\theta})) = 9$.

Kaskade: Indre sløyfen: Vi har nå fra halv-regelen at $\hat{P}_i(s) = \frac{1}{6.2s+1}e^{-0.2s}$. Ved å ta $\tau_{ci} = 2 \cdot 0.2 = 0.4$ får vi $k_{py} = 10.33$ og $T_{Ii} = 4 \cdot 0.6 = 2.4$. Siden SIMC er en intern-modellkontroll-metode (se § 7.2) så gir dette da en indre lukket-sløyfe tilsvarende $\frac{1}{0.4s+1}e^{-0.2s}$.

Ytre sløyfen: Vi approksimerer den indre lukkede sløyfen som en ren tidsforsinkelse med $\hat{\theta}_i = 0.2 + 0.4 = 0.6$. ved å ta i bruk halv-regelen igjen på den ytre sløyfen får vi at $\hat{\theta}_y = 1 + 0.6 + 0.6 = 2.2$, slik at $\hat{P}_y(s) = \frac{1}{6s+1}e^{-2.2s}$. Ved å ta $\tau_{cy} = \hat{\theta}_y$, får vi fra SIMC-reglene at $k_{py} = 1.36$ og $T_{Iy} = 6$.

Simulering: Plottet viser sammenligningen av regulatorene med og uten kaskade, hvor et sprang i forstyrrelsen ($d = 6$) inntreffer ved tiden $t = 50$ s. Det er tydelig at forstyrrelsesresponsen er betraktelig bedre med kaskaderegulering enn uten.



10.2. Forholdsregulering og synkronisering

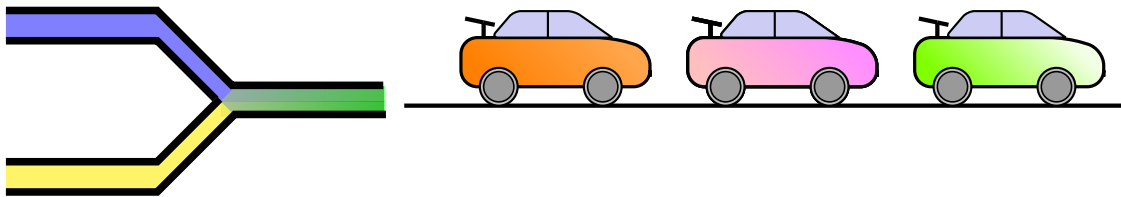


Alternative kilder: §15.2 i [Seborg et al., 2016]; [Hägglund, 2001]

Forholdsregulering: (eng. ratio control) man ønsker å opprettholde en ønsket ratio (forhold) mellom to eller flere variabler (prosess-, forstyrrelses- og/eller pådragsvariabler).

Merk: I [Seborg et al., 2016] er fokuset mest på forholdet mellom en forstyrrelsesvariabel og en manipulert variabel (f.eks. to væskestrømmer), altså en slags foroverkobling. Vi vil dog se litt bredere på det, siden konseptet bak forholdsregulering er høyst relevant til andre typer problemer, som **synkronisering** av robot-ledd og multiagent-systemer.

Forholdsregulering er en vanlig reguleringsmetode i prosessindustrien, hvor man ofte ønsker å opprettholde et ønsket blandingsforhold. Et typisk eksempel i denne sammenheng, hvor man kan ta i bruk forholdsregulering er forbrenningsprosesser, hvor man bør ha en ideell ratio mellom oksygen og forbrenningsmaterialet.



- a) Blandingsforhold: riktig fordeling av blått og gult slik at man får riktig grønnfarge. b) Platooning: hver førerløse bil bestemmer posisjonen/hastigheten til bilen bak seg.

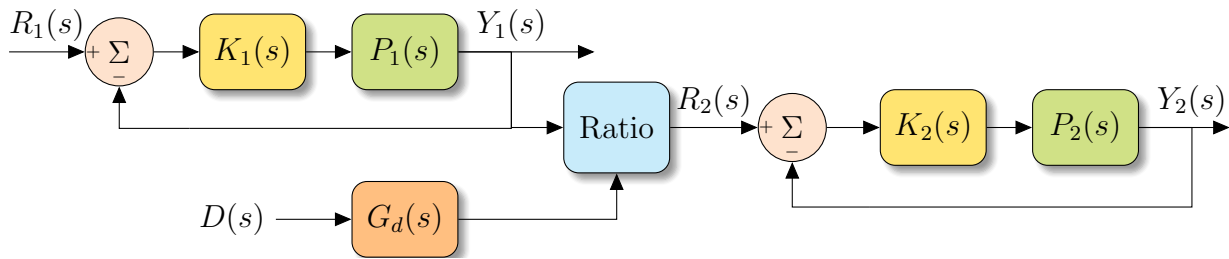
Figur 10.2: Applikasjoner egnet for forholdsregulering/synkronisering.

Blandingsforhold: Ta for eksempel “systemet” vist i a) figur 10.2. Ønsket er å blandet to stoff, **A** (blått) og **B** (gult), i et ønsket forhold slik at vi får riktig blanding $C = \frac{A}{B}$ (grønn). Det kan være seg at vi kan regulere/manipulere bare én eller begge av disse størrelsene (f.eks. væskestrømmer relatert til **A** og **B**), hvor den førstnevnte situasjonen gir oss en form for foroverkobling, mens den andre gir oss en kaskade-lignende struktur (merk at det dog ikke standard kaskader-regulering).

Platooning: De samme ideene bak dette konseptet kan, som tidligere nevnt, også brukes til å synkronisere flere autonome “agenter”. I platooning av biler (se b) i figur 10.2), for eksempel, hvor man ønsker å lage et “tog” av førerløse biler, så ønsker man kanskje at bil tre (den oransje) skal ha samme fart som bil nummer to (den rosa), som velger sin fart i forhold til farten til den første bilen (den grønne). Eventuelt, så kanskje ønsker man at posisjonen til bil 2, p_2^o , skal være $p_2^o = p_1 - 10$, men posisjonen til bil tre skal være $p_3^o = p_2 - 12$, etc.^a

^aDet er selvsagt mye annet man må tenke på i et slikt system, men ideen bak forholdsregulering/synkronisering er uansett meget relevant.

En illustrasjon av et reguleringsystem med forholdsregulering er vist i figur 10.3. Den blå blokken i midten kalles ofte for en **ratio/blandings-stasjon** (eng. ratio station), hvor



Figur 10.3: Forholdsregulering: en ratio (mer generelt, en funksjon) av den ene prosessvariabelen og/eller en forstyrrelse (uregulert variabel) er referansen til den andre reguleringsløyfen.

referansen, $r_2(t)$, til den andre sløyfen bestemmer av utgangen, y_1 , til den første sløyfen og/eller en målt forstyrrelse $\hat{d}(t)$ ($\hat{D}(s) = G_d(s)D(s)$). Vi har er to relevante tilfeller:

1. $r_2(t) = k_r \hat{d}(t)$: en slags “foroverkobling”, hvor i stedet for å foroverkoble forstyrrelsen for å bestemme deler av pådraget direkte, så brukes den til å indirekte bestemme pådraget ved å sette referansen.
2. $r_2(t) = k_r y_1(t)$: en slags serie-tilbakekoblingsstruktur (lik kaskade, men er ikke det), hvor utgangen til den ene reguleringsløyfen (masteren/lederen) setter referansen til den andre sløyfen (slaven/følgeren).

Merk: et alternativ til scenario 2 er å ta $r_2(t) = k_r r_1(t)$, noe som fører til to separate tilbakekoblingsløyfer. Ulempen med dette er at man ikke vil prøve å opprettholde et ønsket forhold/synkronisering under transienter. På den annen side, så unngår man med denne metoden at uønskede forstyrrelser som påvirker utgangen til sløyfe 1 (altså y_1) også vil påvirke sløyfe 2, slik den vil ved forholdsreguleringen $r_2(t) = k_r y_1(t)$. I [Hägglund, 2001] foreslås også generalisering at dette, nemlig $r_2(t) = k_r [\gamma y_1(t) + (1 - \gamma)r_1(t)]$ hvor γ er et tall mellom 0 og 1 ($\gamma \in [0, 1]$). La oss se på et eksempel som motiverer denne mer generelle metoden:

Eksempel 10.3. (Hvordan velge ratio-stasjon^a) Gitt et system på formen i figur 10.3 med overføringsfunksjoner

$$P_1(s) = \frac{1}{(10s + 1)^2} \quad \text{og} \quad P_2(s) = \frac{1}{(2s + 1)^2}.$$

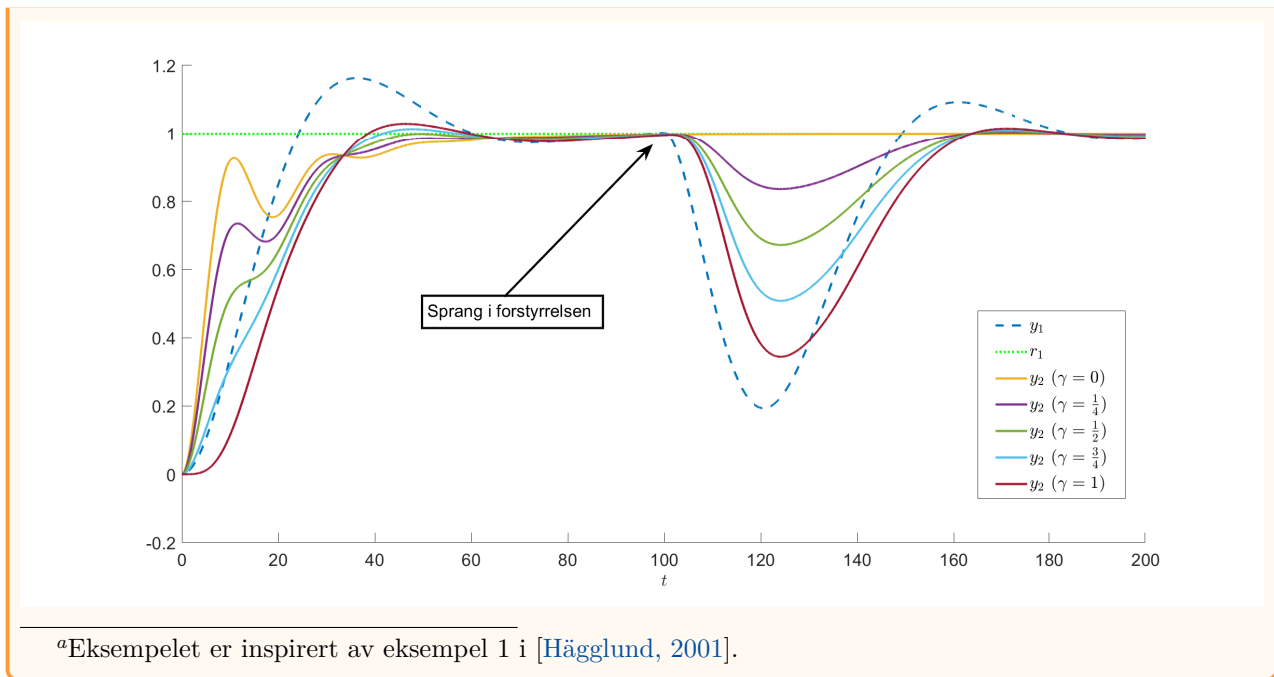
Regulatorerne er begge av typen PI:

$$K_1(s) = 1 + \frac{1}{10s} \quad \text{og} \quad K_2(s) = 1 + \frac{1}{10s}.$$

Det er ønskelig at $y_2(t) = y_1(t)$. Vi skal derfor prøve

$$r_2(t) = \gamma y_1(t) + (1 - \gamma)r_1(t)$$

for forskjellige verdier av γ mellom 0 og 1.



Fun facts, bemerkninger og annet dill dall (you may skip)

Multi-agent systemer og konsensus: TODO [Olfati-Saber et al., 2007].

10.3. Parameterstyring

▶ uHsfj2-cRIA

Alternative kilder: §16.5.1 i [Seborg et al., 2016]; [Åström and Wittenmark, 2013, kap. 9]; [Rugh and Shamma, 2000]; Brian Douglas video; Div. MATLAB-eksempler.

Nåværende problem: Alle ekte reguleringsystemer er ulineære på en eller annen måte, og regulatorer designet kun vha. lineære betraktninger vil derfor ikke fungere optimalt for alle arbeidspunkter i et stort arbeidsområde. Så hvordan lage en regulator som fungerer over et vidt spekter av arbeidspunkter?

Tross dette, vil ofte lineære reguleringsstrategier kunne gjøre det bra gitt at enten

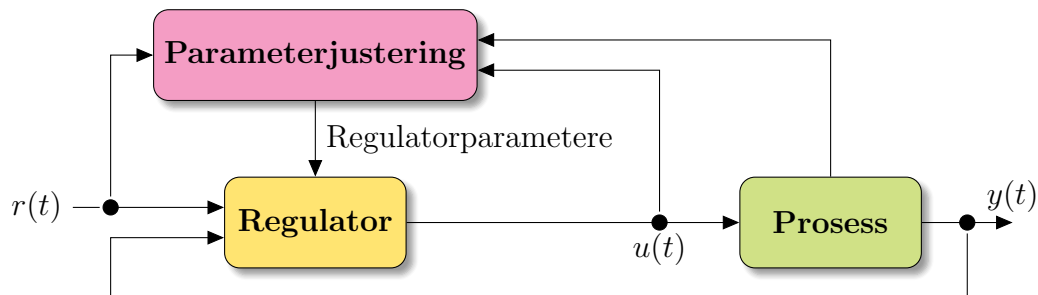
1. ulineariteten er «milde», slik at regulatoren klarer å kompensere for dem; eller
2. man kun opererer i et lite område rundt et gitt arbeidspunkt.

Vi skal nå se på en metode, kalt parameterstyring (eng. “Gain/parameter scheduling”) som kan la en regulator være effektiv i et større arbeidsområde selv for ulineære systemer.

Parameterstyring: Velg regulatorparameterne ut fra en eller flere justeringsvariabler^a, slik som illustrert i figur 10.4. Justeringsvariabelen(e) kan for eksempel være referansen, regulatorutgangen, **endogene** prosessvariabler som prosessutgangen, eller **eksogene** prosess-

variabler som endringer i omgivelsene (“forstyrrelser”) til systemet eller tilstander man ikke tar direkte hensyn til i sin forenklete modell av systemet.

“På engelsk kaller man disse parameterne for scheduling variable(s), noe man kan løst oversette til “planleggingsvariabel”, som egentlig er et passende navn, men justeringsvariabler er bedre etter min mening.



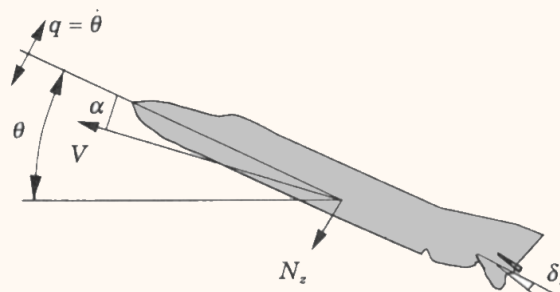
Figur 10.4: Illustrasjon av en reguleringsløyfe med parameterstyring: regulatorparameterne justeres baserte på referansen og/eller pådragsignalet og/eller prosess-relaterte variabler.

Et bruksområde hvor parameterstyring tidlig ble tatt i bruk var auto-pilot-systemer for jagerfly tidlig på 1950-tallet. Slike fly må kunne operere over et vidt spekter av forskjellige hastigheter og flygehøyder. Det ble dog observert at selv om en lineær regulator kunne fungere godt i et visst arbeidsområde, ville den ikke fungere godt for andre. Løsning på dette var derfor å ta i bruk parameterstyring.

Eksempel 10.4. Jagerfly (eksempel og figurer fra [Åström and Wittenmark, 2013]):

En illustrasjon av et jagerfly er vist i figuren til høyre, hvor

- θ : “pitch”-vinkelen
- $q = \dot{\theta}$: pitch-raten
- N_z : normal-akselerasjonen
- δ_e : høyderor-vinkelen

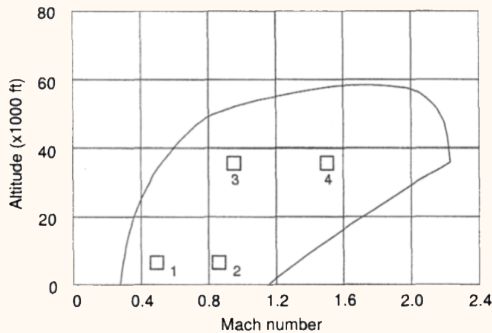


Tilstandsvektoren til systemet er $x = [N_z, \theta, \delta_e]^T$, mens pådraget u kan påvirke høyderorsvinkelen. Hvis man antar at flyet er et stivt legeme, så kan systemets dynamikk tilnærmes som

$$\dot{x} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ 0 & 0 & -a \end{bmatrix} x + \begin{bmatrix} b_1 \\ 0 \\ a \end{bmatrix} u.$$

hvor parameterne $(a_{11}, a_{12}, a_{13}, a_{21}, a_{22}, a_{23}, b_1)$ avhenger av flygehøyden og mach-tall, se figuren under, mens $a = 14$ er konstant. I tabellen til høyre under kan man se verdiene for forskjellige scenarier (“flight conditions” (FC)), hvor λ_1 og λ_2 er egenverdiene til den åpne sløyfen. Vi ser fra disse at den åpne sløyfen bare er stabil for FC4, men da svært dårlig dempet, og ustabil ellers. Grunnet de store forskjellene i systemparameterne, vil man her

være nødt til å ta i bruk parameterstyring, hvor justeringsvariablene bør være flygehøyden (eng. "attitude") og mach-tallet (eng. "mach number").



	FC 1	FC 2	FC 3	FC 4
Mach	0.5	0.85	0.9	1.5
Altitude (feet)	5000	5000	35000	35000
a_{11}	-0.9896	-1.702	-0.667	-0.5162
a_{12}	17.41	50.72	18.11	26.96
a_{13}	96.15	263.5	84.34	178.9
a_{21}	0.2648	0.2201	0.08201	-0.6896
a_{22}	-0.8512	-1.418	-0.6587	-1.225
a_{23}	-11.39	-31.99	-10.81	-30.38
b_1	-97.78	-272.2	-85.09	-175.6
λ_1	-3.07	-4.90	-1.87	
λ_2	1.23	1.78	0.56	$-0.87 \pm 4.3i$

Valg av justeringsvariabler: I eksempel 10.4 så vi at man kan velge systemvariabler som ikke nødvendigvis er tilstander i ens (forenklete) modell som justeringsvariabler. Det også vanlig å bruke en tilstand, f.eks. prosessutgangen (se f.eks. eks. 10.5), eller til og med referansen og utgagnen fra regulatoren som justeringsvariabel.

Når man skal velge en justeringsvariabel, er det dog noen hensyn man bør da. For eksempel, gitt en parameterstyrt PID-regulator, la oss anta at man bare har én slik justeringsvariabel, si σ . Det betyr det at regulatorparameterne er funksjoner av σ : $k_P(\sigma)$, $T_I(\sigma)$ og $T_D(\sigma)$. Når man implementerer en slik metode, er det svært viktig at disse parameterne ikke endrer seg for raskt, siden dette kan føre til ustabilitet:

Warning! Man har begrensede stabilitetsgarantier med parameterstyring mellom designpunktene. Når man velger justeringsvariablene som setter regulatorparameterne, så bør derfor disse helst variere sakte med tid, samt må målte signaler som inneholder mye støy brukes med forsiktighet hvis de ikke filtreres tilstrekkelig.

Prosedyre: Det er mange måter man kan utvikle en parameterstyrt regulator på avhengig av system, regulator og justeringsvariabler. En vanlig og effektiv strategi er dog følgende:

Lineariserings-basert prosedyre for utvikling av parameterstyrt regulator:

Gitt et ulineært system

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \mathbf{u}, \mathbf{w}), \quad \mathbf{y} = \mathbf{h}(\mathbf{x}),$$

hvor \mathbf{x} er tilstandene, \mathbf{u} er pådragene, og \mathbf{w} er en vektor av eksterne signaler (f.eks. forstyrrelser) og umodellerte tilstander.

1. Velg én eller flere (målbare) justeringsvariabler σ , slik at $(\mathbf{x}_e(\sigma), \mathbf{u}_e(\sigma), \mathbf{w}_e(\sigma))$ representerer et sett av likevektspunkter til systemet for visse verdier av σ : $\mathbf{f}(\mathbf{x}_e(\sigma), \mathbf{u}_e(\sigma), \mathbf{w}_e(\sigma)) = 0$ gitt at $\sigma \in \{\sigma_1, \sigma_2, \dots, \sigma_m\}$;
2. Lineariser systemet om slike likevektspunkter (se § 2.5) slik at σ er en parameter til det lineariserte systemet: $\frac{d}{dt} \delta \mathbf{x} = \mathbf{A}(\sigma) \delta \mathbf{x} + \mathbf{B}(\sigma) \delta \mathbf{u}$, $\delta \mathbf{y} = \mathbf{C}(\sigma) \delta \mathbf{x}$;
3. Design regulatorparametere ved forskjellige verdier av justeringsvariablene;

4. Velg en interpoleringsmetode som automatisk veksler mellom disse regulatorparameterne basert på verdien til justeringsvariablene.

^aEt system på formen $\dot{\mathbf{z}} = \mathbf{A}(\sigma)\mathbf{z} + \mathbf{B}(\sigma)\mathbf{u}$, hvor σ er en vektor av *parametere* som antas å være konstant (selv om den egentlig representerer systemvariabler), kalles for et lineært, parameter-varierende system, hvorfra *parameterstyring* følger.

Hvordan veksle mellom forskjellige regulatorparametere? For å veksle mellom de regulatorparameterne som har blitt utledet for forskjellige verdier av justeringsvariabelen, så bruker man en form for [interpolasjonsmetode](#). Ved kun én variabel er dette relativt enkelt, og vanlige interpolasjonsalternativer er da stykkvis-konstant (fører til hopp i pådraget), stykkvis-lineære eller interpolering vha. høyere-ordens polynomer (f.eks kubiske).

Hvis man har flere enn én justeringsvariabel blir dette fort mer komplisert. Da er det vanligst å ta i bruk en oppslagstabell (eng. “look-up-tables”). I Simulink kan man bruke følgende oppslagstabeller : <https://se.mathworks.com/help/simulink/lookup-tables.html>

Eksempel 10.5. (Tank med varierende areal^a) Gitt en tank hvor tverrsnittarealet er gitt av den vertikale høyden, slik at hvis væsknivået i tanken er h , så er volumet med væske i tanken $V = \int_0^h A(\rho)d\rho$. Det kommer en væskestrøm q_{inn} inn i tanken, mens væskestrømmen ut er gitt ved $q_{ut} = c\sqrt{h}$. Vi antar konstant massetetthet til væsken, slik at vi fra 3.2 får

$$\dot{h} = \frac{1}{A(h)} [q_{inn} - c\sqrt{h}].$$

Om et ønsket arbeidspunkt, med settpunkt h^* og tilsvarende nominelle innstrøm $q_{inn}^* = c\sqrt{h^*}$, kan det lineariserte systemet bli representert ved følgende overføringsfunksjon:

$$P(s) = \frac{b}{s + a}$$

hvor

$$b = \frac{1}{A(h^*)} \quad \text{og} \quad a = \frac{q_{inn}^*}{2A(h^*)h^*} = \frac{c\sqrt{h^*}}{2A(h^*)h^*}.$$

La $u(t) = q_{inn}(t)$ og $e(t) = h^* - h(t) = r - y$. En PI-regulator

$$u(t) = q_{inn}^* + k_P \left(e(t) + \frac{1}{T_I} \int_0^t e(\tau) d\tau \right),$$

med

$$k_P = \frac{2\zeta\omega - a}{b} \quad \text{og} \quad T_I = \frac{2\zeta\omega - a}{\omega^2}$$

fører til en lukket sløyfe med relativ dempningsfaktor ζ og naturlig frekvens (udempet svingefrekvens) ω . En mulig justeringsvariabel her er derfor $\sigma = \hat{r}$ hvor \hat{r} er utgangen til en referanseglatte (se § 8.2) som glatte ut referansen $r = h^*$ (eventuelt kan man også bruke rykkfri overføring her; se § 9.3). En parameterstyrt PI-regulator for systemet er dermed

$$u(t) = c\sqrt{\sigma} + k_P(\sigma) \left(e(t) + \frac{1}{T_I(\sigma)} \int_0^t e(\tau) d\tau \right),$$

hvor

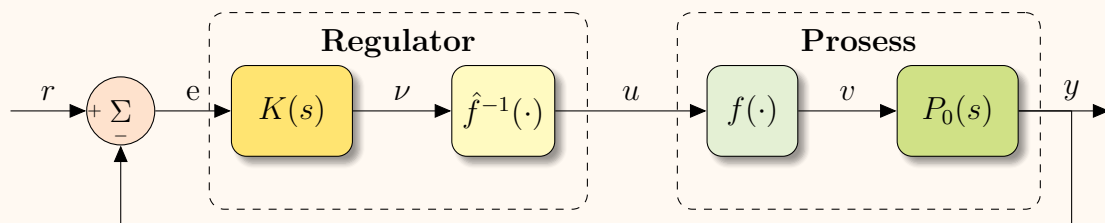
$$k_P(\sigma) = 2\zeta\omega A(\sigma) - \frac{c\sqrt{\sigma}}{2\sigma} \quad \text{og} \quad T_I = \frac{2\zeta}{\omega} - \frac{c\sqrt{\sigma}}{2A(\sigma)\sigma\omega^2}.$$

Hvis c er veldig liten, kan f.eks. $\hat{k}_P = 2\zeta\omega A(\sigma)$ og $T_I = 2\zeta/\omega$ være enkle alternativer.

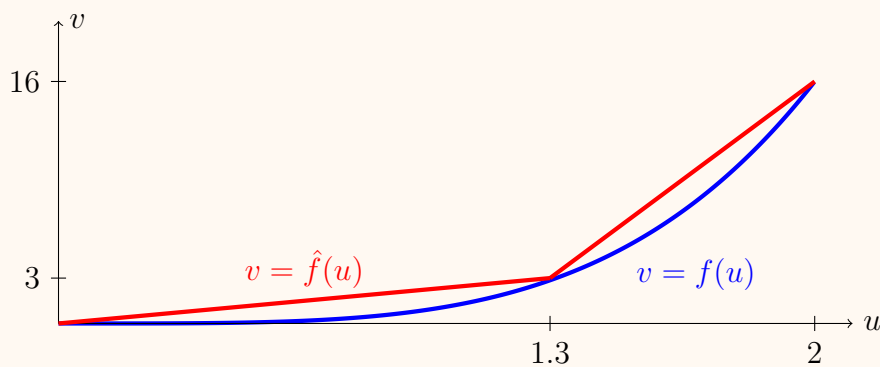
^aEksempelet er inspirert av eksempel 9.2 [Åström and Wittenmark, 2013].

Noen ganger kan det også være hensiktsmessig å bruke regulatorutgangen som en justeringsvariabel. Følgende eksempel viser hvordan man kan bruke lignende ideer til å også kompensere for ulineariteter relatert til pådragsorganet. Selv om dette strengt tatt ikke er parameterstyring i ordets rette forstand, så er denne metoden såpass nært beslektet (og ikke minst nyttig i noen sammenhenger!) at jeg har valgt å ta den med her.

Eksempel 10.6. Ulineært pådragsorgan (eks. fra [Åström and Wittenmark, 2013]):



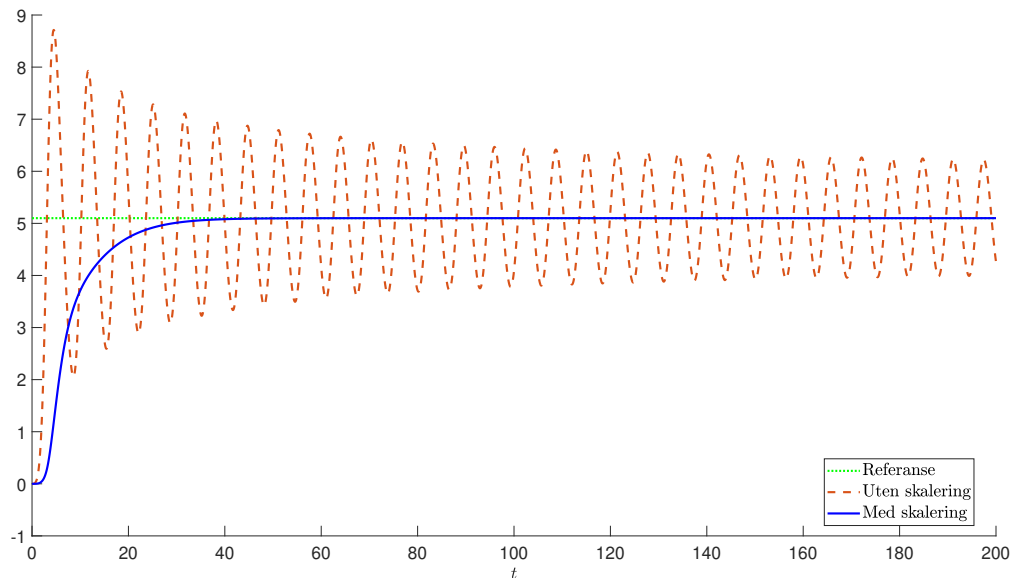
Gitt systemet i figuren over, hvor funksjonen $f(u) = u^4$ tilsvarer et ulineært pådragsorgan (f.eks. en reguleringsventil). **Mål:** Vi ønsker å finne en tilnærming, $\hat{f}^{-1}(\cdot)$, av den inverse funksjonen til $f(\cdot)$ slik at vi får et mest mulig lineært forhold mellom utgangen ν av regulatoren og inngangen v til prosessen, altså ønsker vi ideelt sett at $v = f(\hat{f}^{-1}(\nu)) \approx \nu$.



Anta at vi har begrenset kjennskap til den virkelige funksjonen f , slik at tilnærmingen \hat{f} tilsvarer to rette streker som kobler sammen punktene $(0,0)$ og $(1.3,3)$, samt $(1.3,3)$ og $(2,16)$, slik som vist i figuren over. Dette tilsvarer ca. følgende skaleringsfunksjon:

$$\hat{f}^{-1}(\nu) = \begin{cases} 0.433\nu & \text{når } 0 \leq \nu \leq 3 \\ 0.0538\nu + 1.139 & \text{når } 3 \leq \nu \leq 16. \end{cases}$$

En sammeligning av responsen med og uten denne skaleringen er vist i plottet nedenfor for $P_0(s) = \frac{1}{(s+1)^3}$ med $K(s) = 0.15 \left(1 + \frac{1}{s}\right)$. Det er tydelig at skaleringen er svært fordelaktig i dette tilfellet, selv for en relativt enkel skaleringsfunksjon. Ved behov, kan man lett finne en bedre tilnærming ved å koble sammen flere punkter.



Eksempel 10.7. (Ulineær måling) Gitt følgende system med en ulineær måling:

$$\begin{aligned}\dot{x} &= -x + b \\ y &= \tanh(x).\end{aligned}$$

hvor $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$ er den hyperbolske tangensfunksjonen. Vårt mål er å lage en regulator som fungerer over et bredt spekter av arbeidspunkter ved å ta i bruk parameterstyring.

For en gitt justeringsvariabel σ , la $y_a(\sigma) = \sigma$ være et ønsket arbeidspunkt. Med andre ord så bruker vi σ til parametrisere de ønskede arbeidspunktene. Et slikt arbeidspunkt må naturligvis tilsvare et tvunget likevektspunkt, gitt ved $x_a(\sigma) = u_a(\sigma) = \tanh^{-1}(\sigma)$. Det lineariserte systemet om et slikt arbeidspunkt er

$$\begin{aligned}\dot{\delta x} &= -\delta x + \delta u \\ \delta y &= (1 - \sigma^2)\delta x\end{aligned}$$

hvor $\delta x \approx x - x_a(\sigma)$, etc., nært arbeidspunktet, samt hvor det har blitt brukt at $\frac{d}{dx} \tanh(x) = 1 - \tanh(x)^2$. Ved å la $Y(s) = \mathcal{L}\{\delta y\}$ etc., så tilsvare dette overføringsfunksjonen

$$\frac{Y(s)}{U(s)} = P(s) = \frac{1 - \sigma^2}{s + 1}$$

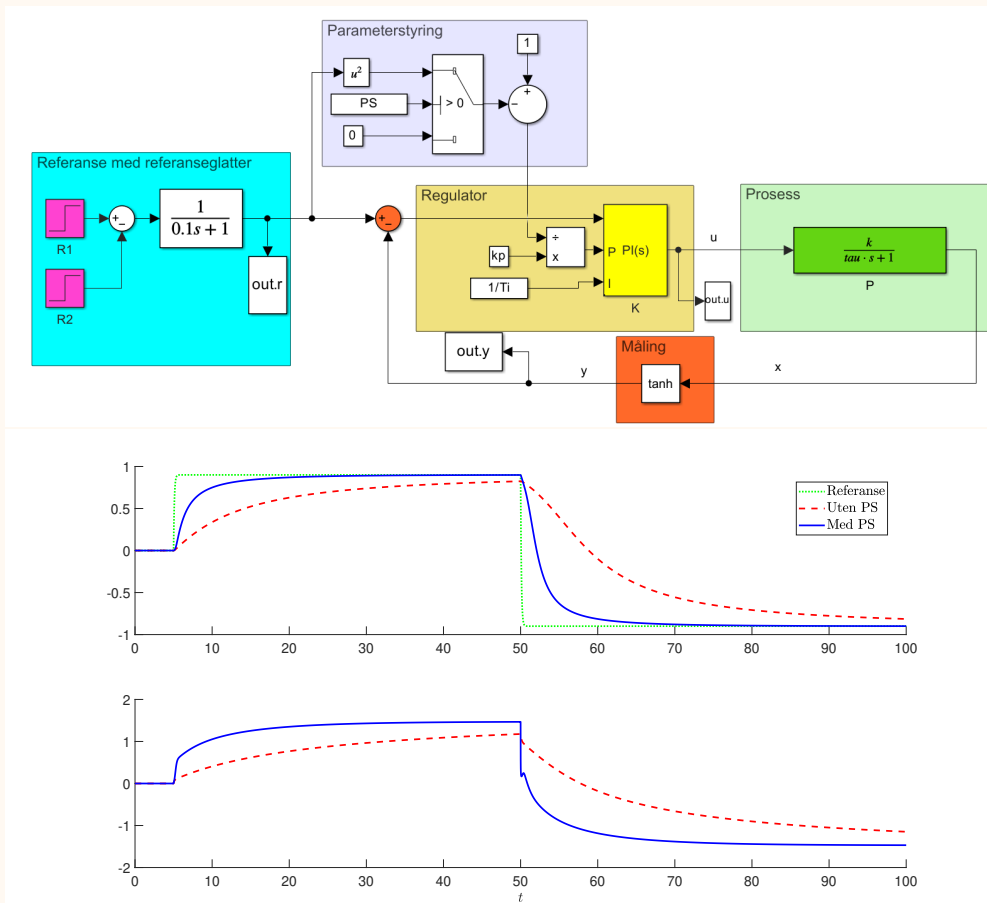
som har forsterkning $k = 1 - \sigma^2$ og tidskonstant $\tau = 1$. Hvis man derfor f.eks. ønsker at den lukkede sløyfen (tilsvarende det lineariserte systemet) skal være

$$G_{LS}(s) = \frac{1}{\tau_c s + 1}$$

så får man fra [Direktesyntese](#) en PI-regulator på [parallellform](#) med

$$k_p(\sigma) = \frac{1}{(1 - \sigma^2)\tau_c} \quad \text{og} \quad T_I = 1.$$

Man kan så ta $\sigma = \hat{r}$ hvor \hat{r} er utgangen til en referanseglatter som glatter referansen r . Et Simulink-diagram av et slikt system med en parameterstyrt PI-regulator er vist under, sammen med et plott som viser ytelsesforskjellen med og uten (σ satt til 0) parameterstyring for $\tau_c = 10$.



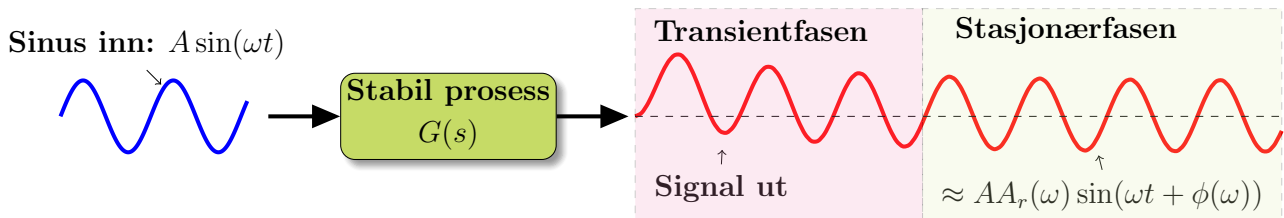
11. Frekvensanalyse

Alternative kilder: Kapittel 14 og vedlegg J i [Seborg et al., 2016]; [Balchen et al., 2016]; §8.2 i [Ogata et al., 2010].

Hva er frekvensanalyse?



Frekvensanalyse baserer seg på et systems frekvensrespons. Med frekvensrespons menes den statiske responsen (altså responsen etter alle transienter har dødd ut) til et system på en sinusformet inngang med en gitt frekvens; se figur 11.1. I frekvensresponsmetoder varierer vi frekvensen på inngangssignalet over et visst område og studere den resulterende responsen.

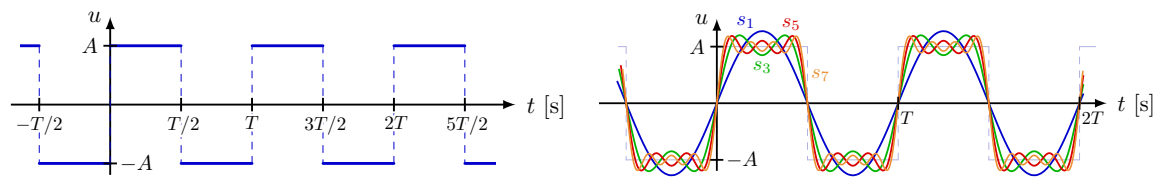


Figur 11.1: Et sinusignal på inngang til et stabilt, lineært, tids-invariant system fører til (etter en transientfase har tatt slutt) et nytt sinussignal ut med samme frekvens, men med en annen amplitude og fase.

Men hvorfor bryr vi oss om det? Jo, pga. Fourier-transformen! [Fourier-rekker](#) kan brukes til å representere «alle» signaler vi bryr oss om (signalet må være stykkvis kontinuert, ha begrenset amplitude, og må kunne deles opp i biter som alle «forelenges» som et periodisk signal); se f.eks. figur 11.2.

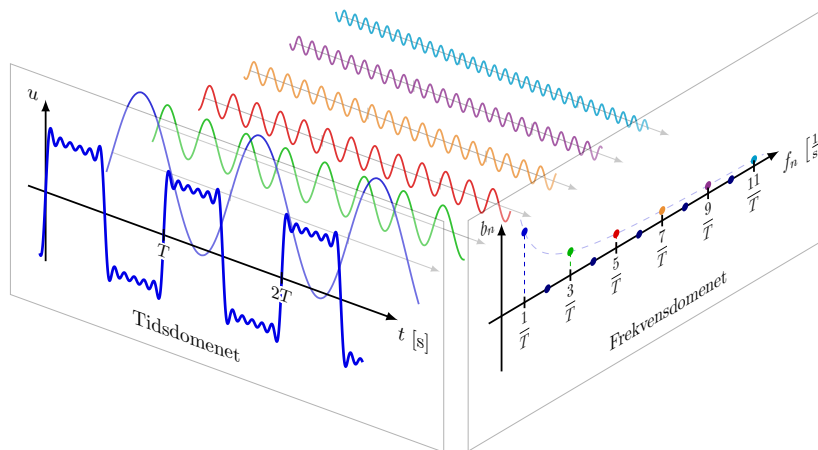
Fun facts, bemerkninger og annet dill dall (you may skip)

For mer om Fourier-transformen og Fourier-serier, så anbefaler jeg følgende mesterverk:



a) Original firkant-puls funksjon.

b) Fourier-rekke approksimasjon.



c) Tidsdomenet vs frekvensdomenet.

Figur 11.2: Fourier-rekke approksimasjon av firkantspuls-funksjon; figur fra https://tikz.net/fourier_series/.

▶ spUNpyF58BY og ▶ 1JnayXHhjlq

Frekvens(respons)-analyse er et nyttig verktøy som kan brukes til flere ting, deriblant:

- **Stabilitetundersøkelser:** man kan analysere stabiliteten til en lukket sløyfe fra frekvensresponsen til den åpne sløyfen. Det typiske scenariet her er å utforske om en regulator $K(s)$ i serie med en $P(s)$ vil føre til en stabil lukket-sløyfe, ved å se på om $1 + K(s)P(s)$ har alle sine nullpunkt i venstre halvplan eller ikke (Nyquists stabilitetskriterium, eventuelt det mindre generelle Bode-kriteriumet).
- **Robusthets-analyse:** Undersøke stabilitetsmarginene til den lukkede sløyfen ved hjelp av den åpne, for å forutsi robusthet med tanke på usikkerhet og endringer i prosessen.
- **Regulatorerdesign:** Designe regulatorer for å oppnå stabilitet og andre kriterier, som f.eks. ønskede marginer.
- **Systemidentifikasjon:** Eksperimentelt tilpasse en prosessmodell fra frekvensresponsen til et system.
- **Nyttig for andre ting også:** F.eks ¹ [redusere oscillasjoner i skysrapere:](#) ▶ flU4SAgy60c

¹følgende video er et mye brukt eksempel på at et system blir utsatt for en periodisk kraft tilsvarende dets resonansfrekvens: <https://youtu.be/kZNjbWY6c7c>. Den spektakulære avslutningen er dog trolig ikke en direkte konsekvens av dette (se <https://youtu.be/mXTSnZgrfxM>).

Merk: Frekvensanalyse tar utgangspunkt i lineære, tidsinvariante systemer, og disse må (med unntak av det teoretiske Nyquist-kriteriet) være stabile i åpen sløyfe.

11.1. Fase og amplitude ▶ ZUjQMIXfy5s&t=248

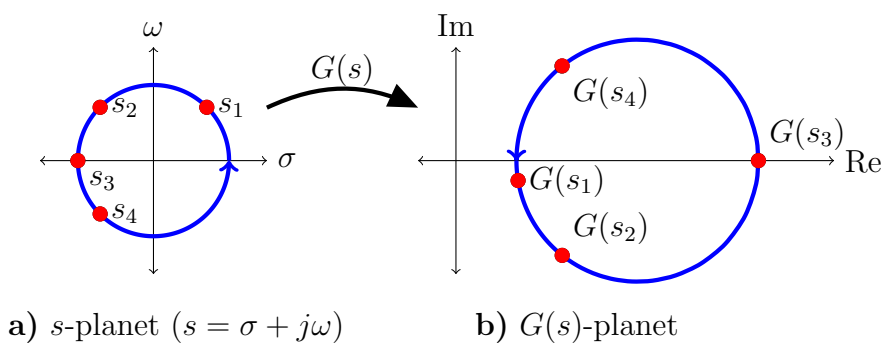
Som tidligere nevnt, så handler frekvensanalyse om å se på systemets respons over et spekter av frekvenser i stedet for responsen over tid. For frekvensanalyser tar man utgangspunkt i at inngangssignalet er en sinussvingning med en gitt amplitude A og økende frekvens ω . Et ideelt system vil klare å følge inngangssignalet $A \sin(\omega t)$ og gi en respons med helt lik frekvens og en amplitude som ikke er avhengig av frekvensen til inngangssignalet. Mer generelt vil dog både utgangssignalet sin amplitude og fase endres med frekvensen til inngangssignalet. Denne endringer er gitt av prosessens overføringsfunksjon for de forskjellige frekvensene. En måte å studere dette på er ved å se på konturplottet til overføringsfunksjonen.

11.1.1 Kontur-plott

En overføringsfunksjon tar et punkt (σ, ω) i det komplekse s -planet ($s = \sigma + j\omega$) og overfører det til det komplekse $G(s)$ -planet. Dette er illustrert ved hjelp av et kontur-plott i figur 11.3, hvor ved å gå *mot* klokken langs enhets sirkelen i dette tilfellet fører til en mot-klokken bevegelse langs en annen sirkel i $G(s)$ -planet.

For frekvensanalyse vil vi se bort fra σ ved å sette den lik $\sigma = 0$, og kun se på konturer tilsvarende $G(j\omega)$. Dette betyr at vi kun beveger oss langs den imaginære, vertikale ω -aksen i s -planet, som regel fra 0 og til ∞ , muligens med en rask svipptur innom $-\infty$ og derfra tilbake til 0. Sistnevnte finner man enkelt fra førstnevnt pga. følgende:

Speiling om den reelle akse: Hvis $G(\sigma + j\omega) = a + jb$, så er $G(\sigma - j\omega) = a - jb$.



Figur 11.3: Kontur-plott av en overføringsfunksjon.

11.1.2 Amplituderatio og fasevinkel

For en overføring funksjon $G(s)$ så definerer vi

$$\mathcal{R}_G(\omega) := \Re_e\{G(j\omega)\} \quad \text{og} \quad \mathcal{I}_G(\omega) := \Im_m\{G(j\omega)\},$$

altså henholdsvis den reelle delen av $G(j\omega)$ og den imaginære delen. Fra dette kan vi definere to viktige størrelser, nemlig amplituderatioen og fasevinkelen, som lar oss relatere et sinussignal med en gitt frekvens på inngangen, til et nytt sinussignal på utgangen (etter alle transienter har dødd ut **hvis** systemet er stabilt), slik vist i figur 11.1:

Amplituderatioen, $A_r(\cdot)$, til en overføringsfunksjon $G(s)$ er

$$A_r(\omega) := |G(j\omega)| = \sqrt{\mathcal{R}_G(\omega)^2 + \mathcal{I}_G(\omega)^2}; \quad (\text{Amplituderatio})$$

fasevinkelen, $\phi(\cdot)$, er gitt ved

$$\phi(\omega) := \angle G(j\omega) = \tan^{-1}(\mathcal{I}_G(\omega)/\mathcal{R}_G(\omega)). \quad (\text{Fasevinkel})$$

Disse størrelsene kan lett regnes ut ved å faktorisere en overføringsfunksjon inn i mindre «basisdeler». For å vise dette, la oss anta at overføringsfunksjonen $G(s)$ har følgende form:

$$G(s) = \frac{G_a(s)G_b(s)G_c(s)\cdots}{G_1(s)G_2(s)G_3(s)\cdots}.$$

Dens **Amplituderatio** for en gitt frekvens ω er da gitt av

$$A_r(\omega) = |G(j\omega)| = \frac{|G_a(j\omega)| |G_b(j\omega)| |G_c(j\omega)| \cdots}{|G_1(j\omega)| |G_2(j\omega)| |G_3(j\omega)| \cdots}, \quad (11.1)$$

mens tilsvarende **Fasevinkel** finner man fra

$$\begin{aligned} \phi(\omega) = \angle G(j\omega) &= \angle G_a(j\omega) + \angle G_b(j\omega) + \angle G_c(j\omega) + \cdots \\ &\quad - \angle G_1(j\omega) - \angle G_2(j\omega) - \angle G_3(j\omega) - \cdots. \end{aligned} \quad (11.2)$$

Merk også at ved å ta den naturlige logaritmen av uttrykket til amplituderatioen på begge sider så får man

$$\begin{aligned} \ln |G(j\omega)| &= \ln |G_a(j\omega)| + \ln |G_b(j\omega)| + \ln |G_c(j\omega)| + \cdots \\ &\quad - \ln |G_1(j\omega)| - \ln |G_2(j\omega)| - \ln |G_3(j\omega)| - \cdots. \end{aligned}$$

Ved å multiplisere dette med skaleringsfaktoren $20/\ln(10)$ får man amplituderatioen i **desibel**:²

$$A_r^{dB}(\omega) = 20 \ln(A_r(\omega)) / \ln(10) = 20 \log_{10}(A_r(\omega)).$$

Overføringsfunksjonene vi er interessert i kan bygges opp fra 5 basis deler: forsterkning, nullpunkter, reelle poler, komplekse poler, og tidsforsinkelser. Vi skal derfor se på effekten av hver av disse på amplituderatioen og fasevinkelen.

Effekt av forsterkning, nullpunkter, poler, og tidsforsinkelser:

Forsterkning En overføringsfunksjon med ren forsterkning,

$$G(s) = k,$$

har triviell **Amplituderatio**, $A_r = |n|$, og **Fasevinkel** ϕ lik 0 når $k > 0$ eller 180° når $k < 0$.

²Vi skal som regel **ikke** bruke desibel i disse notatene.

Nullpunkt: Gitt en overføringsfunksjon med et enkelt (reelt) nullpunkt:

$$G(s) = s + a, \quad a \in \mathbb{R}.$$

Vi har $G(j\omega) = j\omega + a$, slik at

$$\mathcal{R}_G(\omega) = a \quad \text{og} \quad \mathcal{I}_G(\omega) = \omega$$

Fra (Amplituderatio) og (Fasevinkel) er det tydelig at ($\text{sgn}(\cdot)$ er fortegnfunksjonen)

$$A_r(\omega) = |G(j\omega)| \rightarrow \infty \quad \text{og} \quad \angle G(j\omega) \rightarrow \pm \text{sgn}(a) \cdot 90^\circ \quad \text{når} \quad \omega \rightarrow \pm\infty.$$

Reelle poler: Gitt en overføringsfunksjon med én enkelt reell pol:

$$G(s) = \frac{1}{s + b}, \quad b \in \mathbb{R}.$$

Vi har

$$G(j\omega) = \frac{1}{j\omega + b} = \frac{b - j\omega}{\omega^2 + b^2} \quad \implies \quad \mathcal{R}_G(\omega) = \frac{b}{\omega^2 + b^2} \quad \text{og} \quad \mathcal{I}_G(\omega) = \frac{-\omega}{\omega^2 + b^2}$$

slik at

$$A_r(\omega) = |G(j\omega)| \rightarrow 0 \quad \text{og} \quad \angle G(j\omega) \rightarrow \mp \text{sgn}(b) \cdot 90^\circ \quad \text{når} \quad \omega \rightarrow \pm\infty.$$

Komplekse poler: Gitt en overføringsfunksjon med et par komplekskonjugerte poler:

$$G(s) = \frac{1}{(s^2 + 2\zeta\omega_0 s + \omega_0^2)}, \quad -1 < \zeta < 1.$$

Det er tydelig at vi har $A_r(\omega) = |G(j\omega)| \rightarrow 0$ når $\omega \rightarrow \infty$ siden

$$G(j\omega) = \frac{1}{-\omega^2 + 2\zeta\omega_0\omega j + \omega_0^2} = \frac{\omega_0^2 - \omega^2 - 2\zeta\omega_0\omega j}{4\zeta^2\omega_0^2\omega^2 + (\omega_0^2 - \omega^2)^2},$$

mens $\mathcal{I}_G(\omega)/\mathcal{R}_G(\omega) = 2\zeta\omega_0\omega/(\omega^2 - \omega_0^2)$, slik at $\angle G_0(j\omega) \rightarrow \pm \text{sgn}(\zeta) \cdot 180^\circ$ når $\omega \rightarrow \pm\infty$.

Resonanstopp: komplekse poler fører til en resonanstopp ved resonansfrekvensen ω_r , som tilsvarer frekvensen hvor uttrykket $(\frac{\omega_0^2 - \omega^2}{\omega})^2 + (2\zeta\omega)^2$ har sin minimumsverdi. For $0 \leq \zeta \leq 1/\sqrt{2}$, er denne gitt ved $\omega_r = \omega_0\sqrt{1 - 2\zeta^2}$.

Tidsforsinkelser: En ren tidsforsinkelse har overføringsfunksjon $G(s) = e^{-\theta s}$, slik at:

$$G(\omega j) = e^{-j\theta\omega} = \cos(\omega\theta) - j \sin(\omega\theta).$$

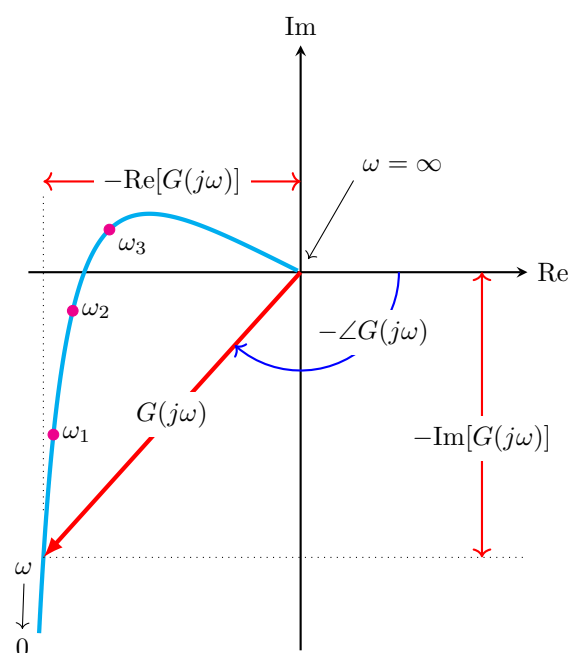
Dermed er $A_r(\omega) = \sqrt{\cos^2(\omega\theta) + \sin^2(\omega\theta)} = 1$ og $\phi(\omega) = \angle e^{-j\theta\omega} = \tan^{-1}(-\sin(\omega\theta)/\cos(\omega\theta)) = -\omega\theta$.

11.2. Grafisk analyse: Nyquist-, Bode- og Nichols-diagram

Alternative kilder: [Brian Douglas video](#) (QAfk8TuOM68).

11.2.1 Nyquist-diagram [ZUjQM1Xfy5s&t=826](#)

MATLAB-kommando: `nyquist` eller `nyquistplot`; se også [nyqlog](#) for dB-skalering.



Figur 11.4: Illustrasjon av et Nyquist-plott til en overføringsfunksjon $G(\cdot)$. Legg merke til negative fortegn foran blant annet $\angle G(j\omega)$ siden vi definerer denne som positiv målt mot klokken fra den horisontale reelle aksens.

Et Nyquist-diagram (også kalt et polar-plott) slik som illustrert i figur 11.4, er et konturplott hvor konturen i s -planet går fra 0 oppover den imaginære aksens til $\lim_{\omega \rightarrow \infty} \omega j$, og derfra, i en uendelig stor bue om $+\infty$ på den reelle aksens til $-\infty j$ på den imaginære, og derfra tilbake til 0 langs den imaginære aksens. Den fulle prosedyren for å lage et slikt plott er gitt under:

Hvordan lage et Nyquist diagram for en overføringsfunksjon $G(s)$:

Mål: Lage en kontur som omkranser hele høyre halvplan med klokken.

Steg 1. Vi starter i origo $(0, j \cdot 0)$, går opp den imaginære aksens $(0, j \cdot \omega)$ til $\omega = +\infty$;

Steg 2. Tar rask liten rundtur om $(+\infty, j \cdot 0)$ til $(0, -j \cdot \infty)$ og derfra opp til origo langs den imaginære aksens igjen.

Steg 3. Plotter verdien til $G(j \cdot \omega)$ i det komplekse plan langs denne konturen, feks. vha. **Amplituderatio** $A_r(\omega)$ og **Fasevinkel** $\phi(\omega)$ som **polarkoordinater**.

Merk følgende:

1. Siden verdien til $G(s)$ er speilet om den reelle (horisontale) akse (se § 11.1.1), så trenger vi bare regne ut $G(j\omega)$ og ta $G(-j\omega) = -G(j\omega)$.
2. Hvis $G(s)$ er strengt proper, så har vi $G(j\omega) \rightarrow 0$ når $\omega \rightarrow +\infty$.
3. Retningen (med- eller mot klokken) $G(j\omega)$ lager i det komplekse plan når vi går langs konturen kan være viktig (spesielt for Nyquists stabilitetskriterium).

Noen fordeler og ulemper med Nyquist-diagrammer følger:

Fordeler:

- Kompakt representasjon av frekvensresponsen;
- Allsidig og generaliserbar (kan også brukes på ustabile systemer);
- Bedre bilde av robusthet enn fra et Bode-plott (mer om det senere).

Ulemper:

- Alt på et plott, så vanskelig å se sammenheng mellom respons og frekvens;

11.2.2 Bode-diagram [ZUjQM1Xfy5s&t=1717](https://www.youtube.com/watch?v=ZUjQM1Xfy5s&t=1717)

MATLAB-kommando: bode eller bodeplot.

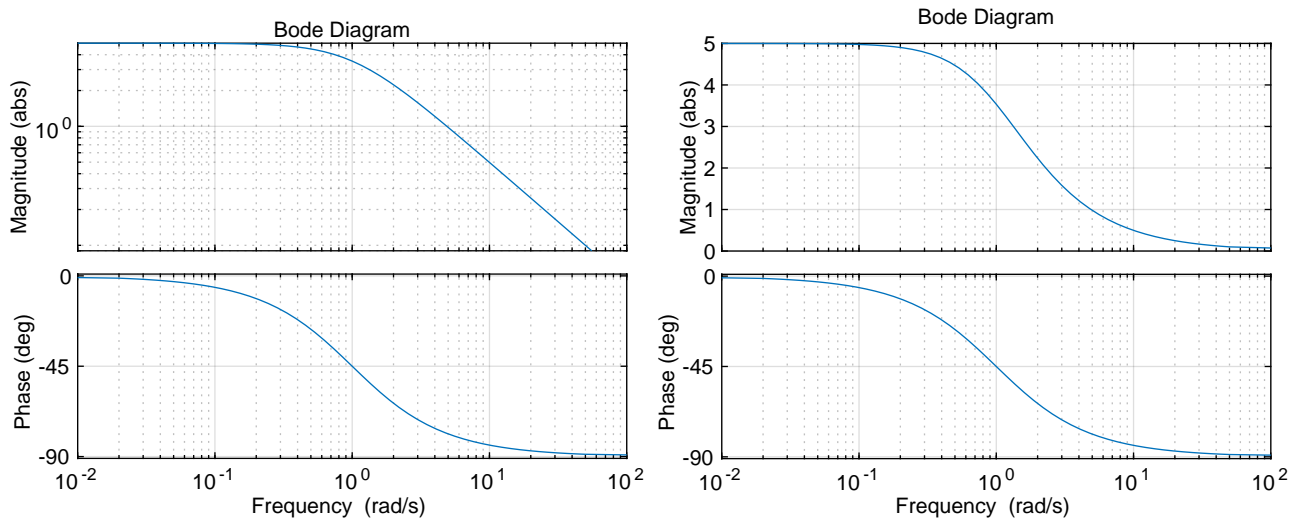
Bode-plotts er du nok godt kjent med allerede. Her har man et eget plott for både magnituden/amplituderatioen, $|G(j\omega)|$ (på en desibel-skala), og fasevinkelen, $\angle G(j\omega)$, med frekvensen, ω (i radianer per sekund), representert logaritmisk på den horisontale akse.

Skal vi bruke desibel? Magnitude-plottet er ofte (f.eks. i MATLABs **bode**-kommando) representert i enheten desibel (dB), altså $A_r^{dB} = 20 \log A_r(\omega)$. Vi skal dog normalt sett **ikke bruke desibel**³ i disse notatene, men i stedet vise magnituden på en logaritmisk-skalert akse (base 10). I MATLAB kan du enkelt kovertere mellom disse vha. kommandoene **mag2db** og **db2mag**

I MATLAB kan man gi en ekstra innstilling til **bode**-kommandoen for å vise absolutt amplitude i stedet for desibel, både med og uten et logaritmisk y-akse-skalering. Et slikt plott er vist i figur 11.5. Se kodesnutt 11.1 for hvordan du kan generere denne typen bode-plott.⁴

³Grunnen til at vi ikke skal bruke desibel handler både om at dere hatt det før, samt er dette basert på min (og andres) mening at desibel ikke akkurat er veldig intuitivt (mennesker sliter som regel med å få en intuitiv forståelse av eksponentielle og logaritmiske representasjoner). Dermed er det ikke ideelt å ta det i bruk i en pedagogisk sammenheng. På den annen side, så brukes desibel veldig ofte innen mange felt (også reguleringsteknikk), så man må også mestre dette og kunne konvertere ved behov.

⁴Hvis du også ønsker et Bode-plott med asymptoter (med med desibel-skala), kan du se se: <https://se.mathworks.com/matlabcentral/fileexchange/10183-bode-plot-with-asymptotes>.



Figur 11.5: Bode-diagram av første-ordens system. Venstre: med log-skala; høyre: uten.

Kodesnutt 11.1: Bode-plott av første-ordens system i MATLAB uten amplitude gitt i decibel.

```
s = tf('s');
G = 5/(1+s);
plotoptions = bodeoptions;
plotoptions.MagUnits='abs'; % fra decibel til absoluttverdi
plotoptions.MagScale='log'; % amplituden paa en log-skala
plotoptions.Grid = 'on';
bode(G, plotoptions)
```

Noen fordeler og ulemper med Bode-plotts følger:

Fordeler:

- Visuelt enkelt å se sammenheng mellom frekvens og både amplituden og fasen;
- Med logaritimisk skala (med eller uten desibel) kan man enkelt summere bidragene til magituden ved forskjellige frekvenser, og skissere asymptote-plotts;
- Enkelt å lese av marginer og kryssfrekvenser.

Ulemper:

- Vanskelig å se sammenhengen mellom fase- og forsterkningsmarginer i forhold til robusthet;
- Begrenset nytteverdi for stabilitetsundersøkelser ved ustabile systemer og hvis det er flere kryssfrekvenser.

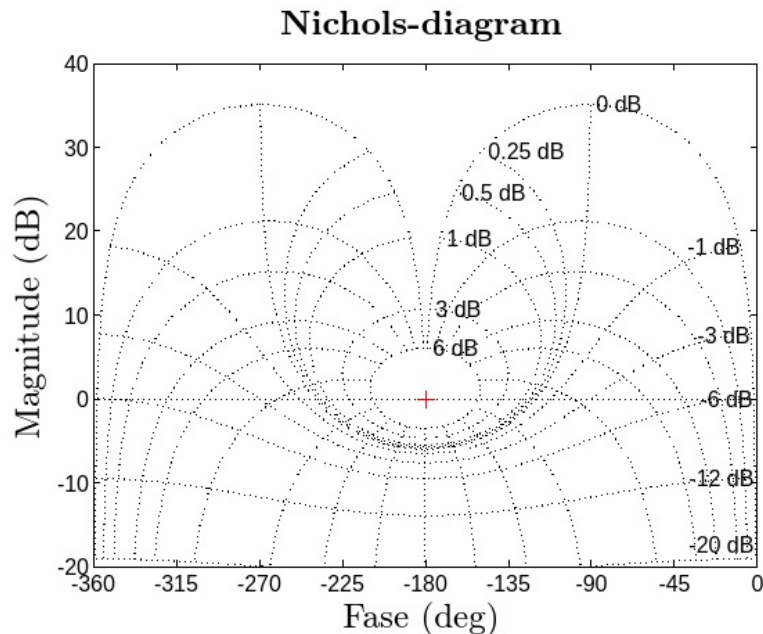
11.2.3 Nichols-diagram



MATLAB-kommando: nichols eller nicholsplot.

Nichols-plott er en annen grafisk metode for å studere frekvensresponsen til et system. Her plotter man fasen langs den horisontale asken, og magnituden (ofte i desibel) langs den vertikale

aksen. Et slikt diagram er vist i figur 11.6. Ulikt Bode-plottet, så lar dessverre ikke MATLAB en bytte fra desibel-skala til kun en logaritmisk-skala for Nichols-plottet (man kan selvsagt lage sin egen variant hvis man vil).



Figur 11.6: Et MATLAB-generert Nichols-diagram med koter som tilsvarer den lukkede sløyfen.

Tanken bak Nichols-diagrammet er at man kan finne både magnituden/amplituderatioen og fasen til den lukkede sløyfen, $G_{LS}(s)$, ved en gitt frekvens, fra den åpne sløyfen, $G_{AS}(s)$, sin magnitude og fase ved samme frekvens siden

$$G_{LS}(s) = \frac{G_{AS}(s)}{1 + G_{AS}(s)}.$$

Mer spesifikt, så har vi

$$A_r^{LS} = A_r^{AS} / \left| 1 + A_r^{AS} e^{j\phi^{AS}} \right|$$

og

$$\phi^{LS} = \phi^{AS} - \angle(1 + A_r^{AS} e^{j\phi^{AS}}).$$

Dette gjør at vi kan tegne opp sammenhengen mellom (A_r^{AS}, ϕ^{AS}) og (A_r^{LS}, ϕ^{LS}) som koter i forkant i diagrammet, slik som vist i figur 11.6.

Noen fordeler og ulemper med Nichols-diagrammer følger:

Fordeler:

- Alt på et plott;
- Effekt av endret forsterkning lett synlig;
- Enkelt å se på flere robusthets-marginer;
- Kan brukes for åpent-ustabile systemet (vha. Nyquist-kriteriet).

Ulemper:

- Frekvensinformasjon ikke direkte synlig;
- Kan ta litt tid å bli vant med.

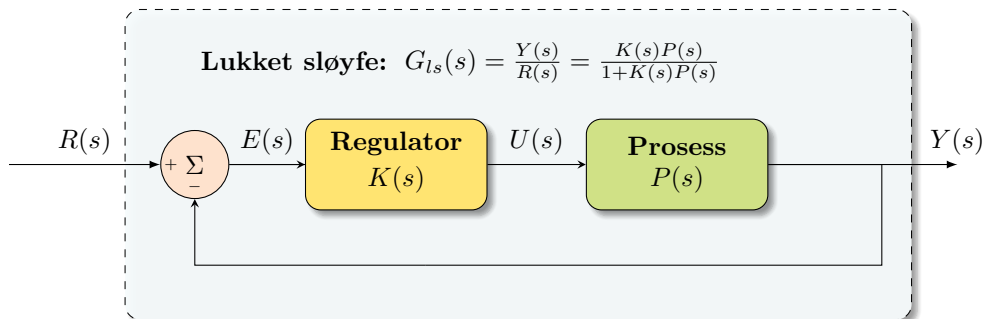
11.3. Lukket-sløyfe stabilitet ▶ ZUjQM1Xfy5s&t=2726

Nåværende problem: Gitt en (marginalt/kritisk) stabil overføringsfunksjon $G_{\text{ÅS}}(s)$, når er den den lukkede-sløyfen $G_{\text{LS}}(s) = \frac{G_{\text{ÅS}}(s)}{1+G_{\text{ÅS}}(s)}$ stabil?

Husk: et lineært system er marginalt/kritisk stabilt hvis dets overføringsfunksjon har én pol i origo og/eller flere forskjellige komplekskonjugerte pol-par på den imaginære akse, mens eventuelle resterende poler er i venstre halvplan.^a

^aEn annen måte å si dette på er at alle poler som ikke er i venstre halvplan er *enkle* (det er bare én slik pol) og ligger på den imaginære akse; f.eks. så har $\frac{1}{(s+2)(s+1)}$ to enkle poler, $s = -1$ og $s = -2$, mens $\frac{1}{(s+1)^2(s^2+1)}$ har de enkle polene $s = \pm\sqrt{-1}$ og en dobbel pol $s = -1$.

11.3.1 Det kritiske punktet ▶ ZUjQM1Xfy5s&t=2836

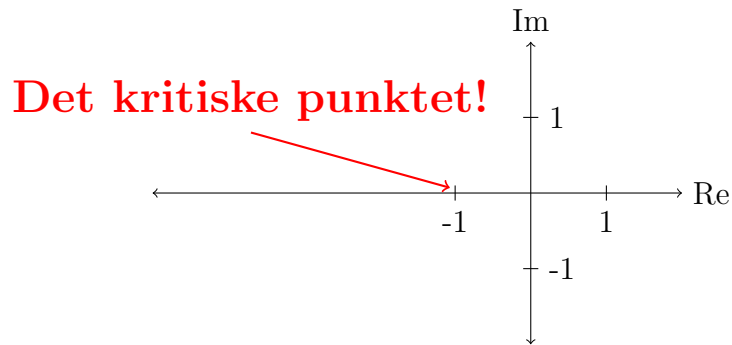


Figur 11.7: Enkel lukket reguleringsløyfe.

Gitt systemet i figur 11.7, hvor: $E(s) = \mathcal{L}\{e(t)\}$, $R(s) = \mathcal{L}\{r(t)\}$ og $Y(s) = \mathcal{L}\{Y(t)\}$, etc. Overføringsfunksjonen til den lukkede sløyfen er dermed

$$G_{\text{LS}}(s) = \frac{K(s)P(s)}{1 + K(s)P(s)}. \quad \text{(Lukket sløyfe)}$$

Legg merke til at nevneren blir null hvis $G_{\text{ÅS}}(s) = K(s)P(s) = -1$. Punktet $(-1, 0)$ i det komplekse plan kalles derfor for det **kritiske punktet** (se figur 11.8), og er av stor betydning når det kommer til å studere stabiliteten til den lukkede sløyfen fra frekvensresponsen til den lukkede sløyfen. Grunnen til dette er løst forklar som følger: Hvis $G_{\text{ÅS}}(j\omega) = -1$ for én frekvens, si ω_k , så har den lukkede sløyfen én pol på den imaginære akse, og er dermed på stabilitetsgrensen.



Figur 11.8: Det kritiske-punktet i $G_{\text{ÅS}}(s) = K(s)P(s)$ -planet. Hvis $G_{\text{ÅS}}(j\omega)$ går igjennom dette punktet, så er den lukkedesløyfen, $G_{\text{LS}}(s)$, enten stabil eller er på stabilitetsgrensen.

Eksempel 11.1. (Eks. 7-20 i [Ogata et al., 2010]) Gitt et system med den åpne sløyfen gitt ved

$$G_{\text{ÅS}}(s) = \frac{K}{s(s+1)(s+5)}.$$

Den lukkede sløyfen er Stabil for $K = 10$, men ikke for $K = 100$. Vi kan konkludere med dette ved å utlede at polene til den lukkede sløyfen $G_{\text{LS}}(s) = \frac{G_{\text{ÅS}}(s)}{1+G_{\text{ÅS}}(s)}$ tilsvarer røttene til tredjeordens polynomet $s^3 + 6s^2 + 5s + K$. Du kan løse dette med for eksempel WolframAlpha for å finne ut at $K = 10$ gir alle polene i venstre halvplan mens $K = 100$ fører et komplekskonjugert polpar i høyre halvplan.

11.3.2 Bodes stabilitetskriterium

▶ ZUjQM1Xfy5s&t=2982

For Bodes stabilitetskriterium trenger vi å introdusere kryssfrekvenser.

Kryssfrekvenser: Gitt et overføringsfunksjon $G(s)$ med [Amplituderatio](#) og [Fasevinkel](#).

Amplitudekryssfrekvensen: Anta at det finnes én frekvens, ω_A , slik at

$$\omega_A \implies A_r(\omega_A) = 1. \quad (\text{Amplitudekryssfrekvens})$$

Vi kaller denne frekvensen ω_A for *amplitudekryssfrekvensen*.

Fasekryssfrekvensen: Anta at det finnes én frekvens, ω_{180} , slik at

$$\omega_{180} \implies \phi(\omega_{180}) = -\pi = -180^\circ. \quad (\text{Fasekryssfrekvens})$$

Vi kaller denne frekvensen ω_{180} for *fasekryssfrekvensen*.

Bodes stabilitetskriterium: Gitt en strengt proper åpen-sløyfe overføringsfunksjon $G_{\text{ÅS}}(s)$ som har alle polene i venstre halvplan, med mulig unntak av én enkelt pol i origo. Anta videre at $G_{\text{ÅS}}(s)$ har bare én amplitudekryssfrekvens, ω_A , og én fasekryssfrekvens, ω_{180} .

Den lukkede sløyfen,

$$G_{LS}(s) = \frac{G_{\text{AS}}(s)}{1 + G_{\text{AS}}(s)},$$

er da stabil hvis **Amplituderatio** til $G_{\text{AS}}(s)$ tilfredsstillers $A_r(\omega_{180}) < 1$ (evt. $\omega_{180} > \omega_A$), hvor ω_{180} og ω_A er henholdsvis **Fasekryssfrekvens** og **Amplitudekryssfrekvens**, og ustabil ellers.

La oss nevne noen fordeler og ulemper med dette stabilitetskriteriet:

Fordeler:

- Stabilitet bestemmes fra den åpne sløyfen
- Kan brukes også for systemer med tidsforsinkelser (i motsetning til f.eks. Routh-Hurwitz-kriteriet);

Ulemper:

- Kan ikke direkte brukes på åpent ustabile systemer eller systemer med flere kryssfrekvenser.

Ulempene med Bodes stabilitetskriterium gjelder dog ikke storebror, nemlig Nyquist sitt stabilitetskriterium.

11.3.3 Nyquists stabilitetskriterium ▶ ZUjQM1Xfy5s&t=3229

Alternative kilder: [Brian Douglas video](#) (sof3meN96MA)

La oss begynne med et eksempel som viser begrensningene med Bodes kriterium:

Eksempel 11.2. La den åpne sløyfen være gitt ved

$$G_{\text{AS}}(s) = \frac{2}{s - 1}.$$

Denne har en pol i høyre halvplan, og er dermed ustabil. Vi har at

$$1 + G_{\text{AS}}(s) = 1 + \frac{2}{s - 1} = \frac{s - 1 + 2}{s - 1} = \frac{s + 1}{s - 1}.$$

Den lukkede sløyfen, gitt ved

$$G_{LS}(s) = \frac{G_{\text{AS}}(s)}{1 + G_{\text{AS}}(s)} = \frac{2}{s + 1},$$

har derfor en pol i venstre halvplan og er dermed stabil.

Eksempelet over viser at et åpent ustabil system kan bli gjort stabilt ved tilbakekobling. Denne konklusjonen kunne vi ikke fått fra Bode sitt stabilitetskriterium, som krever at den åpne sløyfen er (marginalt) stabil.

Nyquist sitt stabilitets kriterium (som er storebror til Bode-kriteriet) lar en sjekke stabiliteten til den lukket sløyfe ved å sjekke frekvensresponsen til den åpne sløyfen selv om denne er ustabil:

Nyquists stabilitetskriterium: Gitt en åpen-sløyfe overføringsfunksjon $G_{\text{ÅS}}(s)$ uten noen ustabile pol-nullpunkt-kanselleringer og med N_p poler i høyre halvplan.

La N_o være antall ganger Nyquist-kurven^a (se § 11.2.1) omkranser det kritiske punktet $(-1, 0 \cdot j)$ med klokken, slik at f.eks. $N_o = -1$ ved én mot-klokken omkransing av $(-1, 0 \cdot j)$.

Den lukkede sløyfen, $G_{\text{LS}}(s) = \frac{G_{\text{ÅS}}(s)}{1+G_{\text{ÅS}}(s)}$, er da stabil hvis, og bare hvis, $N_o + N_p = 0$.

^aVed poler på den imaginære akse kan man bruke visse trikserier som vi ikke skal se på i disse notatene.

En forenklet variant for åpent stabile systemer er som følger:

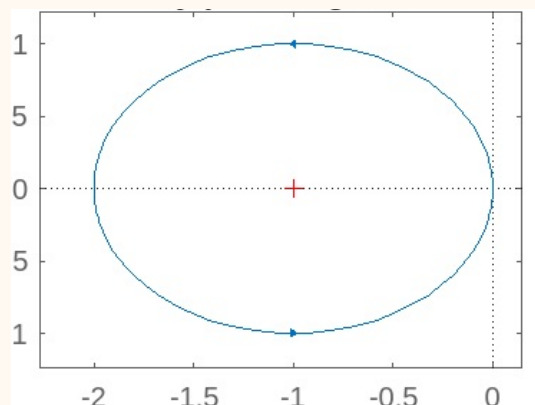
Forenkling av Nyquists stabilitetskriterium: For at et system $G_{\text{ÅS}}(s)$ som er åpent stabilt skal forbli stabilt når det lukkes, må kurven som er formet av $1 + G_{\text{ÅS}}(s)$ når s gjennomløper en kontur rundt hele høyre halvplan med urviseren få null vinkeldreining. Dette er det samme som å kreve at punktet $(-1, 0)$ ligger på utsida av kurven til $G_{\text{ÅS}}(s)$ i Nyquist-diagrammet.

Tre viktig scenarier:

1. Det er ikke noen omringning av $-1 + j0$: Dette betyr at stabiliteten til $G_{\text{LS}}(s)$ tilsvarer stabiliteten til $G_{\text{ÅS}}(s)$.
2. Det er én eller mer omringninger av $-1 + j0$ mot klokka: Dette betyr at G_{LS} er stabilt hvis antall omringninger tilsvarer antallet av polene til $G_{\text{ÅS}}(s)$ som ligger i høyre halvplan; ellers er systemet ustabil.
3. Det er én eller mer omringninger $-1 + j0$ med klokka: Dette betyr at systemet er ustabil.

Eksempel 11.3. (Forst. av eks. 11.2)

Vi så tidligere at selv om den åpne sløyfen $G_{\text{ÅS}}(s) = 2/(s - 1)$ var ustabil på grunn av en pol i høyre halvplan, slik at $N_p = 1$, så ble den lukkede sløyfen stabil: $G_{\text{LS}}(s) = 2/(s + 1)$. Dette kunne vi ikke se fra et Bode plott. Fra Nyquist-diagrammet til høyre kan vi derimot se at vi har én omkransing av det kritiske punktet mot klokka, slik at $N_o = -1$. Dermed er $N_o + N_p = 0$, og vi kan konkludere fra Nyquistkriteriet at den lukkede sløyfen er stabil.



11.3.4 Lukket-sløyfe-stabilitet fra grafiske betraktninger

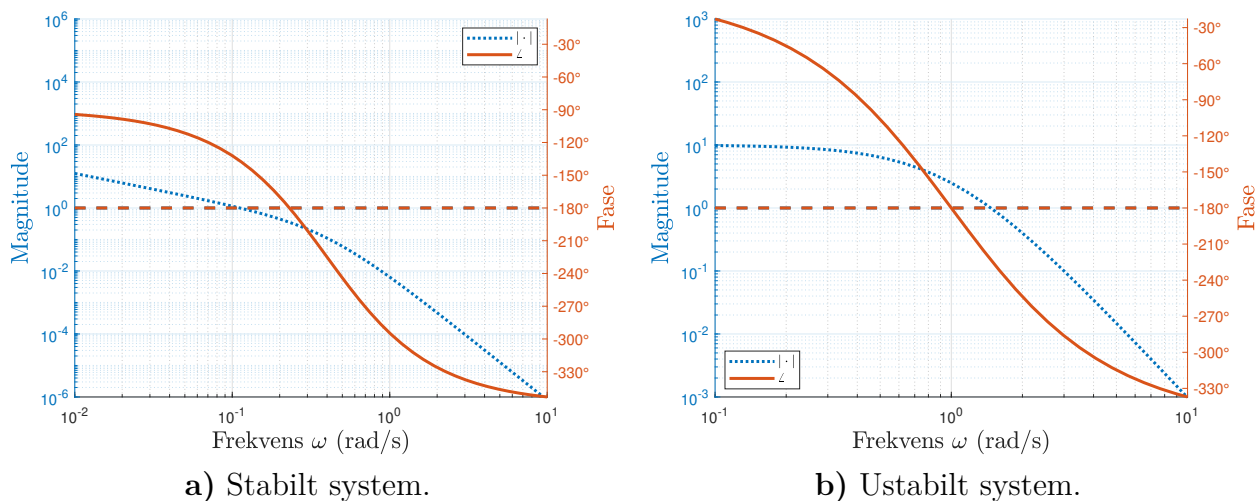
Vi skal nå se på hvordan vi kan se fra frekvensresponsen til den åpne sløyfen om den lukkede sløyfen er stabilt, ved hjelp av de tre grafiske metodene vi har sett på: Bode-, Nyquist- og Nichols-diagrammer. Vi skal hovedsakelig kun ta hensyn åpent stabile systemer ($G_{\text{ÅS}}(s)$ er (marginalt) stabil). Vi vil for enkelthets skyld også se på systemer hvor vi kan ta i bruk Bode

sitt stabilitetskriterium ved at det ikke er flere kryssfrekvenser, men vil også nevne hvilke av metodene som også lar en ta i bruk Nyquistkriteriet.

Bode-diagram [▶ ZUjQM1Xfy5s&t=3483](#)

⚠ Husk: Bode-diagrammer er ikke egnet til å sjekke stabiliteten til en lukket sløyfe hvis den åpne sløyfen er ustabil (har poler i høyre halvplan)!

Figur 11.9 gir et eksempel på et stabilt og et ustabil system i et Bode-diagram. For det stabile systemet ser vi at $\omega_A \approx 0.1 < \omega_{180} \approx 0.25$.



Figur 11.9: Illustrasjon av Bode-diagrammene til et stabilt system ($\omega_A < \omega_{180}$) til venstre og et ustabil system ($\omega_A > \omega_{180}$) til høyre.

Nyquist-diagram [▶ ZUjQM1Xfy5s&t=3703](#)

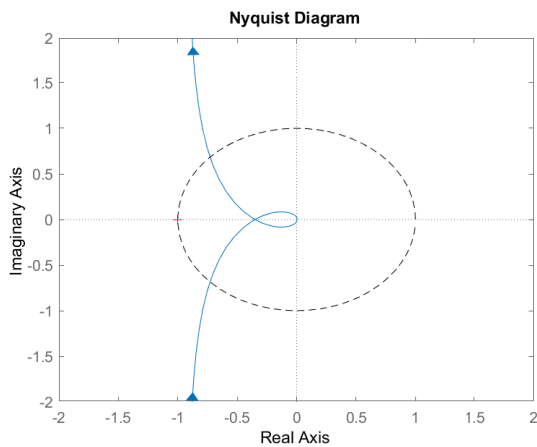
Figur 11.10 gir et eksempel på et stabilt og et ustabil system i et Nichols-diagram.

Noen nyttige kjennetegn: Stabile, første-ordens systemer (uten tidsforsinkelse) krysser aldri den vertikale (imaginære) akse, og forblir dermed i høyre-halvplan.

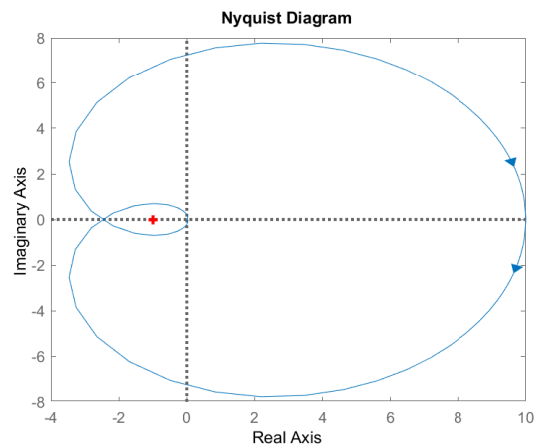
Stabile, andre-ordens systemer krysser den vertikale akse, men delene av kurven som entrer det venstre halvplanet konvergerer alltid til origo uten å krysse den horisontale (reelle) akse.

Nichols-diagram [▶ ZUjQM1Xfy5s&t=3872](#)

Figur 11.11 gir et eksempel på et stabilt og et ustabil system i et Nichols-diagram. I forhold til Nyquist sitt stabilitetskriterium, så teller en kryssing av -180 -aksen *over* 0 dB-punktet som en med-klokken omkransing hvis den går fra venstre mot høyre (ved økende frekvens).

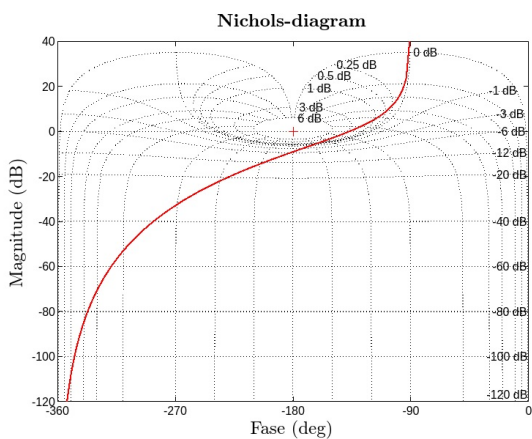


a) Stabilt system.

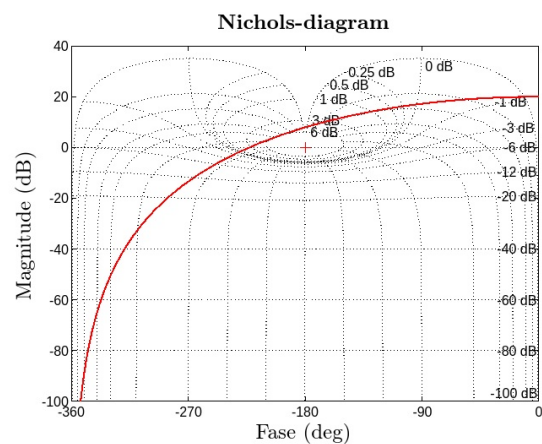


b) Ustabilt system.

Figur 11.10: Illustrasjon av Nyquist-diagrammene til et stabilt system til venstre (omkranser ikke det kritiske punktet) og et ustabilt system til høyre (åpen sløyfe er stabil, men omkranser det kritiske punktet to ganger med klokken).



a) Stabilt system.



b) Ustabilt system.

Figur 11.11: Illustrasjon av Nichols-diagrammene til et stabilt system til venstre (passerer under det kritiske punktet) og et ustabilt system til høyre (passerer over det kritiske punktet).

11.4. Stabilitetsmarginer og sensitivitetsanalyse

Alternative kilder: [Fint MATLAB-eksempel](#); [Brian Douglas' video om robust regulering \(A7wHSr6GRnc\)](#).

I figurene 11.9, 11.10 og 11.11 viste vi eksempler på hvordan man kunne bekrefte en prosess sin stabilitet eller ustabilitet ved hjelp av Bode-, Nyquist- og Nichols-diagrammer (spoiler alert: alle plottene var laget fra de samme prosessene). For de stabile prosessen er det også interessant

å spørre: Hvor stabile er de? Det vil si, hvor mye skal til for «ødelegge» stabiliteten i form av for eksempel usikkerhet eller endringer i prosessmodellen vår. Her kommer (stabilitets-)marginer inn i bildet.

11.4.1 Fase- og forsterknings-marginer

▶ ZUjQM1Xfy5s&t=3906

Alternative kilder: §14.7 i [Seborg et al., 2016].

Nåværende problem: Generelt betyr ufullkommen prosess-modellering at både forsterkning og fase ikke er kjent nøyaktig. Fordi modelleringsfeil er mest skadelig nær amplitudekryssfrekvensen (frekvens der åpen-sløyfeforsterkning er lik én), har det også betydning hvor mye fasevariasjon som kan tolereres ved denne frekvensen.

Fase- og forsterkningsmargin er to velkjente marginer som lett kan leses av fra Nyquist-, Bode- og Nichols-diagrammene til den åpne sløyfen til et LTI system. Sammen gir disse to tallene et visst estimat på «sikkerhetsmarginen» for stabilitet i lukket sløyfe: Jo mindre stabilitetsmarginene er, desto skjørere er stabiliteten. På den annen side kan veldig store marginer tilsvare dårlig ytelse, i form av treg respons, siden den lukkede sløyfen blir i overkant konservativ.

Vi antar nå at systemet tilfredsstiller Bodes stabilitetskriterium (se 11.3.2).

Forsterkningsmargin: Hvor mye kan forsterkningen øke (uten endringer i fasen) før systemet blir ustabil. Denne størrelsen, med symbol GM for «gain margin», er gitt ved

$$GM = \frac{1}{A_r(\omega_{180})} \quad (\text{Forsterkningsmargin})$$

hvor $A_r(\cdot)$ er Amplituderatioen og ω_{180} er Fasekryssfrekvensen.⁵

Merk: Forsterkningsmarginen til et første- eller andreordens stabilt system er uendelig.

Tommelfingerregel: Normalt ønsker man $1.5 \leq GM \leq 4$, og $GM \geq 2$ (≈ 6 dB) er vanlig.

Fasemargin: Hvor mye ekstra fase-etterslep (uten endring i forsterkningen) man kan ha før systemet blir ustabil. Denne størrelsen, med symbol PM for «phase margin», er gitt ved

$$PM = 180^\circ + \phi(\omega_A) \quad (\text{Fasemargin})$$

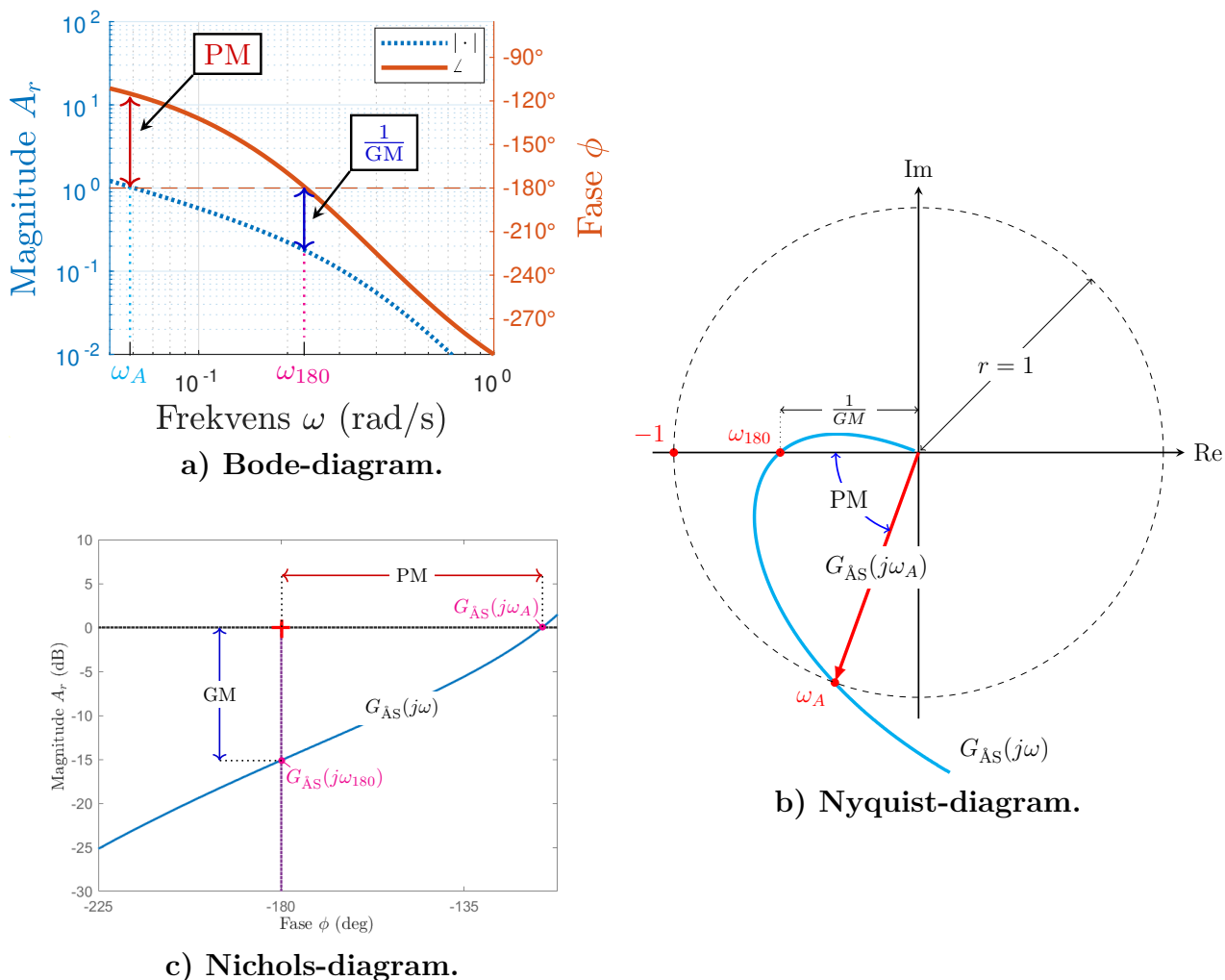
hvor $\phi(\cdot)$ er Fasevinkelen (i grader) og ω_A er Amplitudekryssfrekvensen.

Tommelfingerregel: Normalt ønsker man $30^\circ \leq PM \leq 45^\circ$.

Forsinkelsesmargin: Sterkt relatert til fasemarginen er (tids-)forsinkelsesmarginen, som sier hvor mye ren tidsforsinkelse systemet tåler før det blir ustabil. Denne størrelsen er gitt ved

$$\Delta\theta_{\max} = \left(\frac{PM}{\omega_A} \right) \left(\frac{\pi}{180^\circ} \right). \quad (\text{Forsinkelsesmargin})$$

⁵Siden amplituderatioen i desibel er $A_r^{dB} = 20 \log_{10} A_r$, så er forsterkningsmarginen i desibel $GM^{dB} = -A_r^{dB}(\omega_{180})$. Vi kan derfor direkte lese av GM^{dB} fra Bode- og Nichols-diagrammer med desibelskala.



Figur 11.12: Illustrasjon av forsterknings- og fasemargin i Bode-, Nyquist - og Nichols-diagrammer. Bode- og Nichols-plottene er av samme prosess, mens Nyquist-plottet tilsvarer en annen prosess. Merk at fasen er målt relativt til den venstre vertikale aksens i Bode-diagrammet, samt at magnituden i Nichols-diagrammet er gitt i desibel.

MATLAB-kommandoer: margin (grafisk i Bode-plott) og allmargin .

Er fase- og forsterkningsmarginen en god indikasjon på robusthet?

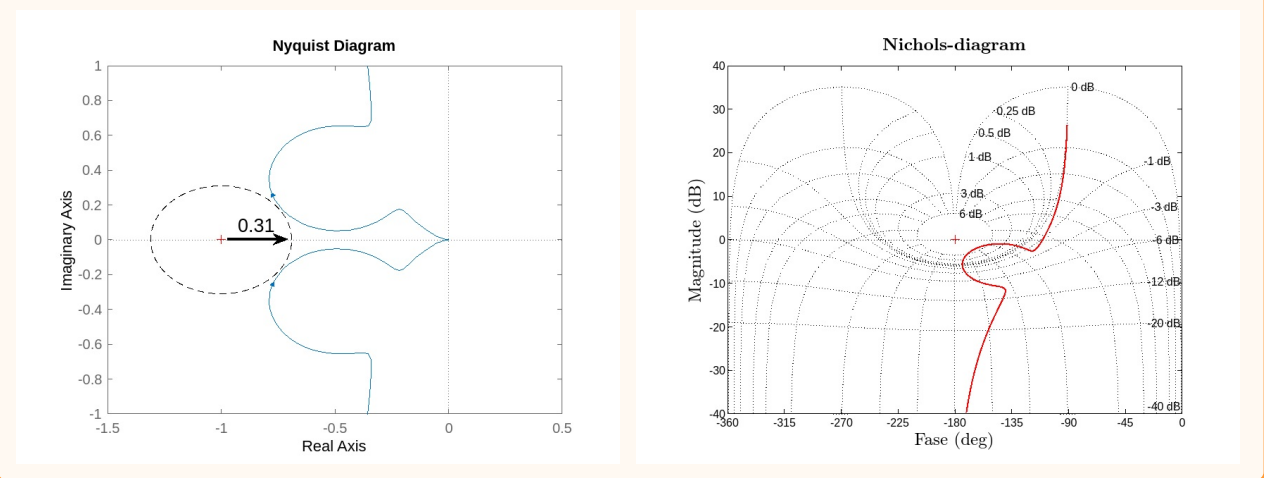
Figur 11.12 b) viser hvordan man kan lese av forsterkningsmarginen, GM , og fasemarginen, PM , fra et Nyquist-plott. Man kan også se fra figuren at Nyquist-kurven er nærmest det kritiske punktet et annet sted enn hvor vi måler disse marginene. Dette demonstrerer visse begrensninger disse marginene har som et mål på den lukkede sløyfens robusthet i forhold til usikkerhet. Følgende eksempel demonstrerer dette enda tydeligere:

Eksempel 11.4. Eks. fra [Åström and Murray, 2021], se også [▶ A7wHSr6GRnc?t=420](#) Gitt

følgende overføringsfunksjon for den åpne sløyfen:

$$G_{\text{ÅS}}(s) = \frac{0.38(s^2 + 0.1s + 0.55)}{s(s + 1)(s^2 + 0.065s + 0.5)}.$$

Ved hjelp av f.eks. `allmargin` i MATLAB kan man finne at man har uendelig forsterkningsmargin (fasen blir aldri mindre enn -180°), mens fasemarginen er ca. 68.43°). Tross disse marginene, så er allikevel den lukkede sløyfen meget sårbar til en kombinasjon av endret fase og forsterkning. Dette kan man se fra Nyquist-plottet under, hvor man ser at Nyquist-kurven kommer svært nært det kritiske punktet. Noe tilsvarende kan man også lett lese av fra Nichols-diagrammet. Faktisk er man så nært dette punktet, at den lukkede sløyfen til $\sqrt{2}G_{\text{ÅS}}(s)e^{-0.2s}$ er ustabil!



Diskmargin*

Et alternativ til fase- og forsterkningsmargin er **diskmargin**, hvor man i stedet ser på både fase- og forsterkningsendringer samtidig. **MATLAB-kommando:** `discmarginplot`. Se f.eks. følgende video for mer om dette: [▶ XazdN6eZF80](#) .

11.4.2 Sensitivitets-analyse [▶ ZUjQM1Xfy5s&t=4388](#)

Alternative kilder: Vedlegg J i [Seborg et al., 2016] (finnes på nett); §5.2 i [Skogestad and Postlethwaite, 2007]; Brian Douglas video (BAWdZvF1O40).

Overføringsfunksjonen

$$S(s) := \frac{1}{(1 + K(s)P(s))} \quad (\text{Sensitivitetsfunksjon})$$

kalles for *sensitivitetsfunksjonen* (eller *Avviksforholdet* [Balchen et al., 2016]), og tilsvarende $Y_{ls}(s)/Y_{os}(s)$ (altså forholdet mellom lukket- og åpen-sløyfe responsene); mens

$$T(s) := 1 - S(s) = \frac{K(s)P(s)}{(1 + K(s)P(s))} \equiv G_{\text{LS}}(s), \quad (\text{Følgeforsholdet})$$

kalles for *følgeforholdet* eller den *komplementære sensitivitetsfunksjonen*.

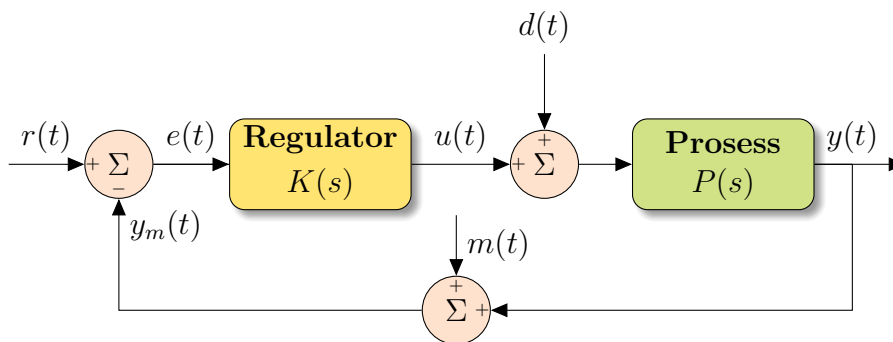
Disse overføringsfunksjonene gir et mål på følsomheten til den lukkede sløyfen i forhold til endringer i prosessen. Vi ønsker generelt sett at

- $S(s)$ er så *liten* som mulig for å maksimere effekten av tilbakekobling (små avvik for sprang i referanse og/eller forstyrrelser);
- $T(s)$ er så *liten* som mulig for å minere effekt av målestøy pga. tilbakekoblingen.

Men: $S(s) + T(s) = 1$, så man må komme frem til et visst kompromiss mellom disse. På grunn av den såkalte **“vannseng-effekten”**, så kan man heller ikke gjøre S liten over et bredt spekter av frekvensen. Dette følger Bodes integralformel, som for et åpent stabilt system sier at (se f.eks. [Åström and Murray, 2021] for mer detaljer rundt dette)

$$\int_0^\infty \ln |S(j\omega)| d\omega = 0.$$

Med andre ord, hvis vi senker verdien til $S(s)$ ved en frekvens, så må den øke ved en annen, slik som hvis du flytter deg rundt på en vannseng.



Figur 11.13: Reguleringsystem som overføringsfunksjoner i Laplace-domenet.

Hvor kommer disse overføringsfunksjonene fra fra? Gitt det monovariabelt systemet (altså med én utgang og ett pådrag, $y, u \in \mathbb{R}$) vist i figur 11.13, så får man

$$Y(s) = \frac{P(s)}{1 + K(s)P(s)} D(s) + \frac{K(s)P(s)}{1 + K(s)P(s)} (R(s) - M(s)), \tag{11.3}$$

hvor d er en forstyrrelse og m er målestøy. Dette tilsvarer

$$E(s) = \underbrace{\frac{1}{1 + K(s)P(s)}}_{S(s)} R(s) - \underbrace{\frac{P(s)}{1 + K(s)P(s)}}_{P(s)S(s)} D(s) + \underbrace{\frac{K(s)P(s)}{1 + K(s)P(s)}}_{T(s)} M(s)$$

hvor man naturlig nok ønsker at $e(t) \rightarrow 0$. Dermed vil $S(s)$ og $T(s)$, sammen med $\frac{P(s)}{1+K(s)P(s)}$ (som også er en viktig sensitivitetsfunksjon i seg selv!), gi alle overføringsfunksjonene foran $D(s)$, $M(s)$ og $R(s)$. Legg også merke til at $1 + K(s)P(s)$ er nevneren i alle leddene.

La oss nå anta at $r = 0$. I åpen sløyfe ($K = 0$), så har vi da at

$$Y_m^{\text{ÅS}} = M(s) + P(s)D(s),$$

mens for den lukkede sløyfen så har vi

$$Y_m^{\text{LS}} = \frac{1}{1 + K(s)P(s)} (M(s) + P(s)D(s)) = S(s)Y_m^{\text{AS}}.$$

Dette betyr at hvis $|S(s)| < 1$ så hjelper tilbakekoblingen, mens hvis $|S(s)| > 1$ så gjør vi responsen faktisk verre!

La oss nå også introdusere noen andre viktige størrelser relatert til systemets sensitivitet:

Den nominelle sensitivitetstoppen: Definer

$$M_S = \max_{0 \leq \omega < \infty} |S(j\omega)| = \max_{0 \leq \omega < \infty} \left| \frac{1}{1 + K(j\omega)P(j\omega)} \right|, \quad (\text{Sensitivitetstoppen})$$

altså den største absoluttverdien til sensitivitetsfunksjonen $S(j\omega)$ for alle positive frekvenser. Tallet M_S tilsvarende den inverse av den korteste avstanden fra det kritiske punktet, $(-1, 0)$, til Nyquist-konturen til den åpne sløyfens overføringsfunksjon, $G_{\text{AS}}(s) = K(s)P(s)$. Dermed blir det regulerede systemet desto mer robust i lukket sløyfe desto mindre M_S er. Tallet M_S er også relatert til **Forsterkningsmarginen**, GM , og **Fasemarginen**, PM , via følgende ulikheter (se [Skogestad and Postlethwaite, 2007, s. 36]):

$$GM \geq \frac{M_S}{M_S - 1} \quad \text{og} \quad PM \geq 2 \arcsin \left(\frac{1}{2M_S} \right) \geq \frac{1}{M_S} [\text{rad}].$$

Tommelfingerregel: Normal sett ønsker man $1.2 \leq M_S \leq 2.0$.

Resonanstoppen: Tilsvarende som for **Sensitivitetstoppen**, så definerer vi

$$M_T = \max_{0 \leq \omega < \infty} |T(j\omega)| = \max_{0 \leq \omega < \infty} \left| \frac{K(j\omega)P(j\omega)}{1 + K(j\omega)P(j\omega)} \right|. \quad (\text{Resonanstoppen})$$

Tallet M_T sier derfor hva som er den maksimale forsterkningen til den lukkede sløyfen. Tallet M_T er relatert til **Forsterkningsmarginen**, GM , og **Fasemarginen**, PM , via følgende ulikheter (se [Skogestad and Postlethwaite, 2007, s. 36]):

$$GM \geq 1 + \frac{1}{M_T} \quad \text{og} \quad PM \geq 2 \arcsin \left(\frac{1}{2M_T} \right) \geq \frac{1}{M_T} [\text{rad}].$$

Tommelfingerregel: Normal sett ønsker man $1.0 \leq M_T \leq 1.5$.

Kodesnutt 11.2: Eksempel: Hvordan regne ut den nominelle sensitivitetstoppen.

```
% Eksempel fra https://youtu.be/BAWdZvF1O40
s = tf('s');
P = 0.38/(s*(s+1));
K = (s^2+0.1*s+0.55)/(s^2+.06*s+0.5);
S = feedback(1,K*P)
[mag,~,w]=bodeone(S);
[Ms,i]=max(mag);
disp(strcat('M_s=', num2str(Ms), ' ved w_s=', num2str(w(i))))
```

Kriterier for den lukkede sløyfens ytelse

Hvis vi hovedsakelig er ute etter ytelse har vi fra [Seborg et al., 2016, J.5.3] følgende kriterier:

1. For å eliminere statisk avvik må $|T(j\omega)| \rightarrow 1$ når $\omega \rightarrow 0$;
2. $|T(j\omega)|$ bør være lik 1 opp til en så høy så frekvens som mulig. Denne sikrer en rask settpunktrespons (kortlevde transienter) til en ny referanse;
3. Som nevnt i en tidligere tommelfingerregel, bør M_T være slik at $1,0 < M_T < 1,5$;
4. Båndbredden ω_{bb} (se B.2.3) og frekvensen ω_T tilsvarende M_T bør være så store som mulig, siden store verdier tilsvarer raske responser for den lukkede sløyfen.

Merk dog at å tilfredsstille disse kriteriene krever vanligvis et kompromiss. For eksempel vil kravet om at $M_T < 1,5$ bety at regulatorforsterkningen ikke kan være for stor; men lavere regulatorforsterkning fører igjen til mindre verdier for ω_{bb} og ω_T .

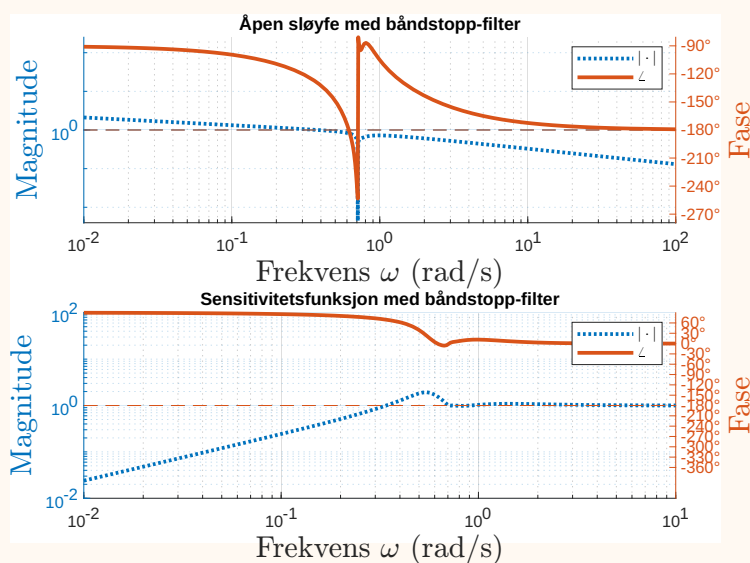
Eksempel 11.5. (Eksempel 11.4 revisited) Vi skal igjen se på prosessen hvor overføringsfunksjonen til den åpne sløyfen er

$$G_{AS}(s) = \frac{0.38(s^2 + 0.1s + 0.55)}{s(s + 1)(s^2 + 0.065s + 0.5)}$$

Som i videoen [▶ A7wHSr6GRnc?t=450](#) skal vi bruke et båndstopp-filter (eng. «notch-filter»), se § 12.3, for å redusere magnituden rundt sensitivitetstoppfrekvensen $\omega_s \approx 0.7 \text{ rad s}^{-1}$. Vi bruker derfor et [Båndstoppfilter](#) med Q -faktor 2, hakkdybde lik 1 og hakkfrekvens lik 0.7 rad s^{-1} :

$$N(s) = \frac{s^2 + 0.7^2}{s^2 + \frac{0.7}{2}s + 0.7^2}$$

Dette gir følgende forbedrede sensitivitet:



Del V

Filtere, signaltilpassing og sampling

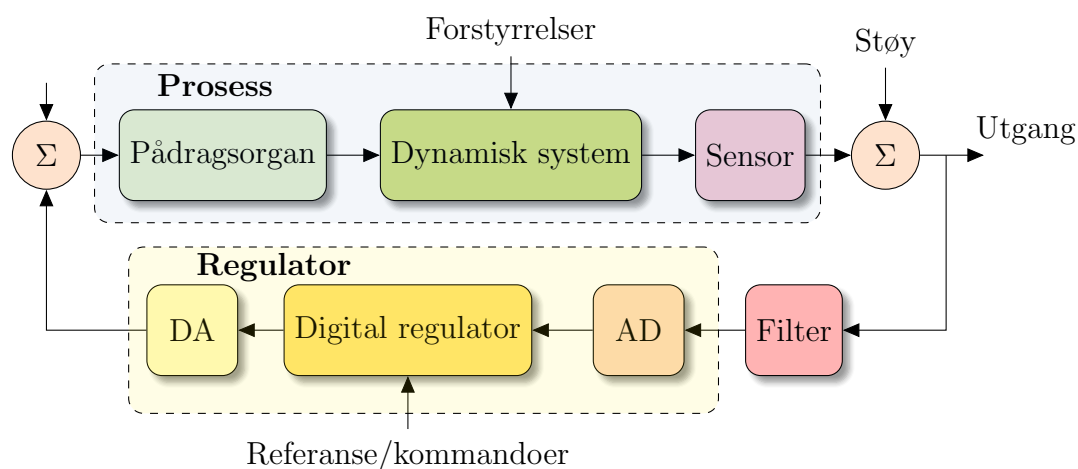
12. Filtre og digitale reguleringsystemer

De fleste regulatorer er i dag implementert digitalt i en eller annen form for datamaskin (f.eks. en [PLS](#) eller [mikrokontroller](#)). I motsetning til de kontinuerlige (*analoge*) signalene vi har jobbet med til nå, må vi derfor også ta hensyn til effekten av at de skal bli omgjort til *digitale* signaler i regulatoren. Vi skal derfor i dette kapittelet se på *digital regulering*, som omhandler bruk av digitale signaler til å regulere (analoge) dynamiske systemer. Her spiller også *filtering* av signalene en viktig rolle; filtrering fokuserer blant annet på å forbedre kvaliteten på signaler ved å redusere uønsket støy og forstyrrelser, samt fjerne uønskede effekter som ned-folding/aliasing.

12.1. Digitale reguleringsystemer

▶ [yM4z44-vHT4&t=21](#)

Alternative kilder: Kap. 17 i [Seborg et al., 2016].



Figur 12.1: Illustrasjon av et reguleringsystem med en digital regulator.

De fleste reguleringsystemer som blir implementert i dag er digitale (se fig. 12.1). Dette bringer med seg en rekke utfordringer og aspekter man må ta hensyn til, blant annet regulator diskretisering, tidsforsinkelser fra sampling og prosessering, samt signalbehandling.

Hvorfor skal du lære dette? Hvorfor bør du lære dette? Når man skal implementere en regulator digitalt så leser man av (måler) verdiene til tilstandene man vil regulere ved gitte tidspunkt.^a Dette kalles tasting/punktprøving (eng.: sampling), og tiden mellom hvert avlesnings-tidspunkt er det vi kaller tastetiden. Dette kan føre til noen viktige elementer man bør ta hensyn til:

- **Effektiv tidsforsinkelse:** Både filtrering og samplingen av de målte signalene, samt beregningene og diskretiseringen av de ønskede regulatorpådragene fører til tidsforsinkelser som man må ta hensyn til i regulatorsyntesen.
- **Folding/aliasing:** Hvis vi taster for sjelden (tastetiden er for lang) i forhold til de høyeste frekvensene i signalet/tilstanden vi vil måle, så kan disse høy-frekvente delene (litt løst forklart) “ødelegge” målingene pga. av et fenomen som kalles *foldning/aliasing*. Dette problemet kan delvis unngås ved å implementere et analogt lavpass-filter med riktig cut-off-frekvens, et så kalt *foldingsfilter*, før man taster signalet.

^aOfte kan man bare måle noen av tilstandene man trenger for å implementere sin regulator, slik at man *estimere* de øvrige tilstandene.

Noen viktige punkter/observasjoner fra [Wittenmark et al., 2002] relatert til dette følger:

- Informasjon kan gå tapt gjennom sampling hvis signalet inneholder frekvenser høyere enn den såkalte [Nyquist-frekvensen](#).
- Sampling kan skape nye foldede frekvenser (aliaser).
- Aliasing/frekvens-foldning gjør det nødvendig å (analogt) filtrere signalene vha. av et folding-filter før de blir samplet.
- Dynamikken i foldingsfilteret må tas hensyn til ved design av den digitale regulatoren. Dynamikken kan dog neglisjeres hvis samplingtiden er kort nok.
- En standard DA-omformer kan beskrives som et zero-order-hold-element. Det er spesielle omformere som gir første-ordens hold, noe som fører til et jevnere kontrollsignal.

12.1.1 Effektiv tidsforsinkelse ved digital regulering

Ved diskret regulering bør man ta hensyn til tidsforsinkelsen som oppstår pga. dette. Anta at man sampler/taster et eller flere signaler med tastid h [s], og at pådraget/regulatorutgangen bestemmes mhp. på dette. En slik tidsforsinkelse, θ_D , vil være innen følgende interval:

$$\frac{h}{2} \leq \theta_D \leq \frac{3}{2}h \quad (\text{Effektiv tidsforsinkelse ved digital regulering})$$

Dette intervallet kommer fra å slå sammen følgende to tidsforsinkelser:

- **Forsinkelse fra hold-element:** Hvis regulatorutgangen blir holdt konstant mellom hver iterasjon (“zero-order hold” (ZOH)), så vil dette før til en effektiv tidsforsinkelse på ca. halvparten av tastetiden, altså $e^{-\frac{\hbar}{2}s}$ i Laplace-domenet.
- **Forsinkelse fra beregning:** Fra man sampler (fra AD-konverteren) til et nytt regulatorsignal blir bestemt (til DA-omformerens) tar det også en viss tid som bør regnes som en effektiv tidsforsinkelse. Denne forsinkelsen er **minimalt** 0 sekunder, som tilsvarer at regulatorutgangen blir satt umiddelbart; og **maksimalt** \hbar sekunder, som tilsvarer at man bruker hele tidsintervallet til å bestemme/regne ut regulatorutgangen.

12.1.2 Tasting/sampling og Nyquist frekvensen



Fun facts, bemerkninger og annet dill dall (you may skip)

[Steve Bruntion video](https://youtu.be/FcXZ28BX-xE) (FcXZ28BX-xE); Denne er også litt relevant, og i hvert fall interessant: <https://youtu.be/rEoc0YoALt0>,

La et signal bli avlest/tastet/samlet (eng.: sampled) med faste tids mellomrom Δt . Vi kaller \hbar *tastetiden*, mens $f_s := 1/\hbar$ (eventuelt $\omega_s := 2\pi/\hbar$) er *tastefrekvensen* (eventuelt sampling- eller prøvetakings-rate). Vi bruker enheten Hertz [Hz=1/s] for f_s og [rad/s] for ω_s .

For at vi skal kunne bevare all vesentlig informasjon i et signal, så må vi ha en høy nok taste-frekvens i forhold til den maksimale (ønskede) frekvensen i signalet. Det er faktisk slik at vi kan gjenskape et analogt signal fra dets målte verdier hvis vi har en tastefrekvens som er høyere enn den såkalte Nyquist-frekvensen:

(Nyquist–Shannon–Kotelnikov samplingsteorem) Et analogt, bånd-begrenset^a signal kan gjenskapes fra diskret målinger med konstant tastetid \hbar hvis tastefrekvensen $f_s := 1/\hbar$ er over dobbelt så stor som den høyeste frekvensen, f_{\max} , i signalet:

$$f_s > 2f_{\max}.$$

For en gitt tastetid \hbar , så kaller vi grensefrekvensen $f_N := \frac{1}{2}f_s = \frac{1}{2\hbar}$ for **Nyquist-frekvensen**, som også kan skrives som

$$\omega_N = 2\pi f_N = \frac{\pi}{\hbar}. \quad \text{(Nyquist-frekvensen)}$$

^aEt signal sies å være **båndbegrenset** hvis det har en endelig maksimal frekvens, og dermed kan representeres med en trunkert Fourier-serie.

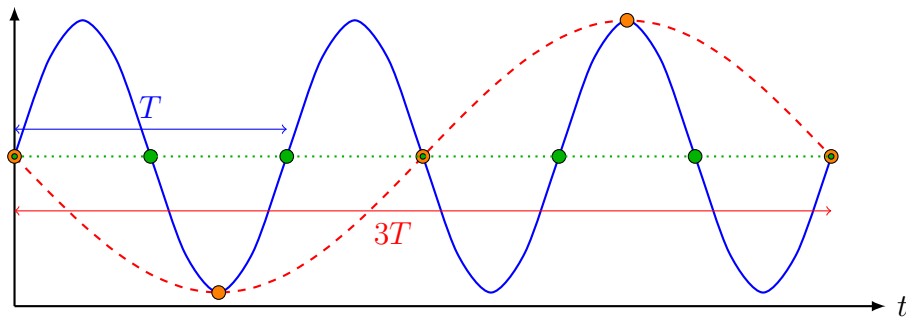
Så hva kan skje hvis vi har et signal som har frekvenskomponenter som er høyere enn Nyquist-frekvensen tilsvarende tastetiden utover at vi ikke kan gjenskape signalet digitalt? Jo, det kan føre til et fenomen kalt (ned-)folding/aliasing.

12.1.3 Fenomenet frekvens-folding/aliasing



Alternative kilder: [Wikipedia](#).

Begrepet folding refererer til symmetrien til et signal om Nyquist-frekvensen i frekvens-domenet.



Figur 12.2: Illustrasjon av folding: Et signal (blått) med periode T blir samplet med tastetid $2T/3$ (se de oransje punktene) $T/2$ (se de grønne punktene). Ut fra disse målingene er disse identiske med henholdsvis et signal tilsvarende den stiplede linjen (i rødt) som har periodetid $3T$ og den prikkede, grønne null-linjen (begge har altså lavere frekvens).

Frekvenser over Nyquist-frekvensen «foldes» (brettes) over Nyquist-frekvensen (derfor blir denne også kalt foldingfrekvensen). Dette fører til at det målte signalet fremstår som at det har lavere frekvens enn det originale analoge signalet, slik som det er illustrert i figur 12.2. Som vist i figur 12.3, så kan også fenomenet oppstå i digitale bilder; se også [▶ yr3ngmRuGUc](#).

Fra dette kan vi gjøre følgende viktige observasjon:

⚠ Aliasing/folding: Hvis vi taster et signal som har komponenter med frekvens høyere enn Nyquist-frekvensen, altså $f_{max} \geq f_N$, så vil disse komponentene fremstå som komponenter med lav frekvens (altså lavere enn Nyquist-frekvensen) i det samplede signalet.

Dette er jo ikke bra! Så hva kan vi gjøre? Jo, vi kan prøve å fjerne komponentene med frekvens høyere enn Nyquist-frekvensen før vi taster signaler ved hjelp av et analogt lavpassfilter, et såkalt *folding-filter*, også ofte kalt *anti-aliasing-filter*.

Folding-filtre

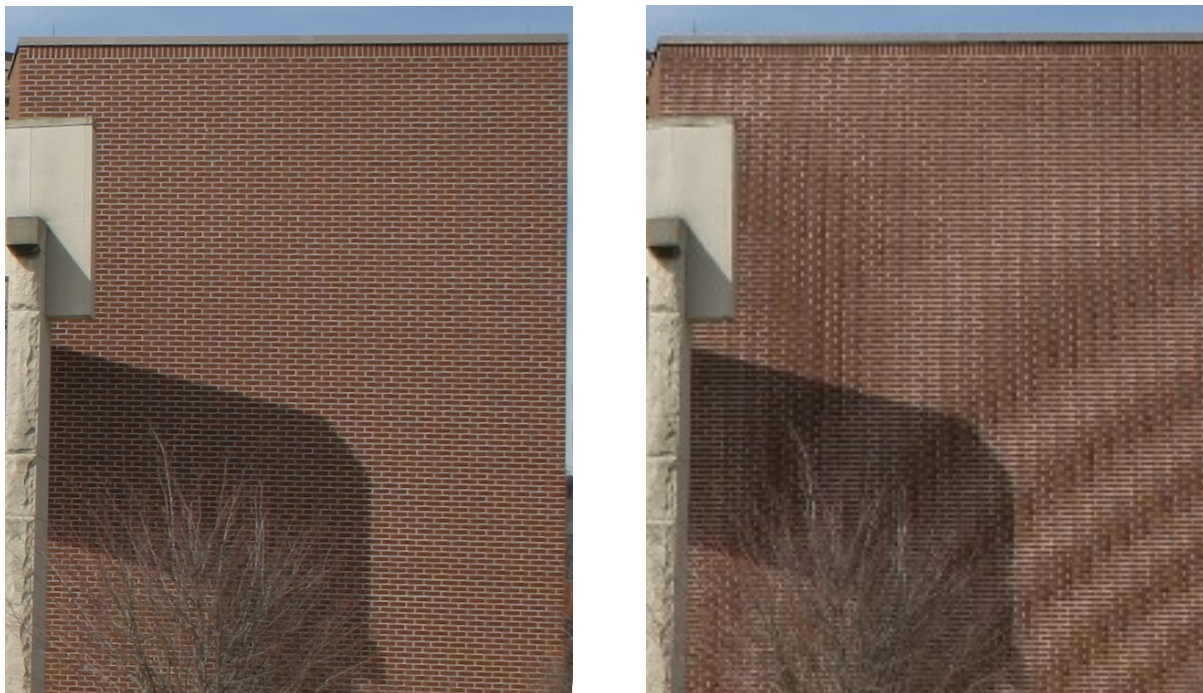
Formålet med et folding-filtre (også kalt anti-aliasing-/konvolusjons-filtre) er som sagt å forhindre at frekvensinnhold nær eller over Nyquist-frekvensen blir med videre i en signalbehandling (for bruk i digital regulering for vår del) etter tasting. Flere kilder [Dessen, 2019, Wittenmark et al., 2002] anbefaler brukene av slike analoge (lavpass-)filtre for digitale reguleringsystemer.

Folding-filtre blir som oftest implementert som analoge elektroniske kretser, noe som gjør dem rimelig enkle. Dette gjelder også designprosedyren, ettersom man normalt sett ikke trenger å ta store hensyn til det digitale reguleringsystemet utover Nyquist-frekvensen.

Hvordan designe og implementere et slikt filter? Man tar utgangspunkt i Nyquist-frekvensen og spesifiserer ønsket demping av det filtrerte signalet ved denne frekvensen. Basert på dette, implementerer man et lavpassfilter (f.eks. et Butterworth-filter) som man igjen konstruerer fysisk ved hjelp av en elektrisk krets.

Tips: I en reguleringsløyfe kan det være lurt å også sende referansen gjennom et lavpassfilter tilsvarende det brukte foldingfiltre. Dette kan dog implementeres digitalt. Dette har den fordel at det tilater varierende forsterkning i passbåndet!

Figur 7 i [Wittenmark et al., 2002]



Figur 12.3: Illustrasjon av folding/aliasing, hvor “bølger” i murveggen oppstår i bildet til høyre siden det har lavere oppløsning enn bildet til venstre. Bilder fra Wikipedia (CC BY-SA 3.0, <https://commons.wikimedia.org/w/index.php?curid=644816>).

12.2. Lavpass- og Butterworth-filtre

Alternative kilder: Kap. 17 i [Seborg et al., 2016]; [Dessen, 2019]; §11.3 i [Balchen et al., 2016]

Hvorfor skal du lære dette? Et lavpass-filter har som oppgave å **1)** fjerne (uønskede) høy-frekvente deler av et signal (f.eks. støy), men samtidig **2)** beholde de delene som er under en gitt «cutt-off» frekvens, og dermed oppnå ønsket båndbredde.

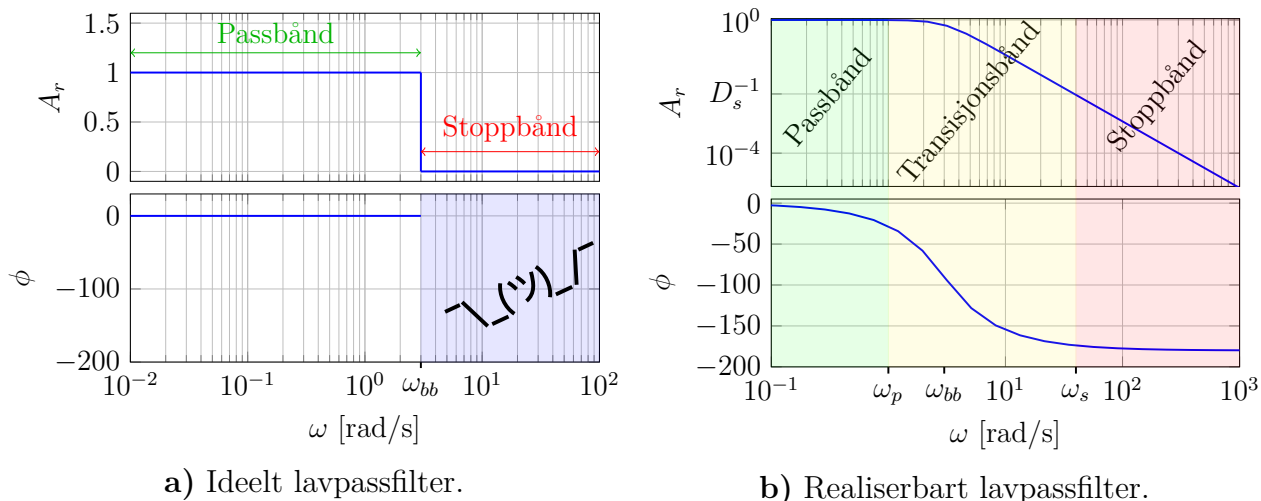
For våre formål i disse notatene, vil dette være spesielt relevant for å kunne designe et såkalt **foldingsfilter** i § 12.1.3.

12.2.1 Lavpassfiltre [yM4z44-vHT4&t=508](https://www.youtube.com/watch?v=yM4z44-vHT4&t=508)

Som nevnt over, så har et lavpassfilter to viktige oppgaver:

1. slippe gjennom lavfrekvente komponenter av et signal mest mulig uforstyrret (**passbånd**);
2. blokkere høyfrekvente komponenter i størst mulig grad (**stoppbånd**).

I skillet mellom passbåndet og stoppbåndet har vi det vi kaller **cut-off-frekvensen** (kanskje **avkutt-frekvens** på godt norsk?). Vi ønsker altså å “kutte” alle frekvenser som er høyere enn

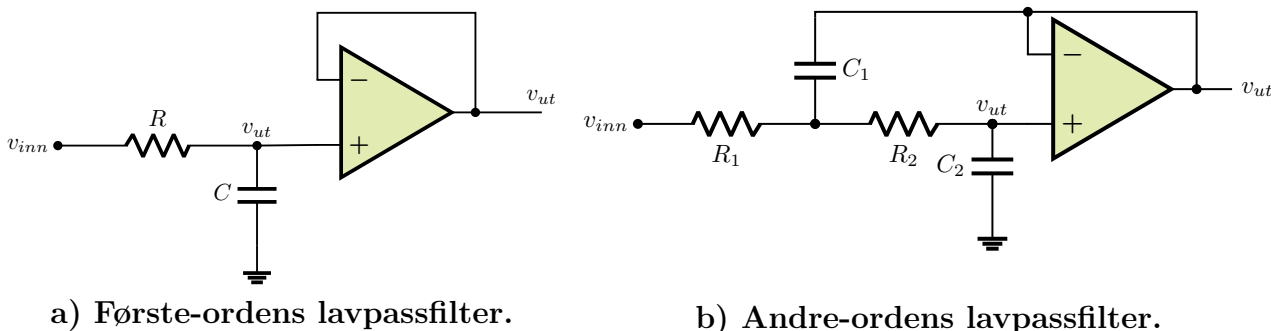


Figur 12.4: Illustrasjon av et ideelt lavpassfilter til venstre og et realiserbart (andre-ordens) filter til høyre.

denne. **I en perfekt verden**, hvor man kunne ha laget et ideelt lavpassfilter, ville grensen mellom stoppbånd og passbånd være helt presist definert, og det er ingen endring i fasen slik som vist i figur 12.4 a). Dette er dog umulig å få til i praksis, og man må alltid kunne leve med et såkalt *transisjonsbånd* mellom pass- og stoppbåndene.

12.2.2 Lavpassfiltre som elektriske kretser

Eksempler på elektriske kretser som tilsvarer et først- og andre-ordens lavpassfilter ([Sallen-Key-topologi](#)) er vist i figur 12.5 (det finnes en rekke andre kretser for dette, såkalte topologier). Merk at man ved å koble disse i serie får et tredje-ordens filter. Legg også merke til at operasjonsforsterkerne (“opampene”) antas som ideelle (altså uendelig inngangsmotstand og uendelig forsterkning), slik at de fungerer som spenningsfølgere hvor spenningen ut tilsvarer spenning på +-inngangen.



Figur 12.5: Eksempler på elektriske kretser tilsvarende første- og andre-ordens lavpassfiltre.

Ved å ta i bruk Kirchoffs strømlov (se § 3.2.4) ved +-inngangen til opampen i første-ordensfilteret (altså **a**) i fig. 12.5, får vi

$$\frac{(v_{ut} - v_{inn})}{R} + C \frac{dv_{ut}}{dt} = 0 \implies \frac{dv_{ut}}{dt} = -\frac{1}{RC} (v_{ut} - v_{inn}).$$

Tilsvarende overføringsfunksjon fra inngang til utgang med cut-off-frekvens/båndbredde $\omega_{bb} = 1/(RC)$ (ofte brukes ω_c i stedet) er dermed som følger:

$$G_{FO}(s) = \frac{\omega_{bb}}{s + \omega_{bb}} = \frac{1}{\frac{s}{\omega_{bb}} + 1}.$$

La oss nå finne tilsvarende ligninger for andre-ordens filteret i figure 12.5-b). Vi starter med å igjen bruke Kirchhoffs strømlov i forgreningspunktet til høyre for R_1 -motstanden, hvor vi antar spenningen er lik v_1 :

$$\frac{(v_1 - v_{inn})}{R_1} + \frac{(v_1 - v_{ut})}{R_2} + C_1 \left(\frac{dv_1}{dt} - \frac{dv_{ut}}{dt} \right) = 0.$$

Ved å også ta i bruk strømloven ved +-inngagen til opampen, får vi $\frac{(v_{ut} - v_1)}{R_2} + C_2 \frac{dv_{ut}}{dt} = 0$, hvorfra vi har at $v_1 = v_{ut} + R_2 C_2 \frac{dv_{ut}}{dt}$, og dermed $\frac{dv_1}{dt} = \frac{dv_{ut}}{dt} + R_2 C_2 \frac{d^2 v_{ut}}{dt^2}$. Ved å sette dette inn i uttrykket over får vi

$$\frac{(v_{ut} + R_2 C_2 \frac{dv_{ut}}{dt} - v_{inn})}{R_1} + \frac{(v_{ut} + R_2 C_2 \frac{dv_{ut}}{dt} - v_{ut})}{R_2} + C_1 \left(\frac{dv_{ut}}{dt} + R_2 C_2 \frac{d^2 v_{ut}}{dt^2} - \frac{dv_{ut}}{dt} \right) = 0.$$

Ved å kansellere leddene markert i rødt og så multiplisere på begge siden med R_1 får man

$$(v_{ut} + R_2 C_2 \frac{dv_{ut}}{dt} - v_{inn}) + R_1 C_2 \frac{dv_{ut}}{dt} + R_1 C_1 R_2 C_2 \frac{d^2 v_{ut}}{dt^2} = 0.$$

Tilsvarende overføringsfunksjon fra $V_{inn}(s) = \mathcal{L}\{v_{inn}\}$ til $V_{ut}(s) = \mathcal{L}\{v_{ut}\}$ er dermed

$$G_{AO}(s) = \frac{V_{ut}(s)}{V_{inn}(s)} = \frac{1}{R_1 C_1 R_2 C_2 s^2 + (R_1 + R_2) C_2 s + 1}.$$

12.2.3 Butterworth-filtre

► yM4z44-vHT4&t=967

De fleste LTI-systemer med en strengt proper og minimum fase overføringsfunksjon kan regnes som et lavpassfilter hvis det har stasjonærforsterkning lik én. Det er dog som regel ønskelig at slike filtre har visse karakteristikk/egenskaper, noe som har gitt oss en rekke «familier» av slike filtre, hvorav følgende trolig er de mest kjente: [Butterworth](#)-, [Chebyshev](#)- og [Bessel](#)-filtre. Vi skal her kun se på førstnevnte, nemlig Butterworth-filtre, oppkalt etter Stephen Butterworth:

“An ideal electrical filter should not only completely reject the unwanted frequencies but should also have uniform sensitivity for the wanted frequencies”.

—Stephen Butterworth

Butterworth-filteret har den karakteristikken at det har en «maksimalt» flatt forsterkningskarakteristikk i passbåndet i tillegg til en monotont avtagende fase i hele spekteret (uten sistnevnte krav kan man lage lavpassfiltere som er «flatere» i passbåndet).

Et n -te-ordens **Butterworth-filtre** med båndbredde, ω_{bb} , har overføringsfunksjon

$$G_{BW}(s) = \frac{1}{B_n(s/\omega_{bb})}, \quad (\text{Butterworth-filtre})$$

hvor $B_n(\cdot)$ er et n -te-ordens Butterworth-polynom (se også tabell 12.1):

$$B_n(s) = \sum_{k=0}^n a_k s^k, \quad a_{k+1} = \frac{\cos(k\gamma_n)}{\sin((k+1)\gamma_n)} a_k, \quad a_0 = 1, \quad \gamma_n = \frac{\pi}{2n}.$$

Merk: *i)* (cut-off-)frekvensen betegnes ofte som ω_c , men vi bruker ω_{bb} siden denne angir filterets båndbredde; *ii)* har den egenskapen at forsterkningen ved ω_{bb} tilsvarer $1/\sqrt{2}$ (≈ -3 dB); *iii)* et n -te ordens Butterworth-filter har følgende forsterkning/**Amplituderatio**:

$$A_r(\omega) = \sqrt{\frac{1}{1 + \left(\frac{\omega}{\omega_{bb}}\right)^{2n}}}.$$

Tabell 12.1: Faktorisert Butterworth-polynom $B_n(s)$ for forskjellige filterordner n , hvor $\varphi = (1 + \sqrt{5})/2$ en **det gyldne snitt**. Tabellen er tatt fra [Wikipedia](#).

n	$B_n(s)$
1	$(s + 1)$
2	$(s^2 + \sqrt{2}s + 1)$
3	$(s + 1)(s^2 + s + 1)$
4	$(s^2 + \sqrt{2 - \sqrt{2}}s + 1)(s^2 + \sqrt{2 + \sqrt{2}}s + 1)$
5	$(s + 1)(s^2 + \varphi^{-1}s + 1)(s^2 + \varphi s + 1)$
6	$(s^2 + \sqrt{2 - \sqrt{3}}s + 1)(s^2 + \sqrt{2}s + 1)(s^2 + \sqrt{2 + \sqrt{3}}s + 1)$

⚠ NB! I større grad enn visse andre lavpass-filtre, har Butterworth-filtre en underdempet sprangrespons med tydelig oversving; dette er det viktig å være klar over for regulerings-tekniske formål.

Et analogt Butterworth-filter kan genereres i MATLAB ved hjelp av koden i kodesnutt 12.1.

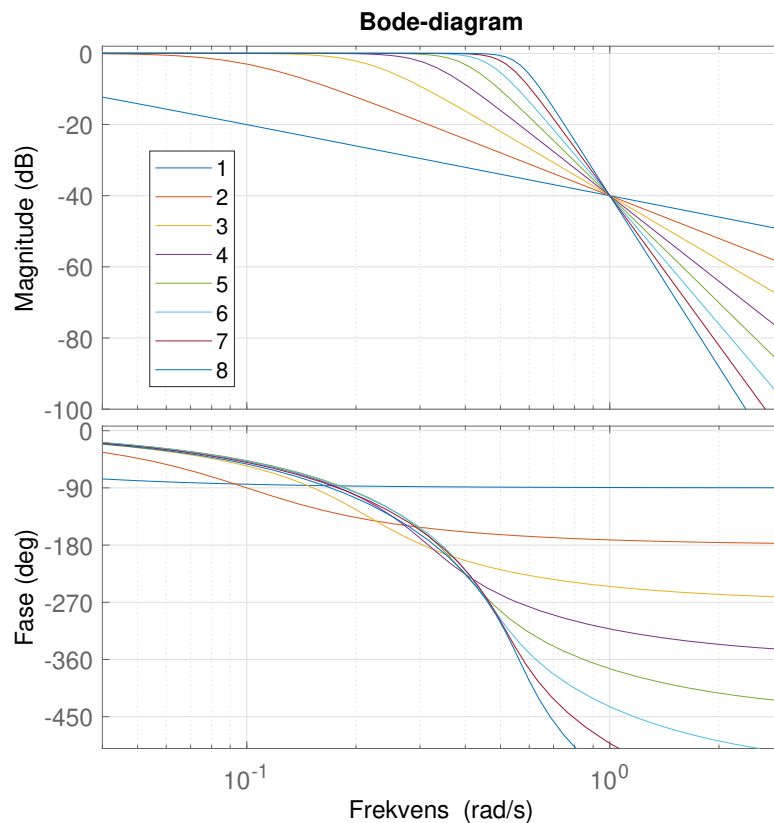
Kodesnutt 12.1: MATLAB-kommando for et analogt Butterworth-filter (se <https://se.mathworks.com/help/signal/ref/butter.html>).

```
% n-te-ordens Butterworth filter med cut-off-frekvens wn:
[T,N] = butter(n,wn,'s'); % legger til 's' for et analogt filter
G=tf(T,N); bode(G)      % Plotter Bode-diagrammet
```

Figur 12.6 viser Butterworth-filtre av orden 1 til 8, alle designet med en demping tilsvarende $D_s = 40$ dB ved $\omega_s = 1$ rad/s.

Fun facts, bemerkninger og annet dill dall (you may skip)

Polene til Butterworth-filtrene kan finnes på følgende måte: For et filter av orden n , tegn en $2n$ -kant med hjørner som ligger på den komplekse enhets sirkelen, slik at den er symmetrisk om både den reelle- og imaginære akse. Behold så bare de polene som ligger i venstre halvplan.



Figur 12.6: Bode-diagram av Butterworth filtere av orden 1-8, alle designet for å gi en dempning tilsvarende 40 dB ved $\omega_s = 1$ rad/s.

Hvordan velge filteret?



Et Butterworth-filter er bestemt av to parametere: båndbredden, ω_{bb} , og filterets orden, n .

Nåværende problem: Bestemme ω_{bb} og n ut fra følgende fire parametere, som er naturlige spesifikasjoner når man ønsker å designe et filter for et gitt problem:

ω_s frekvensen som angir starten på stoppbåndet;

D_s ($\gg 1$) ønsket dempningsratio ved ω_s ;

ω_p frekvensen som angir slutten på passbåndet;

D_p ønsket dempningsratio ved ω_p .

Hvordan sette cut-off-frekvensen/båndbredden:

La oss si at vi ønsker at filter skal ha en dempnings-ratio ved frekvensen ω_s (som angir starten på stoppbåndet) tilsvarende $D_s \gg 1$. Vi får dermed $D_s^2 = 1/A_n^2(\omega_D) = 1 + \left(\frac{\omega_s}{\omega_{bb}}\right)^{2n}$, slik at man får følgende:

Båndbredde fra stoppbånd-spesifikasjon: Gitt et ønsket stoppbånd-par (ω_s, D_s) , ta

$$\omega_{bb} = \frac{\omega_s}{\sqrt[2n]{D_s^2 - 1}} = \omega_s (D_s^2 - 1)^{\frac{-1}{2n}}. \quad (12.1)$$

Eksempel 12.1. (Finne filterets båndbredde) Anta at stoppbåndet er definert ved at dempningen skal være mer enn 60 dB for alle frekvenser over 50 rad/sek. Anta videre at filteret skal ha orden 7. Med andre ord ønskes det at $D_s \geq 60$ dB for alle $\omega \geq \omega_s = 50$ rad/s, med $n = 7$.

Løsning: $\omega_{bb} = 50 / (1000^2 - 1)^{(1/14)} \approx 18.64$ rad/s

Merk: For å finne ω_{bb} fra denne formelen, må vi jo dog allerede vite filterets orden n . Vi skal se på to måter for å velge denne:

1. Den «klassiske» metoden basert på transisjonsbåndet; og
2. Dessens metode, som er basert på minimering av filterets «tidsforsinkelse».

Metoder for å bestemme filterets orden

▶ yM4z44-vHT4&t=2059

Den klassiske metoden: La oss si at vi ønsker å finne ut hva passbåndet er ved å finne ved hvilken frekvens ω_p dempnings-ratioen er tilnærmet $D_p \geq 1$. Dette gir oss jo

$$\omega_p = \omega_{bb} \sqrt[2n]{D_p^2 - 1}. \quad (12.2)$$

Vi kan dog bruke dette til å finne forholdet $\omega_s/\omega_p > 1$ som angir størrelsen på transisjonsbåndet (se fig. 12.4); desto nærmere dette forholdet er lik 1, desto smalere er dette båndet. Si at vi ønsker at forholdet skal være mindre eller lik $\gamma > 1$, altså $\gamma \geq \omega_s/\omega_p$. Vi har da fra (12.1)-(12.2) at

$$\gamma \geq \frac{\omega_s}{\omega_p} = \frac{\omega_s}{\omega_{bb} \sqrt[2n]{D_p^2 - 1}} = \frac{\sqrt[2n]{D_s^2 - 1}}{\sqrt[2n]{D_p^2 - 1}} \implies \ln(\gamma) \geq \frac{1}{2n} \ln\left(\frac{D_s^2 - 1}{D_p^2 - 1}\right).$$

Dermed får vi følgende:

Klassisk metode for å bestemme filterorden: Fra både stoppbånd- og passbånd-spesifikasjonene, gitt ved henholdsvis (ω_s, D_s) og (ω_p, D_p) , la $\gamma \geq \frac{\omega_s}{\omega_p}$ og ta

$$n \geq \frac{1}{2} \frac{\ln\left(\frac{D_s^2 - 1}{D_p^2 - 1}\right)}{\ln(\gamma)} = \frac{\ln(D_s^2 - 1) - \ln(D_p^2 - 1)}{2 \ln(\gamma)}. \quad (12.3)$$

Merk: Den klassiske metoden over tar kun hensyn til filterets forsterkning. Det er helt OK hvis filterets faseforskyvning ikke er av betydning, samt at det er relevant å spesifisere passbåndet.

Dessens metode: Problemet med den klassiske metoden for å bestemme filterets orden fra et regulerings teknisk perspektiv, er at den ikke tar hensyn til den effektive forsinkelsen. Dette er derimot hovedkriteriet i Dessens metode:

Dessens metode for å bestemme filterorden ([Dessen, 2019]): Gitt stoppbåndspesifikasjoner i form av paret (ω_s, D_s) , bruk tabell 12.2^a for å finne den optimale filterordenen n for den spesifiserte dempningen D_s .

Effektiv tidsforsinkelse: Med metoden følger det også en måte å estimere den minste effektive tidsforsinkelsen:^b

$$5 \cdot \omega_s \cdot \theta_{\min} \approx 20 \log_{10}(D_s).$$

^aFor dempningsverdier som faller mellom de oppgitte verdiene finnes det en "tallinje-tabell" i [Dessen, 2019] som kan brukes.

^bMerk at tilsvarende formel for et Bessel-filter er $4 \cdot \theta_{\min} \approx 20 \log_{10}(D_s)$.

Tabell 12.2: Dessen sin metode til å finne ordenen til et Butterworthfilter; [Dessen, 2019].

D_s [dB]	20	30	40	50	60	70	80	90
D_s	10	$\approx 3.16 \cdot 10$	10^2	$\approx 3.16 \cdot 10^2$	10^3	$\approx 3.16 \cdot 10^3$	10^4	$\approx 3.16 \cdot 10^4$
n	3	4	5	6	7	8	9	10

Eksempel 12.2. Anta at man ønsker å konstruere et Butterworth-filter hvor dempningen skal være mer enn 40 dB for frekvenser over 100 rad/sek. Vi vil derfor finne ordenen som minimalisere fasefall/forsinkelse, og så bestemme hva ω_{bb} skal være.

Løsning: Fra tabell 12.2 ser vi at vi bør velge $n = 5$ siden $D_s = 40 \text{ dB} = 100$. Fra (12.1) får vi dermed $\omega_{bb} = 100/(100^2 - 1)^{(1/10)} \approx 39.81 \text{ rad/s}$.

Eksempel 12.3. Anta at dempningen skal være mer enn 70 dB for frekvenser over 30 rad/sek. Vi vil finne filteret med den beste orden for å minimalisere fasefall eller forsinkelse, samt finne de tilsvarende forsinkelsen.

Løsning: Dempningen på 70 dB tilsier at ordenen skal være $n = 8$. Vi får dermed fra (12.1) at $\omega_{bb} \approx 30/(3160^2 - 1)^{(1/(2 \cdot 8))} \approx 11 \text{ rad/s}$. Den minste tidsforsinkelsen er $\theta_{\min} = 70/(5 \cdot 30) \approx 0.47 \text{ s}$.

12.2.4 Alternative lavpass-filtre: Chebyshev, Bessel og elliptiske*

Chebyshev-filter, Bessel-filter og Elliptisk filter

[Wikipediafigur for sammenligning av filtre](#)

12.3. *Høypass-, båndpass og båndstopp-filtre*

Alternative kilder: §11.3 i [Balchen et al., 2016]

Høypass: ta et lavpass-filter og bytt s med ω_c^2/s .

Båndpass: ta et lavpass-filter i serie med et høypass-filter.

Båndstopp: en måte å få et båndstoppfilter, også kalt hakk-filter (eng. “notch filter”), er å ta et lavpass-filter parallelt med et høypass-filter. Noe vanligere er dog å bruke filtere med såkalt Q -faktor i serie. Et slikt filter har følgende form: [referanse her?](#)

$$N(s) = \frac{s^2 + \frac{(1-c_d)}{Q}\omega_h s + \omega_h^2}{s^2 + \frac{\omega_h}{Q}s + \omega_h^2} \quad (\text{Båndstoppfilter})$$

hvor ω_h er hakk-frekvensen, Q er Q -faktoren og c_d er hakk-dybden.

Disse er ganske nyttige (se f.eks. denne [Brian Douglas videoen](#)), og de brukes også tidvis som en tuning-parameter i visse lukkede sløyfer (se f.eks. figur 1.5).

Sensorfusjon*

Nåværende problem: Vi ønsker å måle en prosessvariabel, y . Vi har to sensorer tilgjengelig, som gir oss målingene^a $y_m^1 = y + m_1$ og $y_m^2 = y + m_2$, hvor m_1 og m_2 er målestøyene til de to sensorene. F.eks. kan y_m^2 være en rask måling hvor m_1 har liten varians, men muligens varierende (drivende) middelværdi; mens y_m^1 er en tregere sensor med stor varians i støyen, men med stabil middelværdi. **Mål:** kombinere (fusjonere) målingene de to målingene til å oppnå en ny, bedre måling.

^aSensorene trenger ikke nødvendigvis måle den ønskede prosessvariabelen direkte, en indirekte måling hvorfra man kan estimere den er et vanlig scenario som man også kan dekkes av metodene vi skal se på.

12.3.1 Komplementærfilter

Gitt problemstillingen over (f.eks. y_m^1 er en IMU¹ og y_m^2 er en GPS-måling), så kan et *komplementærfilter* være en mulig strategi:

Komplementærfilter: Gitt to målinger, $y_m^1 = y + m_1$ og $y_m^2 = y + m_2$, hvor m_1 og m_2 målestøyene. Ta, for en positiv τ_m ,

$$\hat{y}_m = \frac{1}{\tau_m s + 1} y_m^1 + \frac{\tau_m s}{\tau_m s + 1} y_m^2$$

Dermed en kombinasjon av et lasspassfilter og en høypassfilter: Vi bruke lavpassfilteret til å beholde den stasjonære verdien fra den trege målingen y_m^1 , målingen med stor varians, mens vi for den raske, men drivende målingen y_m^2 bare beholder den de høypassfilteret delene.

12.3.2 Kalmanfilter*



¹Et vanlig bruksområde til et komplementærfilter er fusjonering av gyroskopmålinger i en IMU med dens akselerometermålinger.

Del VI

Regulering av multi-variable, koblede systemer

13. Multivariable, koblede systemer

Vi skal nå se på hvordan man kan regulere multivariable, koblede prosesser, det vil si systemer med flere innganger og utganger, såkalte MIMO-systemer.

13.1. Hva er multivariable, koblede systemer?



Vi starter like greit med noen definisjoner.

Multivariable systemer: Et multivariabelt (MIMO) reguleringsystem har minst to pådrag og minst to utganger.

Koblede (coupled) systemer: Systemer med én eller flere krysskoblinger, altså reguleringsystemer hvor en tilstand påvirker dynamikken til en annen tilstand

Koblet MIMO-system: Følgende system med to tilstander og utganger $(y_1, y_2) = (x_1, x_2)$,

$$\dot{y}_1 = -y_1 + u_1 + \underbrace{2y_2 - u_2}_{\text{Krysskoblinger}}$$

$$\dot{y}_2 = u_2,$$

sies å være koblet siden \dot{y}_1 avhenger av / har en kobling til både y_2 og u_2 . Med andre ord, hvis vi bruker u_2 til å regulere y_2 , så vil dette ha en innvirkning på dynamikken til y_1 .

Dekoblet MIMO-system:

$$\dot{y}_1 = -y_1 + u_1$$

$$\dot{y}_2 = u_2$$

Multivariable systemer kalles ofte bare for MIMO-systemer, mens monovariable systemer kalles for SISO-systemer:

SISO=Single-Input-Single-Output. Med andre et system med **én inngang** og **én utgang**.

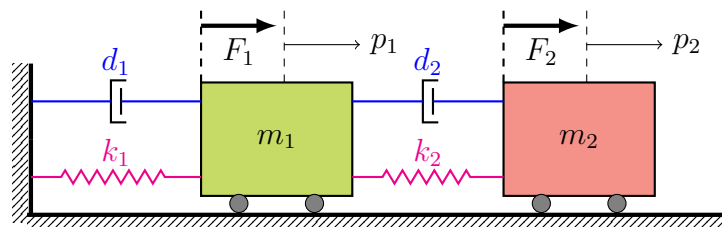
MIMO=Multiple-Input-Multiple-Output. Med andre et system med **flere innganger** og **flere utganger**.

De fleste systemer man møter på i prosessindustrien er MIMO-systemer (selv om de noen ganger kan reguleres som om de var SISO-systemer). Noen eksempler følger:

- kjemiske reaktorer (temperatur, nivå, konsentrasjon)
- blandekar (nivå og konsentrasjon)
- destillasjonskolonner (mange tilstander)
- dampkjeler (nivå og damptrykk)
- varmevekslere (temperatur og strømning)
- innløpskasser i papirindustrien (trykk og nivå)

La oss starte med et motiverende eksempel som ikke er prosessindustri-relatert, men som meget tydelig fremhever utfordringene slike systemer ofte har fra et reguleringsteknisk perspektiv.

13.1.1 Motiverende eksempel: To koblede traller



Figur 13.1: To traller koblet sammen via en fjær og en demper.

Gitt et dynamisk system bestående av to traller koblet til flere masse-fjære-dempere, som vist i figur 13.1. La p_i være posisjonen til tralle i , og $v_i = \frac{dp_i}{dt}$ dens hastighet.

I eksempel 3.7 fant vi at ved å ta $(x_1, x_2, x_3, x_4) = (p_1, p_2, v_1, v_2)$ og $(u_1, u_2) = (F_1, F_2)$, så kunne systemets dynamiske ligninger skrives på tilstandsromform:

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}$$

hvor

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ b_{31} & 0 \\ 0 & b_{42} \end{bmatrix}, \quad \mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}.$$

for et sett av parametere (a_{ij}, b_{ij}) som avhenger av massene, samt fjær- og dempningskon-

stantene. La oss anta at vi har to utganger, $y_1 = x_1$ og $y_2 = x_2$, eller tilsvarende:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \mathbf{x} = \mathbf{C}\mathbf{x}.$$

Vi ønsker nå å finne tilsvarende system representert i Laplace-domenet ved hjelp av metoden vi så på i § 2.6.3. La oss anta at $\mathbf{x}_0 = \mathbf{x}(0) = 0$, noe som da gir

$$\mathbf{Y}(s) = \mathbf{C}(\mathbf{I}_4s - \mathbf{A})^{-1}(\mathbf{B}\mathbf{U}(s) + \mathbf{x}_0) = \mathbf{P}(s)\mathbf{U}(s)$$

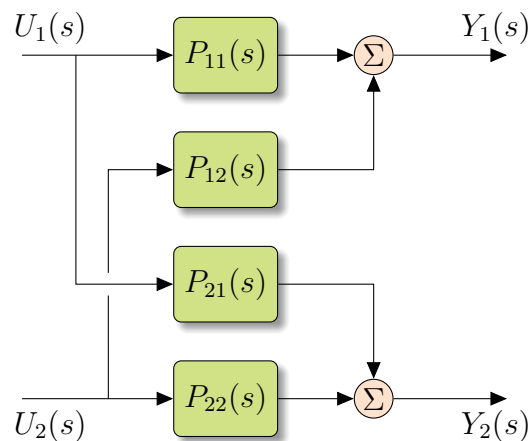
hvor $\mathbf{P}(s)$ er en såkalt *transfermatrise*, gitt ved

$$\mathbf{P}(s) = \underbrace{\mathbf{C}}_{2 \times 4} \underbrace{(\mathbf{I}_4s - \mathbf{A})^{-1}}_{4 \times 4} \underbrace{\mathbf{B}}_{4 \times 2} = \underbrace{\begin{bmatrix} P_{11}(s) & P_{12}(s) \\ P_{21}(s) & P_{22}(s) \end{bmatrix}}_{2 \times 2}.$$

Dette tilsvarer igjen

$$\begin{aligned} Y_1(s) &= P_{11}(s)U_1(s) + P_{12}(s)U_2(s) \\ Y_2(s) &= P_{21}(s)U_1(s) + P_{22}(s)U_2(s). \end{aligned}$$

En illustrasjon av dette systemet er vist i figur 13.2. Denne figuren illustrerer en viktig egenskap med MIMO-systemer: det eksisterer ofte vekselvirkninger mellom de ulike pådragene og utgangene. **Problem:** Hvis disse krysskoblingene (altså $P_{12}(s)$ og $P_{21}(s)$) er store, så må vi ta hensyn til dem når vi skal designe reguleringsstrategi for systemet.



Figur 13.2: Enkelt blokkdiagram for 2×2 MIMO-system

13.1.2 Hvordan regulere koblede, multivariable systemer?

Nåværende problem: Ofte møter man på systemer med flere innganger og utganger. Det er også ofte krysskoblinger (f.eks. hvordan dynamikken til de to trallene påvirker hverandre

gjennom fjæren og demperen som kobler de sammen), slik at utgangene i utgangspunktet ikke kan reguleres helt uavhengig fra hverandre. Fra et regulerings teknisk perspektiv, så er det noen viktige spørsmål som dukker opp:

1. Hvordan kan vi designe regulatorer for slike systemer?
2. Hva må vi i så fall ta hensyn til?
3. Og (når) kan vi bruke metodene vi har sett på for monovariabel systemer (cf. kap. 7)?

Hvis vi er veldig heldig, og vekselvirkningene er meget svake, så trenger vi ikke ta noen ekstra hensyn; men hvordan vet vi at de er svake, og hvilken inngang skal styre hvilken utgang? På den annen side, når vekselvirkningene er sterke, så kan det oppstå problemer som ikke dekkes av grunnleggende regulerings teori.

NB! I dette kapitlet skal vi fokusere på metoder¹ hvor man prøver å regulere hver utgang med én inngang, slik at man til en viss grad direkte kan bruke strategiene vi har sett på tidligere i disse notatene for hver enkelt sløyfe.

13.1.3 Generell form til et koblet, multivariabelt system

Dynamikken til et generelt lineært, koblet, multivariabelt (MIMO) system med p utganger og m innganger er gitt av overføringsfunksjonene $P_{ij}(s)$:

$$\begin{aligned} Y_1(s) &= P_{11}(s)U_1(s) + P_{12}(s)U_2(s) + \dots + P_{1m}(s)U_m(s) \\ Y_2(s) &= P_{21}(s)U_1(s) + P_{22}(s)U_2(s) + \dots + P_{2m}(s)U_m(s) \\ &\vdots \\ Y_p(s) &= P_{p1}(s)U_1(s) + P_{p2}(s)U_2(s) + \dots + P_{pm}(s)U_m(s) \end{aligned} \quad (13.1)$$

Ved å introdusere en *transfermatrise*², $\mathbf{P}(s) \in \mathbb{C}^{r \times m}$, kan dette skrives på den følgende, mer kompakte formen:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_r \end{bmatrix} = \mathbf{Y}(s) = \mathbf{P}(s)\mathbf{U}(s) = \begin{bmatrix} P_{11} & P_{12} & \dots & P_{1m} \\ P_{21} & P_{22} & \dots & P_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ P_{r1} & P_{r2} & \dots & P_{rm} \end{bmatrix} \begin{bmatrix} U_1 \\ U_2 \\ \vdots \\ U_m \end{bmatrix}. \quad (13.2)$$

For enkelhetens skyld, skal vi i dette kapitlet fokusere på systemer hvor $p = m = 2$:

13.1.4 Vårt fokus: desentralisert regulering av koblede 2x2 systemer

For et system med to innganger og to utganger, altså

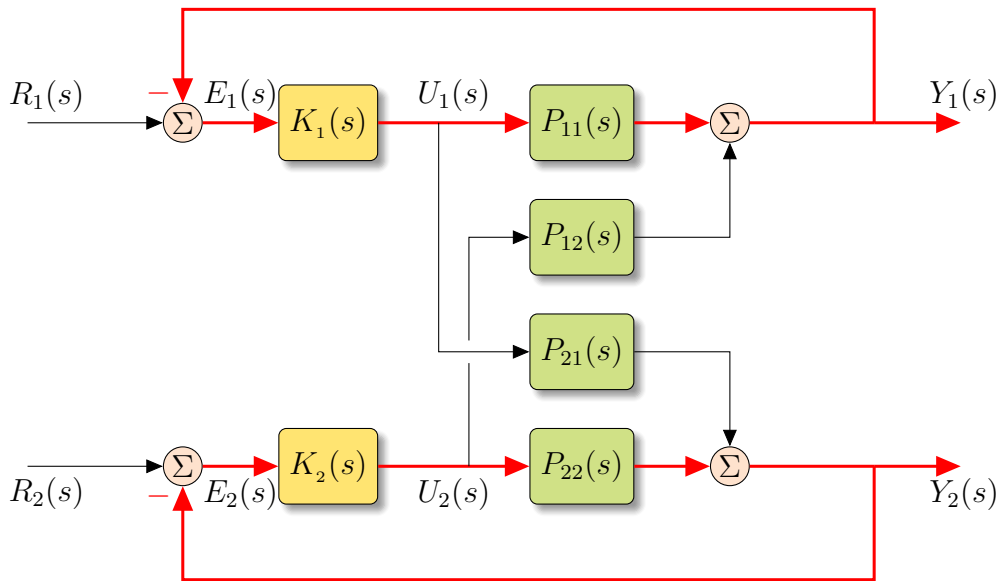
$$Y_1(s) = P_{11}(s)U_1(s) + P_{12}(s)U_2(s)$$

¹Samtlige av metodene vi skal se på i dette kapitlet baserer seg på bruken av overføringsfunksjoner. I kapittel 13.3.2 (WIP) skal vi derimot se på tilstandsrom-baserte metoder for generell multivariabel regulering, hvor man prøver å regulere systemet med hjelp av full-tilstands-basert tilbakekobling (altså uten å prøve å dele systemet om i flere enkeltsløyfer).

²Man kan her også si “overførings-matrisefunksjon” fremfor “transfermatrise”, men vi holder oss til sistnevnte siden det er det som oftest er brukt.

$$Y_2(s) = P_{21}(s)U_1(s) + P_{22}(s)U_2(s).$$

slik som vist i figur 13.2, kan det være sterke vekselvirkninger mellom utgangene og pådragene gjennom krysskoblingene gitt av overføringsfunksjonene $P_{12}(s)$ og $P_{21}(s)$. Hvis derimot $P_{12}(s)$ og $P_{21}(s)$ er små, kan vi designe regulatorer som om $Y_i(s)/U_i(s) \approx P_{11}(s)$, altså som to SISO-systemer som vist i figur 13.3. Dette kalles for **desentralisert regulering**, siden vi separerer sløyfene og dermed desentraliserer reguleringsproblemet. Dette er i motsetning til en *sentralisert strategi*, hvor man for eksempel tar $\mathbf{U}(s) = \mathbf{K}(s)\mathbf{E}(s)$ gitt en 2×2 regulatormatrise $\mathbf{K}(s)$ som ikke nødvendigvis bare har elementer på hoveddiagonalen.



Figur 13.3: Blokkdiagram for 2×2 MIMO-system med forslag til tilbakekobling.

Hvordan en slik strategi påvirker systemets dynamikk kommer frem av følgende oppgave:

Oppgave 13.1. Vis at diagrammet i figur 13.3 gir følgende overføringsfunksjoner:

$$Y_1(s) = \left[P_{11}(s) - P_{12}(s) \frac{K_2(s)P_{21}(s)}{1 + K_2(s)P_{22}(s)} \right] U_1(s) + P_{12}(s) \frac{K_2(s)}{1 + K_2(s)P_{22}(s)} R_2(s)$$

$$Y_2(s) = \left[P_{22}(s) - P_{21}(s) \frac{K_1(s)P_{12}(s)}{1 + K_1(s)P_{11}(s)} \right] U_2(s) + P_{21}(s) \frac{K_1(s)}{1 + K_1(s)P_{11}(s)} R_1(s).$$

Fra oppgaven over, har man derfor følgende relasjoner for systemet i figur 13.3:

$$Y_1 = G_{11}(s)R_1 + G_{12}(s)R_2, \tag{13.3a}$$

$$Y_2 = G_{21}(s)R_1 + G_{22}(s)R_2 \tag{13.3b}$$

hvor

$$G_{11}(s) := \frac{K_1 P_{11} + K_1 K_2 (P_{11} P_{22} - P_{12} P_{21})}{\Delta(s)}$$

$$G_{12}(s) := \frac{K_2 P_{12}}{\Delta(s)}$$

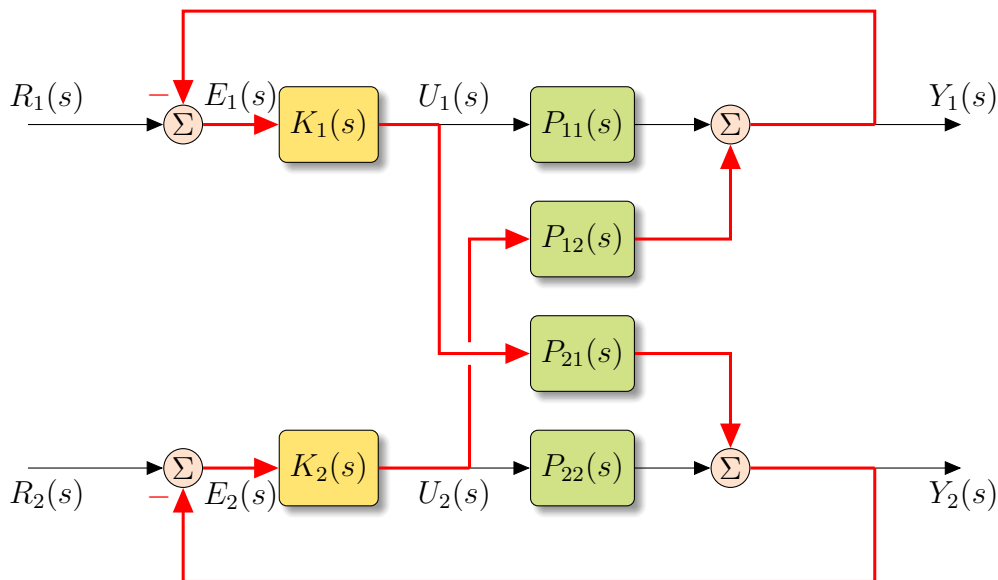
$$G_{21}(s) := \frac{K_1 P_{21}}{\Delta(s)}$$

$$G_{22}(s) := \frac{K_2 P_{22} + K_1 K_2 (P_{11} P_{22} - P_{12} P_{21})}{\Delta(s)}$$

der

$$\Delta(s) := (1 + K_1 P_{11})(1 + K_2 P_{22}) - K_1 K_2 P_{12} P_{21}.$$

Hvis krysskoblingene P_{12} og P_{21} ikke er små, så vil vekselvirkningene kunne gi flere «skjulte» tilbakekoblinger som illustrert figur 13.4. Generelt for et MIMO-system med n pådrag og n utganger er det faktisk $n!$ mulige tilbakekoblinger!



Figur 13.4: Blokkdiagram for 2×2 MIMO-system med «skjult» tilbakekobling.

For det enkle 2×2 -tilfellet i figur 13.3 kan vi observere følgende:

Observasjoner om vekselvirkninger for systemet i figur 13.3:

- Endrer vi referanse i en tilbakekoblingssløyfe vil både y_1 og y_2 endre seg siden G_{12} og G_{21} begge er forskjellig fra 0 hvis P_{12} og P_{21} er det.
- Stabiliteten til systemet er gitt av $\Delta(s)$ siden denne er nevneren i alle overføringsfunksjonene; denne avhenger av begge regulatorene, K_1 og K_2 , samt alle fire prosess-overføringsfunksjonene $P_{11}, P_{12}, P_{21}, P_{22}$.
- I det spesielle tilfellet $P_{12} = 0$ og/eller $P_{21} = 0$ får vi

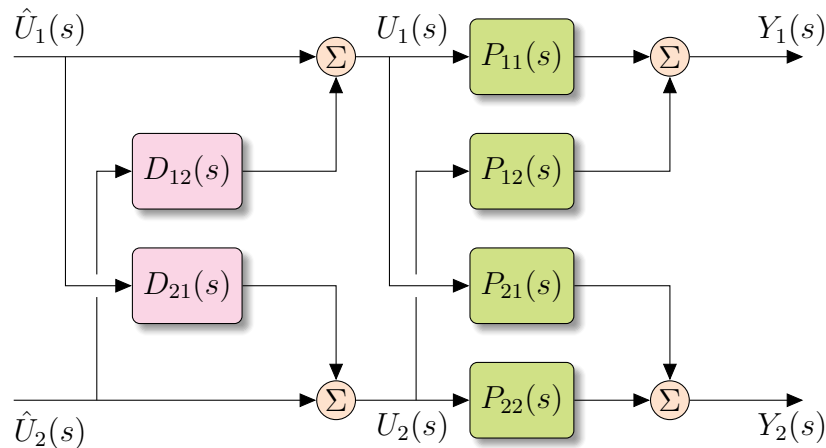
$$\Delta(s) = (1 + K_1 P_{11})(1 + K_2 P_{22}), \tag{13.4}$$

slik at stabiliteten til systemet da kun er avhengig av den individuelle stabiliteten til begge de to tilbakekoblingssløyvene.

13.2. Dekobling ▶ 1NvF-3HGUa4

Alternative kilder: §18.5 i [Seborg et al., 2016]

Nåværende problem: Konstruere et sett med **dekoblingsfiltre** som ideelt sett skal fjerne alle krysskoblingene i prosessen. Dette vil i så fall gjøre at vi kan designe en regulator hvor hvert del-system individuelt ved hjelp av for eksempel en egnet metode fra Del IV.



Figur 13.5: Blokkdiagram for 2×2 system med dekoblingsfilter.

La oss igjen fundere litt rundt 2×2 -systemet vi har sett på tidligere, men nå med et part dekoblingsfilter, $D_{12}(s)$ og $D_{21}(s)$, koblet til som vist i figur 13.5. For å få til vårt ønskede mål om ingen krysskoblinger, må følgende holde:

Ideelle dekoblingsfiltre for 2×2 -systemer: Et ideelt dekoblingsfilter for systemet vist i figur 13.5 er gitt ved følgende overføringsfunksjoner:

$$D_{12}(s) = -\frac{P_{12}(s)}{P_{11}(s)} \quad \text{og} \quad D_{21}(s) = -\frac{P_{21}(s)}{P_{22}(s)}. \quad (\text{Ideelle dekoblingsfiltre})$$

Disse er realiserbare hvis de er proper (se § 2.6.4), ikke har poler i høyre halvplan, og ikke krever invertering av en tidsforsinkelse.

De ideelle filtrene fører til følgende to separate monovariabel systemer:

$$Y_1(t) = G_1(s)\hat{U}_1(s) =: P_{11}(s)(1 - G_{\times}(s))\hat{U}_1(s)$$

$$Y_2(t) = G_2(s)\hat{U}_2(s) =: P_{22}(s)(1 - G_{\times}(s))\hat{U}_2(s)$$

hvor

$$G_{\times}(s) := \frac{P_{12}(s)P_{21}(s)}{P_{11}(s)P_{22}(s)} \quad (\text{Krysskoblingsgrad})$$

er **krysskoblingsgraden** i systemet.

Selv med slike filtre dukker det opp et viktig spørsmål:

Hvordan stille inn regulatorene gitt slike filtre? Hvis vi har laget et fungerende (ideelt) dekoblings-filter, så er det uansett viktig å ta høyde for at regulatorene blir påvirket av hverandre uansett. Innstilling av regulatorene må derfor som oftest gjøres som en iterativ prosedyre, hvor man bytter mellom å etterjustere de forskjellige regulatorene.

Når kan vi ikke bruke ideelle dekoblingsfiltere?

Et ideelt dekoblingsfilter kan **ikke** realiseres hvis

1. Det har ledd av typen $e^{\theta s}$ (filteret er en prediktor, og dermed anti-kausalt);
2. Det ikke er en proper overføringsfunksjon (høyere orden i teller enn i nevner);
3. Det har poler i høyre halvplan (ustabilt filter).

Hva kan vi gjøre hvis det ikke er realiserbart? Det ideelle filteret må da tilnærmes ved hjelp av tilsvarende metoder som i § 8.1.5, for eksempel kan man gjøre én eller en kombinasjon av følgende:

1. Bruke en statisk tilnærming (kan være problematisk i transientfasen);
2. Fjerne/tilnærme positive tidsforsinkelser fra transferfunksjonen;
3. Kombinere nullpunkt (f.eks. lage et *lead-lag*-element, hvor tidskonstanten til *lead*-elementet er summen av de sammenslåtte nullpunktene).

Eksempel 13.1. Gitt en 2×2 prosess med følgende overføringsfunksjoner:^a

$$P_{11} = \frac{12.8e^{-s}}{1 + 16.7s}, \quad P_{12} = \frac{-18.9e^{-3s}}{1 + 21s}, \quad P_{21} = \frac{6.6e^{-7s}}{1 + 10.9s}, \quad P_{22} = \frac{-19.4e^{-3s}}{1 + 14.4s}. \quad (13.5)$$

De ideelle dekoblingsfiltrene er da gitt som:

$$D_{12} = -\frac{P_{12}}{P_{11}} = -\frac{-18.9e^{-3s}}{1 + 21s} \cdot \frac{1 + 16.7s}{12.8e^{-s}} = 1.48 \frac{1 + 16.7s}{1 + 21s} e^{-2s} \quad (13.6)$$

$$D_{21} = -\frac{P_{21}}{P_{22}} = -\frac{6.6e^{-7s}}{1 + 10.9s} \cdot \frac{1 + 14.4s}{-19.4e^{-3s}} = 0.340 \frac{1 + 14.4s}{1 + 10.9s} e^{-4s}$$

Filteret er realiserbart, men siden polene ligger så nær nullpunktene, kan det være naturlig å forsøke med en stasjonær dekopling:

$$\hat{D}_{12} = 1.48, \quad \hat{D}_{21} = 0.340 \quad (13.7)$$

^aDette tilsvarer en destillasjonskolonne med verdier tatt fra [Wood and Berry, 1973].

Merk: I følge [Seborg et al., 2016] blir dekopling sjeldent brukt i nyere anlegg i industrien siden metoden blant annet er sensitiv til usikkerhet i prosessmodellen. I stedet er det i stor grad MPC (modell-prediktiv kontroll) som brukes for multivariable systemer.

13.3. Stasjonær analyse: RFM og kondisjontall

Full dynamisk analyse av multivariable systemer kan ofte være krevende på grunn av de mange kompliserte vekselvirkningene. Et alternativ kan derfor være å se på hvilke egenskaper prosessen har

når det har nådd stasjonære tilstander. Denne analysen er raskere og enklere å utføre, og er et godt utgangspunkt for videre undersøkelser.

Som det kommer frem av navnet, så finner man ved stasjonær analyse den stasjonære responsen til et system ved å sette alle de deriverte av prosessvariablene/tilstandene lik 0 (altså ingen endring). I frekvensdomenet tilsvarer dette $s = 0$.

Ved hjelp av noen enkle stasjonære forhåndsundersøkelser av systemet, kan man potensielt få veiledende svar på spørsmål som:

- Hvordan bør reguleringsløyferne kobles?
- Vil det bli lett eller vanskelig å få til god regulering?
- Vil vi kunne oppnå et robust reguleringsystem?
- Vil vi kunne få et stabilt system med PI-regulering i alle sløyfer?

For dette trenger vi følgende:

Stasjonær-respons-modellen: Gitt en transfermatrise, $\mathbf{P}(s) \in \mathbb{C}^{p \times m}$, på formen (13.2), hvor hvert element $P_{ij}(s)$ er proper og stabilt. Den tilhørende *stasjonær-responsen-modellen*,

$$\mathbf{y}^s = \mathbf{k}\mathbf{u}^s,$$

er gitt ved den konstante *stasjonære forsterkningsmatrisen*:

$$\mathbf{k} := \lim_{s \rightarrow 0^+} \mathbf{P}(s) = \begin{bmatrix} k_{11} & k_{12} & \cdots & k_{1m} \\ k_{21} & k_{22} & \cdots & k_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ k_{r1} & k_{r2} & \cdots & k_{rm} \end{bmatrix}. \quad (13.8)$$

I resten av avsnittet skal vi se på noen enkle stasjonære metoder og starter med RFM-metoden.

13.3.1 Relativ forsterkningsmatrise (RFM-analyse)



Alternative kilder: §18.2 i [Seborg et al., 2016]

Nåværende problem: Anta at vi har et multivariabelt system, og vi vil bruke ett pådrag (inngang) til å regulere én prosessvariabel (utgang). Men hvilken inngang skal brukes på hvilken utgang? For eksempel, skal vi bruke u_1 eller u_2 til å regulere y_1 ?

Når man skal velge tilbakekoblingsløyfer for multivariable systemer, så bør det gjøres basert på hva som gir best regulering, ikke nødvendigvis hva som er intuitivt. Så hvordan kan vi avgjøre hvilke tilbakekoblinger som gir best regulering?

En mulig metode man kan bruke for å prøve å avgjøre dette er ved å analysere den *relative forsterkningsmatrisen*, abbreviert **RFM-analyse**, (eng. (Relative Gain Array (RGA)) til systemet.³ Denne metoden ble introdusert av Bristol i 1966 (se [Bristol, 1966]) som et stasjonært mål på vekselvirkninger i desentraliserte reguleringsystemer (altså én separat regulator per tilbakekoblingsløyfe).

³Merk at selv om vi skal fokusere på stasjonær-responsen (13.8), så finnes det også videreutviklinger som lar en bruke denne metoden også for å analysere dynamiske responser, altså når $\omega \neq 0$.

Idéen bak RFM-analyse: La oss ta for oss følgende hovedtrekk ved RFM-analysemetoden:

- Ta et mulig inngang-utgang par (u_j, y_i) , og anta at vi ønsker å benytte u_j til å regulere y_i ;
- For en hver individuell sløyfe eksisterer det to ekstreme tilfeller:
 1. Alle andre sløyfer er åpne: $u_k = 0, \forall k \neq j$
 2. Alle andre sløyfer er lukket med perfekt regulering: $y_k = 0, \forall k \neq i$
- For å forenkle prosedyren ser man ofte kun på stasjonære situasjoner, og benytter derfor den stasjonære forsterkningsmatrisen $\mathbf{k} = \mathbf{P}(0)$.
- Vi ønsker å finne forholdet/ratioen mellom disse to ekstreme tilfelle:

$$\lambda_{ij} = \frac{\text{Forsterkning (fra } u_j \text{ til } y_i) \text{ med alle andre sløyfer åpne}}{\text{Forsterkning (fra } u_j \text{ til } y_i) \text{ med alle andre sløyfer lukket}} = \frac{\left(\frac{\partial Y_i}{\partial U_j}\right) \Big|_{U_k=0, k \neq j}}{\left(\frac{\partial Y_i}{\partial U_j}\right) \Big|_{Y_k=0, k \neq i}}$$

- Matrisen $\mathbf{\Lambda}$ med elementer λ_{ij} er den **relative forsterkningsmatrisen**.

Vi tar for oss et eksempel for å se hvordan dette gjøres:

Eksempel 13.2. (Eks. fra [Skogestad and Postlethwaite, 2007]) Gitt følgende 2×2 -system:

$$\begin{aligned} Y_1 &= P_{11}(s)U_1 + P_{21}(s)U_2, \\ Y_2 &= P_{21}(s)U_1 + P_{22}(s)U_2. \end{aligned}$$

Vårt mål er å kontrollere y_1 ved hjelp av u_1 . Ved å følge prosedyren over, starter vi med å anta at $U_2 = 0$ i den første ligningen, noe som gir oss

$$Y_1 = P_{11}(s)U_1.$$

Vi antar så perfekt regulering i den andre sløyfen, og setter $y_2 = 0$ i den andre ligningen, noe som da gir oss

$$0 = P_{21}(s)U_1 + P_{22}(s)U_2 \implies U_2 = -(P_{21}(s)/P_{22}(s))U_1.$$

Ved å sette dette inn i den første ligningen får vi dermed

$$Y_1 = \underbrace{\left(P_{11}(s) - P_{12}(s) \frac{P_{21}(s)}{P_{22}(s)} \right)}_{=: \hat{P}_{11}(s)} U_1.$$

Dette betyr at vi endrer koblingen fra $P_{11}(s)$ til $\hat{P}_{11}(s)$ når vi lukker den andre sløyfen, slik at den tilsvarende stasjonære forsterkningen endres fra $k_{11} = P_{11}(0)$ til $\hat{P}_{11}(0)$. Den stasjonære *relative forsterkningen* mellom disse er dermed

$$\lambda_{11} = \lim_{s \rightarrow 0^+} \frac{P_{11}(s)}{\hat{P}_{11}(s)} = \lim_{s \rightarrow 0^+} \frac{P_{11}(s)P_{22}(s)}{P_{11}(s)P_{22}(s) - P_{12}(s)P_{21}(s)} = \frac{k_{11}k_{22}}{k_{11}k_{22} - k_{12}k_{21}}.$$

Hvert element i den relative forsterkningsmatrisen forteller oss dermed om det spesifikke inngang-utgang paret u_j og y_i vil føre til god regulering. Analysen lar oss derfor velge tilbakekoblingsløyper slik at de aktuelle λ_{ij} er mest optimale.

Definisjon og huskereglar:

Relativ forstærkningsmatrise (RFM): ([Skogestad and Postlethwaite, 2007]) Gitt en kvadratisk ($m \times m$), ikke-singular (inverterbar) kompleks matrise $\mathbf{k} \in \mathbb{C}^{m \times m}$, så er den tilhørende *relative forsterkningsmatrisen* (RFM), $\mathbf{\Lambda} \in \mathbb{C}^{m \times m}$, gitt ved ^a

$$\mathbf{\Lambda}(\mathbf{k}) := \mathbf{k} \circ (\mathbf{k}^{-1})^\top = \begin{bmatrix} \lambda_{11} & \lambda_{12} & \cdots & \lambda_{1m} \\ \lambda_{21} & \lambda_{22} & \cdots & \lambda_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{m1} & \lambda_{m2} & \cdots & \lambda_{mm} \end{bmatrix}. \tag{13.9}$$

Her bruker vi \circ til å betegne det elementvise- (Hadamard-) produktet, altså $(A \circ B)_{ij} = A_{ij} \cdot B_{ij}$ for to matriser $A, B \in \mathbb{C}^{m \times m}$ (hvor subscriptet ij betyr rad i og kolonne j); MATLAB: $\mathbf{A}.*\mathbf{B}$.

^a**Utleddning:** For et gitt par (u_j, y_i) , så er overføringsfunksjonen fra $U_j(s)$ til $Y_i(s)$ gitt ved $P_{ij}(s)$ når alle andre sløyfer er åpne/skrudd av ($u_k = 0, \forall k \neq j$). Anta så at alle andre sløyfer er lukket, med perfekt regulering, slik at $y_k = 0, \forall k \neq i$. Siden $\mathbf{U}(s) = \mathbf{P}^{-1}(s)\mathbf{Y}(s)$ og $Y_k = 0, \forall k \neq i$, så har vi $U_j(s) = (P^{-1})_{ji}(s)Y_i(s)$, slik at overføringsfunksjonen fra $U_j(s)$ til $Y_i(s)$ da er gitt ved $1/(P^{-1})_{ji}(s)$. Det åpne scenarioet delt på det lukkede når $s \rightarrow 0$ er dermed $\lambda_{ij} = \lim_{s \rightarrow 0^+} \frac{P_{ij}(s)}{(1/(P^{-1})_{ji}(s))} = \lim_{s \rightarrow 0^+} P_{ij}(s)(P^{-1})_{ji}(s) = \lim_{s \rightarrow 0^+} P_{ij}(s)((P^{-1})_{ij}(s))^\top = \mathbf{k}_{ij}((\mathbf{k}^{-1})_{ij})^\top$, som ved å samle alle kombinasjonene i en matrise gir (13.9).

Legg merke til følgende egenskaper til $\mathbf{\Lambda}(\mathbf{k})$:

- Alle dens rader og kolonner summerer til 1;
- $\mathbf{\Lambda}(\mathbf{k})$ er identitetsmatrisen hvis \mathbf{k} er øvre eller nedre **triangulær**;
- Den er uavhengig av inngangs- og utgangsskalering.

Så hvordan kan vi bruke denne til å fastslå hvilken inngang som skal brukes til å regulere hvilken utgang? Dette kan gjøres ut fra følgende tabell:

Tolkning av den relative forsterkningsmatrisen: (for stasjonære situasjoner)	
$\lambda_{ij} = 1$	Ideelt tilfelle, med ingen vekselvirkninger. Dermed er en forandring i u_j kun overført til y_i gjennom P_{ij} , men merk at denne sløyfen kan påvirke andre regulerte variabler.
$\lambda_{ij} = 0$	u_j har ingen effekt på y_i når de andre regulatorene er i åpen sløyfe; dermed et dårlig valg siden den kun fungerer når de andre regulatorene bidrar.
$0 < \lambda_{ij} < 1$	Den stasjonære sløyfeforsterkningen med de andre sløyfene lukket er sterkere enn den samme prosess forsterkningen med de andre sløyfene åpne.
$\lambda_{ij} < 0$	Åpen og lukket sløyfe forsterkning for dette paret har motsatt fortegn, med andre ord må fortegnet til kontrolleren settes avhengig av modusen til de andre kontrollerne. Ikke en ønskelig situasjon.
$\lambda_{ij} > 1$	Den stasjonære sløyfeforsterkningen med de andre sløyfene åpne er sterkere enn den samme prosessforsterkningen med de andre sløyfene lukket. Hvis $\lambda_{ij} > 5$ kan det være vanskelig å kontrollere utgangen med den konfigurasjonen, da de andre regulatorene overstyrer.
$\lambda_{ij} = \infty$	Den stasjonære sløyfeforsterkningen med de andre sløyfene lukket er 0, og det er dermed ikke mulig å regulere denne sløyfen.

Regler for å koble innganger og utganger: (TL;DR: $\lambda_{ij} \approx 1$ er bra, $\lambda_{ij} < 0$ bør unngås)

1. Foretrekk koblinger som er slik at λ_{ij} for den aktuelle koblingen er så nærme 1 som mulig. Med andre ord skal (evt. etter en permutasjon) den relative forsterkningsmatrisen være så nærme identitetsmatrisen som mulig.
2. Unngå så langt det er mulig koblinger der $\lambda_{ij} < 0$.

⚠ Achtung! Egentlig bør man se på den relativeforsterkningsmatrisen over flere frekvenser; mer spesifikt, bør man se på $\mathbf{P}(j\omega)$ ved viktige frekvens for reguleringsproblemet fremfor bare $\mathbf{k} = \mathbf{P}(0)$.

Eksempel 13.3. Gitt et 2×2 -system med forsterkingsmatrise

$$\mathbf{k} = \begin{bmatrix} k_{11} & k_{12} \\ k_{21} & k_{22} \end{bmatrix},$$

så er den relative forsterkings matrisen

$$\begin{aligned} \Lambda(\mathbf{k}) &= \mathbf{k} \circ (\mathbf{k}^{-1})^T = \begin{bmatrix} k_{11} & k_{12} \\ k_{21} & k_{22} \end{bmatrix} \circ \left(\frac{1}{k_{11}k_{22} - k_{12}k_{21}} \begin{bmatrix} k_{22} & -k_{21} \\ -k_{12} & k_{11} \end{bmatrix} \right) \\ &= \frac{1}{k_{11}k_{22} - k_{12}k_{21}} \begin{bmatrix} k_{11}k_{22} & -k_{12}k_{21} \\ -k_{12}k_{21} & k_{11}k_{22} \end{bmatrix}. \end{aligned}$$

Ved å definere $\lambda := \lambda_{11} = \frac{k_{11}k_{22}}{(k_{11}k_{22} - k_{12}k_{21})}$ kan vi skrive dette som

$$\Lambda(\mathbf{k}) = \begin{bmatrix} \lambda & 1 - \lambda \\ 1 - \lambda & \lambda \end{bmatrix}.$$

Merk: Siden vi for et 2×2 -system kan finne hele RFMen fra tallet λ (se eksempel 13.3), så kan vi gjøre følgende konklusjoner for denne typen systemer:

Huskeregul for 2x2-systemer: Man bør bruke u_1 for å regulere y_1 bare hvis $\lambda \geq 0.5$; ellers bør man bruke u_2 .

Eksempel 13.4. Beregne relative forsterkningsmatrise ved hjelp av MATLAB: La den statiske forsterkning til en prosess, $\mathbf{k} = \lim_{s \rightarrow 0} \mathbf{P}(s)$, være lik

$$\mathbf{k} = \begin{bmatrix} 12.8 & -18.9 \\ 6.6 & -19.4 \end{bmatrix}$$

Dette tilsvarer systemet i eksempel 13.1. Den relative forsterkningsmatrisen beregnes enklest i MATLAB ved å benytte formel (13.9), altså $\Lambda = \mathbf{k} \circ (\mathbf{k}^{-1})^T$. Dette er vist kodesnutt 13.1, og resulterer i

$$\Lambda = \begin{bmatrix} 2.0094 & -1.0094 \\ -1.0094 & 2.0094 \end{bmatrix}.$$

Kodesnutt 13.1: Utregning av RFM i MATLAB.

```
k = [12.8 , -18.9; 6.6 , -19.4];
```

```
Lambda = k.*(pinv(k)).'
% hvor .*= betyr elementvis multiplikasjon,
% mens .' er transponering (evt. kan ''transpose'' brukes).
```

Eksempel 13.5. ([Seborg et al., 2016]) Anta en 4×4 prosessmodell slik at den stasjonære forsterkningsmatrisens $\mathbf{k} \in \mathbb{R}^{4 \times 4}$ RFM er gitt som

$$\mathbf{\Lambda}(\mathbf{k}) = \begin{bmatrix} 0.931 & 0.150 & 0.080 & -0.164 \\ -0.011 & -0.429 & 0.286 & 1.154 \\ -0.135 & 3.314 & -0.270 & -1.910 \\ 0.215 & -2.030 & 0.900 & 1.919 \end{bmatrix}.$$

Vi ønsker nå å velge tilbakekoblinger for dette systemet basert på verdiene av λ_{ij} .

Ifølge reglene over, kan vi umiddelbart eliminere alle koblinger der $\lambda_{ij} < 0$:

$$\mathbf{\Lambda} = \begin{bmatrix} 0.931 & 0.150 & 0.080 & \\ & & 0.286 & 1.154 \\ & 3.314 & & \\ 0.215 & & 0.900 & 1.919 \end{bmatrix}.$$

Vi sitter nå igjen med kun en mulig kobling for y_3 og velger derfor denne:

$$\mathbf{\Lambda} = \begin{bmatrix} 0.931 & 0.150 & 0.080 & \\ & & 0.286 & 1.154 \\ & \boxed{3.314} & & \\ 0.215 & & 0.900 & 1.919 \end{bmatrix}.$$

De resterende koblingene følger deretter av regel nummer 1, velge $\lambda_{ij} \approx 1$:

$$\mathbf{\Lambda} = \begin{bmatrix} \boxed{0.931} & & & \\ & \boxed{3.314} & & \\ & & \boxed{0.900} & \\ & & & \boxed{1.154} \end{bmatrix}.$$

Dette gir oss følgende par-koblinger:

$$(u_1, y_1), \quad (u_4, y_2), \quad (u_2, y_3), \quad (u_3, y_4).$$

Det er viktig å huske at RFM-analysen vi har gjort kun er basert på den stasjonære prosessen. Følgende eksempel demonstrerer visse begrensninger med dette.

Eksempel 13.6. Hva med dynamiske betraktninger? ([Seborg et al., 2016]) For følgende prosess

$$\mathbf{P}(s) = \begin{bmatrix} \frac{-2e^{-s}}{10s+1} & \frac{1.5e^{-s}}{s+1} \\ \frac{1.5e^{-s}}{s+1} & \frac{2e^{-s}}{10s+1} \end{bmatrix}$$

er den stasjonære forsterkningsmatrisen lik

$$\mathbf{k} = \begin{bmatrix} -2 & 1.5 \\ 1.5 & 2 \end{bmatrix}.$$

Siden vi har $\lambda = -4/(-4 - 9/4) = 16/25 = 0.64$ fra formelen i eksempel 13.3, så får vi dermed RFMen

$$\mathbf{\Lambda} = \begin{bmatrix} 0.64 & 0.36 \\ 0.36 & 0.64 \end{bmatrix}.$$

Våre tidligere konklusjoner i anbefaling 13.3.1 sier dermed at vi bør regulere y_1 med u_1 og y_2 med u_2 siden $\lambda \geq 0.5$. Dette er jo dog basert på den stasjonære responsen. Legg derimot merke til at tidskonstanten på diagonalen til $\mathbf{P}(s)$ er ti ganger større enn på anti-diagonalen. Med andre ord, så reagerer jo y_1 betraktelig raskere på u_2 enn på u_1 , noe som er i konflikt med hva vi fant fra den stasjonære RFM-analysen.

RFM-analyse har dog også noen andre potensielle svakheter:

Eksempel 13.7. (Sensitiviteten til RFM-analyse) En 2×2 prosess har stasjonær forsterkningsmatrise

$$\mathbf{k} = \begin{bmatrix} 1 & k_{12} \\ 10 & 1 \end{bmatrix}$$

Vi ønsker å undersøke hvordan egenskapene til prosessen endrer seg når k_{12} endres.

Case 1 ($k_{12} = 0$): Vi begynner med å anta at $k_{12} = 0$:

$$\mathbf{k} = \begin{bmatrix} 1 & 0 \\ 10 & 1 \end{bmatrix}.$$

Siden $\det(\mathbf{k}) = 1 \cdot 1 - 10 \cdot 0 = 1$, så er ikke matrisen singular. Det er også tydelig at \mathbf{k} er en nedre triangulær matrise, og dermed $\mathbf{\Lambda}(\mathbf{k}) = \mathbf{I}$. La oss kontrollere dette ved å direkte regne ut krysskoblingsgraden (se (Krysskoblingsgrad)):

$$\chi_s = \frac{k_{12}k_{21}}{k_{11}k_{22}} = 0 \quad \Rightarrow \quad \lambda_{11} = \frac{1}{1 - \chi_s} = 1.$$

Uten noen krysskobling er tilbakekoblingene dermed ($y_1 \rightarrow u_1$) og ($y_2 \rightarrow u_2$).

Case 2 ($k_{12} = 0.2$): Gjør nå en endring i systemet og setter $k_{12} = 0.2$ slik at

$$\mathbf{k} = \begin{bmatrix} 1 & 0.2 \\ 10 & 1 \end{bmatrix}$$

Beregner først stasjonær krysskoblingsgrad for å finne $\mathbf{\Lambda}(\mathbf{k})$:

$$\chi_s = \frac{k_{12}k_{21}}{k_{11}k_{22}} = 2 \quad \Rightarrow \quad \lambda_{11} = \frac{1}{1 - \chi_s} = -1$$

Dermed blir RFMen:

$$\mathbf{\Lambda}(\mathbf{k}) = \begin{bmatrix} -1 & 2 \\ 2 & -1 \end{bmatrix}$$

Tilbakekoblingene blir ($y_2 \rightarrow u_1$) og ($y_1 \rightarrow u_2$), altså motsatt fra forrige case!

Eksempelen over viser at resultatene fra RFM-analyse kan være sensitive med tanke på modellusikkerhet. Den neste metoden vi skal se på kan blant annet brukes som et mål på denne sensitiviteten

13.3.2 Singulærverdi-analyse og Kondisjonstall



Alternative kilder: §18.3 i [Seborg et al., 2016]

Nåværende problem: Finne ut hvor mye vil utgangen \mathbf{y} vil endre seg når inngangen \mathbf{u} endres en infinitesimal (svært, svært liten) mengde

En måte å måle dette på er ved hjelp av **kondisjonstallet** κ til den stasjonære forsterkningsmatrisen \mathbf{k} (se (13.8)).

Hva er et kondisjonstall? Og hva brukes det til? Gitt et statisk system

$$\mathbf{y} = \mathbf{k}\mathbf{u},$$

hvor $\mathbf{k} \in \mathbb{R}^{n \times n}$ er en ikke-singulær matrise, og hvor \mathbf{y} og \mathbf{u} er henholdsvis utgangen og inngangen. Vi bruker κ til å betegne kondisjonstallet, som er definert som den største *singulærverdien* til \mathbf{k} delt på den minste singulærverdien (vi skal definere singulærverdier om litt). Fra dette tallet kan man trekke følgende konklusjoner:

- i)* Hvis κ er liten (tilnærmet lik 1), så vil en hver endring i \mathbf{u} føre til tilnærmet lik endring i \mathbf{y} ;
- ii)* Hvis κ er stor (mye større enn én), så vil den relative endring i \mathbf{y} kunne variere stort for forskjellige (retnings-) endringer i \mathbf{u} .

En slik sensitivitet i utgangen til endringer i inngangen kan vi finne ved å regne ut

$$\|\mathbf{y}\|^2 = \mathbf{y}^T \mathbf{y} = y_1^2 + y_2^2 + \dots + y_n^2 = \mathbf{u}^T \mathbf{k}^T \mathbf{k} \mathbf{u},$$

hvor da $\|\cdot\|$ er den **Euklidiske vektornormen**. Siden det må være tilfelle at $\mathbf{u}^T \mathbf{k}^T \mathbf{k} \mathbf{u} = \|\mathbf{k}\mathbf{u}\|^2 \geq 0$, så må den symmetriske og kvadratiske matrisen $\mathbf{k}^T \mathbf{k}$ bare ha reelle, positive egenverdier. Disse egenverdiene tilsvarer igjen kvadratet av singulærverdiene til \mathbf{k} : $\sigma_{\min}^2 = \sigma_1^2 \leq \sigma_2^2 \leq \dots \leq \sigma_n^2 = \sigma_{\max}^2$. Vi har derfor

$$\sigma_{\min}^2 \|\mathbf{u}\|^2 \leq \|\mathbf{y}\|^2 \leq \sigma_{\max}^2 \|\mathbf{u}\|^2 \quad \implies \quad \sigma_{\min} \leq \frac{\|\mathbf{y}\|}{\|\mathbf{u}\|} \leq \sigma_{\max},$$

hvorfra konklusjonene i *i)* og *ii)* følger siden $\kappa := \sigma_{\max}/\sigma_{\min}$.

Eksempel 13.8. (Kondisjonstall for sensitivtetsanalyse) I eksempel 13.7 så vi at RFM-analyse kunne være sensitiv til usikkerhet i prosessmodellen. La oss derfor se om kondisjonstallet kan hjelpe oss med å fange opp dette.

La den stasjonære forsterkningen være gitt som i Case 1 i eksempel 13.7:

$$\mathbf{k} = \begin{bmatrix} 1 & 0 \\ 10 & 1 \end{bmatrix}.$$

Før vi finner singulærverdiene, så regner vi først ut egenverdiene til \mathbf{k} (braker α for å unngå konflikt med den relative forsterkningen λ_{11}) til å være $\alpha_1 = \alpha_2 = 1$, noe som følger fra:

$$\det(\mathbf{k} - \alpha I) = |k - \alpha I| = \begin{vmatrix} 1 - \alpha & 0 \\ 10 & 1 - \alpha \end{vmatrix} = (1 - \alpha)^2 = 0.$$

Fra dette ser jo alt bra ut, i form av at vi får lik «forsterkning» i retningene til begge egenvektorene.

La oss nå i stedet beregne singularverdiene ved å ta utgangspunkt i

$$\mathbf{k}^T \mathbf{k} = \begin{bmatrix} 1 & 10 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 10 & 1 \end{bmatrix} = \begin{bmatrix} 101 & 10 \\ 10 & 1 \end{bmatrix}.$$

Eigenverdiene til $\mathbf{k}^T \mathbf{k}$ er $\hat{\alpha}_1 \approx 101.99$ og $\hat{\alpha}_2 \approx 0.01$, noe som følger fra

$$\det(\mathbf{k}^T \mathbf{k} - \hat{\alpha} I) \Rightarrow (101 - \hat{\alpha})(1 - \hat{\alpha}) - 100 = 0.$$

Singularverdiene blir derfor

$$\sigma_1 \approx \sqrt{101.99} \approx 10.1 \quad \text{og} \quad \sigma_2 \approx \sqrt{0.01} = 0.1$$

som betyr at kondisjonstallet til \mathbf{k} er lik

$$\kappa = \frac{\sigma_1}{\sigma_2} \approx \frac{10.1}{0.1} = 101$$

Den statiske prosessmodellen gitt av \mathbf{k} er derfor ekstremt sensitiv til endring i inngangsverdiene, og systemet har mest sannsynlig store reguleringsproblemer slik som eksempel 13.7 halveis hintet til.

Hvorfor er dette nyttig for oss? Som vi så i forrige avsnitt, så er analyse ved hjelp av den relative forsterkningsmatrisen (RFM) nyttig for å bestemme hvilke tilbakekoblinger som gir best regulering. Dette forutsetter dog at man allerede har bestemt systemets innganger og utganger. MIMO-systemer kan ha mange utganger og innganger, og det er viktig å bestemme hvilke av disse utgangene som skal reguleres og hvilke pådrag som skal benyttes. For eksempel kan flere av utgangene være så nært knyttet til hverandre at det ikke er mulig å regulere de separat. RFM-analyse kan derimot **ikke** hjelpe oss med å bestemme hverken dette eller hvor vanskelig systemet er å regulere. Vi vil derfor nå se på et nytt analyseverktøy: singularverdi-analyse.

Singularverdi-analyse (SVA) er et kraftig analytisk verktøy som blant annet kan benyttes til å:

- Velge regulerte, målte og manipulerte variabler;
- Bestemme den beste konfigurasjonen for multisløyferegulering;
- Evaluere robustheten til en foreslått reguleringsstrategi.

Merk: Som med RFM-analysen, så vil vi kun benytte SVA basert på den statiske prosessmodellen til systemet (stasjonær analyse).

Singularverdier er tett knyttet til egenverdier (se vedlegg B.3.2). Før vi kommer til SVA, skal vi derfor først se på hvordan man kan analysere statiske prosessmodeller ved å se på egenverdiene til forsterkningsmatrisen.

Eigenverdi-analyse for prosessmodeller:

Gitt følgende statisk prosessmodell:

$$\mathbf{Y}^s(s) = \mathbf{k} \mathbf{U}^s(s) = \begin{bmatrix} k_{11} & k_{12} & \cdots & k_{1m} \\ k_{21} & k_{22} & \cdots & k_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ k_{n1} & k_{n2} & \cdots & k_{nm} \end{bmatrix} \begin{bmatrix} U_1 \\ U_2 \\ \vdots \\ U_m \end{bmatrix}. \quad (13.10)$$

Denne er ekvivalent med

$$\begin{aligned} Y_1 &= k_{11}U_1 + k_{12}U_2 + \dots + k_{1m}U_m \\ Y_2 &= k_{21}U_1 + k_{22}U_2 + \dots + k_{2m}U_m \\ &\vdots \\ Y_n &= k_{n1}U_1 + k_{n2}U_2 + \dots + k_{nm}U_m \end{aligned} \tag{13.11}$$

hvor k_{ij} er den stasjonære forsterkningen til overføringsfunksjonen P_{ij} (se (13.8)).

Det er ønskelig å undersøke om noen av utgangene i systemet er avhengig av hverandre, det vil si at man ikke kan regulere den ene uavhengig av den andre. Dette tilsvarer at noen av likningene som beskriver utgangene er lineært avhengige. For eksempel er utgangene y_i og y_j lineært avhengige hvis det eksisterer a_1 og a_2 slik at

$$a_1 Y_i + a_2 Y_j = 0. \tag{13.12}$$

Dette tilsvarer

$$a_1(k_{i1}U_1 + k_{i2}U_2 + \dots + k_{im}U_m) + a_2(k_{j1}U_1 + k_{j2}U_2 + \dots + k_{jm}U_m) = 0 \tag{13.13}$$

Dette kan vi igjen skrive mer kompakt som

$$a_1 k_i \mathbf{U} + a_2 k_j \mathbf{U} = 0,$$

som tilsvarer $k_i = -\frac{a_2}{a_1} k_j$. Man må altså undersøke om to av radvektorene i \mathbf{k} er lineært avhengige; Men dette er det samme som å sjekke om \mathbf{k} er singular!

Eigenverdi-analyse: Kan benyttes til å undersøke om prosessen kan reguleres på ønsket måte:

- Utgangene y_i og y_j kan ikke reguleres uavhengig av hverandre hvis de er lineært avhengige;
- De er lineært avhengige hvis \mathbf{k} er singular;
- \mathbf{k} er singular hvis $\det(\mathbf{k}) = 0$;
- $\det(\mathbf{k}) = 0$ hvis en eller flere av egenverdiene til \mathbf{k} er lik 0.

Konsekvens av egenverdier:

- Hvis \mathbf{k} -matrisen er singular vil det være svært vanskelig å regulere prosessen, uansett hvordan tilbakekoblingsløyene velges;
- Hvis en av egenverdiene er veldig liten relativ til de andre, vil det kreve veldig store endringer i en eller flere av pådragene for å regulere prosessen (systemet kalles *stivt*).

Singularverdier og Singularverdi-dekomposisjon

Som tidligere nevnt, så er singularverdier nært beslektet med egenverdier. Faktisk kan de til en viss grad betraktes som en generalisering av konseptet:

Singularverdi: Et ikke-negativt, reelt tall σ er en *singularverdi* til en (ikke nødvendigvis kvadratisk) potensielt kompleks matrise $\mathbf{k} \in \mathbb{C}^{m \times n}$ hvis det finnes to vektorer, $v \in \mathbb{C}^n$ og $w \in \mathbb{C}^m$, med lengde lik én, altså $\|v\| = \|w\| = 1$, slik at $\mathbf{k}v = \sigma w$ og $\mathbf{k}^T w = \sigma v$.

Legg her merke til at:

1. En «inngang» i retningen v fører til (via \mathbf{k}) en utgang i retningen w skalert med σ .

2. De $p \leq \min(m, n)$ singularverdier, $\sigma_1, \sigma_2, \dots, \sigma_p$ til matrisen $\mathbf{k} \in \mathbb{C}^{m \times n}$ er ikke-negative reelle tall slik at $\sigma_i = \sqrt{\alpha_i}$ der α_i er egenverdiene til $\mathbf{k}^T \mathbf{k}$.⁴
3. Hvis $\mathbf{k} \in \mathbb{R}^{n \times n}$ er kvadratisk og symmetrisk med (reelle) egenverdier $\alpha_1, \dots, \alpha_n$, så er $\sigma_i = |\alpha_i|$.

Eksempel 13.9. Matrisen

$$\mathbf{k} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 3 \end{bmatrix}$$

har singularverdierne $\sigma_1 = 3$ og $\sigma_2 = 1$ siden

$$\underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 3 \end{bmatrix}}_{\mathbf{k}} \underbrace{\begin{bmatrix} 0 \\ 0 \\ -1 \end{bmatrix}}_{v_1} = \underbrace{3}_{\sigma_1} \underbrace{\begin{bmatrix} 0 \\ -1 \end{bmatrix}}_{w_1} \quad \text{og} \quad \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 3 \end{bmatrix}}_{\mathbf{k}^T} \underbrace{\begin{bmatrix} 0 \\ -1 \end{bmatrix}}_{w_1} = \underbrace{3}_{\sigma_1} \underbrace{\begin{bmatrix} 0 \\ -1 \end{bmatrix}}_{v_1},$$

samt

$$\underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 3 \end{bmatrix}}_{\mathbf{k}} \underbrace{\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}}_{v_2} = \underbrace{1}_{\sigma_2} \underbrace{\begin{bmatrix} 1 \\ 0 \end{bmatrix}}_{w_2} \quad \text{og} \quad \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 3 \end{bmatrix}}_{\mathbf{k}^T} \underbrace{\begin{bmatrix} 1 \\ 0 \end{bmatrix}}_{w_2} = \underbrace{1}_{\sigma_2} \underbrace{\begin{bmatrix} 1 \\ 0 \end{bmatrix}}_{v_2},$$

Singularverdier er sentrale i lineær algebra; de kan blant annet brukes til å tolke hvor nærme en matrise er til å være singular, og dermed (for vårt formål) benyttes til å avsløre hvor sensitiv utgangen til en matrise er når inngangen endres.

Et annet viktig faktum relatert til singularverdier er følgende: Alle matriser kan gjenskapes ved tre elementære operasjoner: en rotering, fulgt av en skalering, fulgt av en ny rotasjon. Her er skaleringen og rotasjonene relatert til henholdsvis singularverdierne og de tilsvarende enhetsvektorene via en såkalt singularverdi-dekomposisjon (SVD). SVD kan gjøres for alle matriser (reelle og komplekse, kvadratiske og rektangulære), og kan ses på som en generalisering av konseptet diagonalisering (se vedlegg B.3.2).

Singularverdi-dekomposisjon (SVD): Enhver (ikke nødvendigvis kvadratisk) matrise^a $\mathbf{k} \in \mathbb{R}^{n \times m}$ kan dekomponeres som produktet

$$\mathbf{k} = W \Sigma V^T \quad (\text{Singularverdidekomposisjon})$$

der $W \in \mathbb{R}^{n \times n}$ og $V \in \mathbb{R}^{m \times m}$ er ortogonale matriser, slik at

$$W W^T = I \quad \text{og} \quad V V^T = I,$$

mens $\Sigma \in \mathbb{R}^{n \times m}$ er en *rektangulær-diagonal* matrise som inneholder singularverdierne til \mathbf{k} .

MATLAB-kommando: `svd`

^aDette holder også for komplekse matriser, $\mathbf{k} \in \mathbb{C}^{n \times m}$, hvor da W og V også må være komplekse.

Eksempel 13.10. (Reell 2x2 matrise; [Skogestad and Postlethwaite, 2007]) Singularver-

⁴Merk at egenverdiene ikke er negative, altså $\alpha_i \geq 0$, siden $z^T \mathbf{k}^T \mathbf{k} z = \|\mathbf{k}z\|^2 \geq 0$ for alle $z \in \mathbb{R}^m$.

didekomposisjonen til en reell 2×2 matrise \mathbf{k} vil ha følgende form:

$$\mathbf{k} = W\Sigma V^T = \begin{bmatrix} \cos(\theta_1) & -\sin(\theta_1) \\ \sin(\theta_1) & \cos(\theta_1) \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix} \begin{bmatrix} \cos(\theta_2) & \pm \sin(\theta_2) \\ -\sin(\theta_2) & \pm \cos(\theta_2) \end{bmatrix}$$

hvor σ_1 og σ_2 er singularverdiene til \mathbf{k} , og hvor også θ_1 og θ_2 avhenger av matrisen \mathbf{k} .

SVD har en rekke bruksområder, blant annet innenfor maskinlæring, signalprosessering og algoritme-effektivisering (beregning av pseudoinvers), og

Kondisjonstall

Singularverdi-analyse av et system utføres ved å tolke prosessens *kondisjonstall* (eng.: condition number):

Kondisjonstallet $\kappa(\mathbf{k})$ til en matrise $\mathbf{k} \in \mathbb{R}^{n \times m}$ er definert som

$$\kappa(\mathbf{k}) := \frac{\sigma_{\max}}{\sigma_{\min}} \quad (\text{Kondisjonstallet})$$

der σ_{\max} og σ_{\min} er henholdsvis den største og den minste singularverdien til \mathbf{k} .

MATLAB-kommando: `cond`

Merk at:

1. κ er alltid større eller like én siden $\sigma_{\max} \geq \sigma_{\min} \geq 0$;
2. κ er ikke definert («lik» ∞) hvis $\sigma_{\min} = 0$;
3. κ er alltid veldefinert hvis \mathbf{k} er kvadratisk ($m = n$) og ikke er singular. Man har da $\sigma_{\min}(\mathbf{k}) = 1/\sigma_{\max}(\mathbf{k}^{-1})$ og dermed $\kappa(\mathbf{k}) = \sigma_{\max}(\mathbf{k})\sigma_{\max}(\mathbf{k}^{-1})$.

Tommelfingerregel for tolkning av kondisjonstallet:

- Systemet vil være robust og lett å regulere hvis κ er liten (1-4);
- Systemet vil være ømfintlig og vanskelig å regulere hvis κ er stor (over 25).

Skalering av pådrag: kondisjonstall-analyse er (i motsetning til RFM-analyse) avhengig av signalstørrelsene. Det er derfor vanlig å skalere pådragene før man beregner \mathbf{k} . Normalt sett gjøres dette ved å normalisere hver inngang, altså

$$u_i^* = \frac{u_i}{u_i^{\max} - u_i^{\min}},$$

hvor u_i^{\max} og u_i^{\min} henholdsvis er den største og minste verdien u_i kan ta.

Kondisjonstallet kan noen ganger gi noen litt rare svar på grunn av skalering:

Eksempel 13.11. Gitt en stasjonær prosess med

$$\mathbf{k} = \begin{bmatrix} 100 & 0 \\ 0 & 1 \end{bmatrix}.$$

Legg merke til at det ikke er noen krysskoblinger, slik at prosessen bør være lett å regulere. RFMen

er $\Lambda = \mathbf{I}_2$, mens kondisjontallet er $\kappa = \frac{100}{1} = 100$, noe som motsier at prosessen skal være lett å regulere. Dette motiverer det såkalte *optimale kondisjontallet*, κ^* , som er definert som

$$\kappa^*(\mathbf{k}) = \min_{Q_1, Q_2} \kappa(Q_1 \mathbf{k} Q_2) \quad (\text{Optimale kondisjontallet})$$

for to diagonale skaleringsmatriser Q_1 og Q_2 (se [Skogestad and Postlethwaite, 2007]). Ved å ta $Q_1 = \mathbf{k}^{-1}$ og $Q_2 = \mathbf{I}_2$ i eksempelet over, får man f.eks. den mye mer fornuftige verdien $\kappa^*(\mathbf{k}) = 1$.

La oss også hvordan vi kan bruke kondisjontallet til å se hvor lett et system vil være å regulere:

Eksempel 13.12. (Hvordan velge regulerte utganger) Gitt følgende statiske prosessmodell for et 3×3 system:

$$\mathbf{k} = \begin{bmatrix} 0.48 & 0.9 & -0.006 \\ 0.52 & 0.95 & 0.008 \\ 0.9 & -0.95 & 0.02 \end{bmatrix}$$

Vi ønsker å finne mulige tilbakekoblingsløyper for systemet. Vi starter med RFM-analyse:

RFM-analyse: Ved å beregne RFMen ved hjelp av MATLAB (se kodesnutt 13.2) får vi

$$\Lambda(\mathbf{k}) = \begin{bmatrix} 0.7100 & -0.1602 & 0.4501 \\ -0.3557 & 0.7925 & 0.5632 \\ 0.6456 & 0.3677 & -0.0133 \end{bmatrix}$$

slik at vi velger følgende tilbakekoblinger: $(y_3 \rightarrow u_1)$, $(y_2 \rightarrow u_2)$, $(y_1 \rightarrow u_3)$

Kondisjonstall: Beregner nå kondisjontallet som i kodesnutt 13.2. Dette gir singularveridene $\sigma_1 \approx 1.6183$, $\sigma_2 \approx 1.1434$ og $\sigma_3 \approx 0.0097$, og dermed $\kappa \approx 166.5212$. Dette kondisjontallet er meget høyt, og man vil trolig få store problemer med å regulere prosessen. Vi legger også merke til følgende:

Merk: Singularverdien σ_3 er liten relativt til de andre singularverdiene, noe som tyder på at matrisen \mathbf{k} er nær å være singular. Det kan derfor være fornuftig å undersøke om man kan oppnå bedre regulering ved å fjerne 1 inngang og 1 utgang fra reguleringsystemet (hvis dette er en mulighet for den gitte applikasjonen). Dette gjøres ved å fjerne den tilsvarende raden og kolonnen fra \mathbf{k} og beregne et nytt kondisjonstall. Dette gjentas for alle mulige kombinasjoner av 2 innganger og 2 utganger, totalt $3^2 = 9$ muligheter. Resultatet samles i en tabell hvor man får oversikt over systemet.

Kodesnutt 13.2: Utregning av RFM for eksempel 13.12

```
%% RFM: %%
k = [0.48, 0.9, -0.006; 0.52, 0.95, 0.008; 0.9, -0.95, 0.02]
Lambda = k.*(pinv(k)).'
% Resulterer i Lambda =
%    0.7100    -0.1602    0.4501
%   -0.3557     0.7925    0.5632
```

```
%    0.6456    0.3677   -0.0133

%% Singularverdier: %%
Sigma = svd(k)
% Resulterer i Sigma =
%    1.6183
%    1.1434
%    0.0097

%% Kondisjonstall %%
kappa = Sigma(1)/Sigma(end)
% Resulterer i kappa =
% 166.5212
```

*Regulering og estimering av LTI systemer i tilstandsrommet



Lineære differensialligninger

Kanoniske former

Fra [Luenberger, 1967] to come.

Styrbarhet/kontrollerbarhet og observerbarhet

Polplassering

Lineær-kvadratisk regulator (LQR)

Luenberger observeren

Kalman-filteret

Lineær-kvadratisk-Gaussian regulator (LQC)

Robust regulering

13.4. Innstilling av PID-regulatorer via frekvensanalyse*

Frekvensrespons av PID-regulatorer



*Ulineære systemer

Til nå har vi hovedsakelig sett på måter for å designe lineære regulatorer for prosesser som antas å være tilnærmet lineær. I praksis er selvfølgelig dette en overforenkling siden de fleste prosesser er unlineære, En lineær regulator vil derfor vanligvis bare gi god ytelse i et (ofte lite) nabolag av det ønskede settpunktet/banen den ble designet om. For å oppnå god regulering i et større operasjonsområde er det ofte nødvendig å planlegge eller bytte mellom lineære kontrollere designet fra lineariseringer om ulike settpunkter.

Linearisering

Flytte hit og ta med linearisering om bane?

Gain scheduling og parameterstyring

Flytte hit og ta med flere strategier og analyse?

*Lyapunov- og barriere-funksjoner

Modell-prediktiv-kontroll (MPC)

Referanser

- [Åström and Hägglund, 1984] Åström, K. J. and Hägglund, T. (1984). Automatic tuning of simple regulators with specifications on phase and amplitude margins. *Automatica*, 20(5):645–651.
- [Åström and Hägglund, 2006] Åström, K. J. and Hägglund, T. (2006). *Advanced PID control*, volume 461. ISA-The Instrumentation, Systems and Automation Society.
- [Åström and Murray, 2021] Åström, K. J. and Murray, R. M. (2021). *Feedback systems: an introduction for scientists and engineers*. Princeton university press.
- [Åström and Wittenmark, 2013] Åström, K. J. and Wittenmark, B. (2013). *Adaptive control*. Courier Corporation.
- [Balchen et al., 2016] Balchen, J. G., Andresen, T., and Foss, B. A. (2016). *Reguleringsteknikk*. NTNU, Institutt for teknisk kybernetikk.
- [Bjørvik and Hveem, 2014] Bjørvik, K. and Hveem, P. (2014). *Reguleringsteknikk*.
- [Bristol, 1966] Bristol, E. (1966). On a new measure of interaction for multivariable process control. *IEEE transactions on automatic control*, 11(1):133–134.
- [Cengel and Cimbala, 2013] Cengel, Y. and Cimbala, J. (2013). *EBOOK: Fluid Mechanics Fundamentals and Applications (SI units)*. McGraw Hill.
- [Desborough and Miller, 2002] Desborough, L. and Miller, R. (2002). Increasing customer value of industrial control performance monitoring-honeywell’s experience. In *AICHE symposium series*, number 326 in 1, pages 169–189. New York; American Institute of Chemical Engineers; 1998.
- [Dessen, 2019] Dessen, F. (2019). Optimizing order to minimize low-pass filter lag. *Circuits, Systems, and Signal Processing*, 38(2):481–497.
- [Dormand and Prince, 1980] Dormand, J. R. and Prince, P. J. (1980). A family of embedded runge-kutta formulae. *Journal of computational and applied mathematics*, 6(1):19–26.
- [Fridman, 2014] Fridman, E. (2014). *Introduction to time-delay systems: Analysis and control*. Springer.
- [Grimholt, 2018] Grimholt, C. (2018). *Optimal tuning of PID controllers*. PhD thesis, Norwegian University of Science and Technology (NTNU) Trondheim, Norway.
- [Grimholt and Skogestad, 2013] Grimholt, C. and Skogestad, S. (2013). Optimal pid-control on first order plus time delay systems & verification of the simc rules. *IFAC Proceedings Volumes*, 46(32):265–270.
- [Hägglund, 2001] Hägglund, T. (2001). The blend station—a new ratio control structure. *Control Engineering Practice*, 9(11):1215–1220.
- [Haugen, 2023] Haugen, F. A. (2023). *Modelling, Simulation and Control*. Tech Teach (<http://www.techteach.no/control>).
- [Ingimundarson and Hägglund, 2001] Ingimundarson, A. and Hägglund, T. (2001). Robust tuning procedures of dead-time compensating controllers. *Control Engineering Practice*, 9(11):1195–1208.

- [Loría, 2015] Loría, A. (2015). Observers are unnecessary for output-feedback control of lagrangian systems. *IEEE Transactions on Automatic Control*, 61(4):905–920.
- [Luenberger, 1967] Luenberger, D. (1967). Canonical forms for linear multivariable systems. *IEEE Transactions on Automatic Control*, 12(3):290–293.
- [O’duyer, 2009] O’duyer, A. (2009). *Handbook of PI and PID controller tuning rules*. World Scientific.
- [Ogata et al., 2010] Ogata, K. et al. (2010). *Modern control engineering*, volume 5. Prentice hall Upper Saddle River, NJ.
- [Olfati-Saber et al., 2007] Olfati-Saber, R., Fax, J. A., and Murray, R. M. (2007). Consensus and cooperation in networked multi-agent systems. *Proceedings of the IEEE*, 95(1):215–233.
- [Rugh and Shamma, 2000] Rugh, W. J. and Shamma, J. S. (2000). Research on gain scheduling. *Automatica*, 36(10):1401–1425.
- [Samad, 2017] Samad, T. (2017). A survey on industry impact and challenges thereof [technical activities]. *IEEE Control Systems Magazine*, 37(1):17–18.
- [Samad et al., 2020] Samad, T., Bauer, M., Bortoff, S., Di Cairano, S., Fagiano, L., Odgaard, P. F., Rhinehart, R. R., Sánchez-Peña, R., Serbezov, A., Ankersen, F., et al. (2020). Industry engagement with control research: Perspective and messages. *Annual Reviews in Control*, 49:1–14.
- [Seborg et al., 2016] Seborg, D. E., Edgar, T. F., Mellichamp, D. A., and Doyle III, F. J. (2016). *Process dynamics and control*. John Wiley & Sons, 4th edition.
- [Shamsuzzoha and Skogestad, 2010] Shamsuzzoha, M. and Skogestad, S. (2010). The setpoint overshoot method: A simple and fast closed-loop approach for pid tuning. *Journal of Process control*, 20(10):1220–1234.
- [Sharma et al., 2022] Sharma, A., Kosasih, E., Zhang, J., Brintrup, A., and Calinescu, A. (2022). Digital twins: State of the art theory and practice, challenges, and open research questions. *Journal of Industrial Information Integration*, page 100383.
- [Shinskey, 2002] Shinskey, F. G. (2002). Process control: as taught vs as practiced. *Industrial & Engineering Chemistry Research*, 41(16):3745–3750.
- [Skogestad, 2018] Skogestad, C. G. S. (2018). Should we forget the smith predictor? *IFAC-PapersOnLine*, 51(4):769–774.
- [Skogestad, 2003] Skogestad, S. (2003). Simple analytic rules for model reduction and pid controller tuning. *Journal of process control*, 13(4):291–309.
- [Skogestad and Grimholt, 2012] Skogestad, S. and Grimholt, C. (2012). The simc method for smooth pid controller tuning. In *PID control in the third millennium*, pages 147–175. Springer.
- [Skogestad and Postlethwaite, 2007] Skogestad, S. and Postlethwaite, I. (2007). *Multivariable feedback control: analysis and design*. Wiley New York, 2nd edition.
- [Taguchi and Araki, 2000] Taguchi, H. and Araki, M. (2000). Two-degree-of-freedom pid controllers—their functions and optimal tuning. *IFAC Proceedings Volumes*, 33(4):91–96.
- [Tan et al., 2003] Tan, W., Marquez, H. J., and Chen, T. (2003). Imc design for unstable processes with time delays. *Journal of process control*, 13(3):203–213.
- [Tarbouriech and Turner, 2009] Tarbouriech, S. and Turner, M. (2009). Anti-windup design: an overview of some recent advances and open problems. *IET control theory & applications*, 3(1):1–19.
- [van Waarde et al., 2020] van Waarde, H. J., De Persis, C., Camlibel, M. K., and Tesi, P. (2020). Willems’ fundamental lemma for state-space systems and its extension to multiple datasets. *IEEE Control Systems Letters*, 4(3):602–607.

- [Willems et al., 2005] Willems, J. C., Rapisarda, P., Markovsky, I., and De Moor, B. L. (2005). A note on persistency of excitation. *Systems & Control Letters*, 54(4):325–329.
- [Wittenmark et al., 2002] Wittenmark, B., Åström, K. J., and Årzén, K.-E. (2002). Computer control: An overview. *IFAC Professional Brief*, 1:2.
- [Wood and Berry, 1973] Wood, R. and Berry, M. (1973). Terminal composition control of a binary distillation column. *Chemical Engineering Science*, 28(9):1707–1717.

Vedlegg

A. Begreper og konsepter

A.1. Det greske alfabetet

Liten	Stor	Navn
α	A	alfa
β	B	beta
γ	Γ	gamma
δ	Δ	delta
ϵ	E	epsilon
ζ	Z	zeta
η	H	eta
θ	Θ	theta
ι	I	iota
κ	K	kappa
λ	Λ	lambda
μ	M	my
ν	N	ny
ξ	Ξ	ksi
o	O	omikron
π	Π	pi
ρ	P	rho
σ	Σ	sigma
τ	T	tau
υ	Υ	ypsilon
ϕ	Φ	phi
χ	X	chi
ψ	Ψ	psi
ω	Ω	omega

A.2. Mekaniske systemer, frihetsgrader og aktueringsgrad

Mekanisk system: et system hvor dets dynamikk er utledet fra klassisk mekanikk (essensielt Newtons andre lov, og derav Euler ligning (rotasjonsdynamikk)).

Frihetsgraden (eng.: degrees of freedom; [Wikipedai](#)) til et system tilsvarer hvor mange variabler man trenger for å spesifisere konfigurasjonen til et system (se [Wikipedia](#)). Eksempler: en heis har en frihetsgrad (den kan bare gå opp og ned), et fly har 6 frihetsgrader (opp og ned, frem og tilbake, til sidene, samt 3 rotasjoner for å angi dets orientering), du har en hel hane med frihetsgrader (1 for vinkelen for kneet dit, én for hvert fingerledd, etc, etc.). Merk: et system med n frihetsgrader har som regel $2n$ tilstander (én variabel for hver frihetsgrad, samt deres tidsderivater (hastigheter)).

Aktueringsgraden (eng.: degree of actuation) til et system tilsvarer antall uavhengige aktuatorer, altså tallet av aktuatorer (altså pådrag som motorer etc) som påvirker systemet på uavhengig måter; løst forklart kan man si at to uavhengige aktuatorer påvirker dynamikken til to forskjellige tilstandsvariabler. Eksempler: et tog på en skinnegang (1 frihetsgrad) har kun en uavhengig aktuator selv om det er to jetmotorer som kan skubbe det frem og tilbake på skinnene, siden de virker ”i samme retning”; et quadcopter (drone med fire propeller) har som et fly seks frihetsgrader, men da bare fire aktuatorer (én for hver propell), så det har underaktueringsgrad $2=6-4$. Her er underaktueringsgraden (til et mekanisk system) lik systemets frihetsgrader minus antall uavhengige aktuatorer. Men hvorfor er denne graden viktig? Et system som har underaktueringsgrad lik 0 har full aktivering, noe som betyr at man kan (bare til en viss grad selsvasgt) få det til å gjøre hva man vil. De fleste systemer er dog underaktuerte, noe som setter visse begrensinger på det. F.eks. kan du ikke få en normal bil til å bevege seg direkte sidelengs, noe som gjør lukeparkering ganske så utfordrende, og gjør bevegelsesplanlegging veldig viktig som reguleringsproblemet.

A.3. Transienter, statisk respons og likevekt

B. Matematiske verktøy

B.1. Kalkulus

B.1.1 Differensiering

Produktregelen: Gitt to differensierbare funksjoner $f(x)$ og $g(x)$, så

$$\frac{d}{dx}(f(x)g(x)) = f'(x)g(x) + f(x)g'(x). \quad (\text{Produktregelen})$$

Kvotientregelen: Gitt to differensierbare funksjoner $f(x)$ og $g(x)$, så

$$\frac{d}{dx}\left(\frac{f(x)}{g(x)}\right) = \frac{f'(x)g(x) - f(x)g'(x)}{g(x)^2}. \quad (\text{Kvotientregelen})$$

Kjerneregelregeln: Gitt to differensierbare funksjoner $f(x)$ og $g(x)$, så

$$\frac{d}{dx}(f(g(x))) = f'(g(x))g'(x). \quad (\text{Kjerneregelregeln})$$

B.2. Overføringsfunksjoner, Laplace-transformasjonen og s-domenet

B.2.1 Omskriving av strengt propre overføringsfunksjoner

Noen ganger kan det greit å skrive om en overføringsfunksjon som en sum av enklere uttrykk. Vi skal vise en slik prosedyre her.

La o

$$\frac{Y(s)}{R(s)}G(s) = \frac{\mathcal{T}_G(s)}{\mathcal{N}_G(s)} = \frac{\mathcal{T}_G(s)}{(s+p_1)(s+p_2)\cdots(s+p_n)}$$

være strengt proper, slik at $\mathcal{T}_G(s)$ er av orden $n - 1$ eller mindre, og hvor alle polene p_1, \dots, p_n er forskjellige og uten imaginære deler. Det finnes et sett av tall, $\alpha_1, \alpha_2, \dots, \alpha_n$, gitt av

$$\alpha_i := \left[(s + p_i) \frac{\mathcal{T}_G(s)}{\mathcal{N}_G(s)} \right] \Bigg|_{s=-p_i},$$

slik at

$$G(s) = \frac{\alpha_1}{s + p_1} + \frac{\alpha_2}{s + p_2} + \dots + \frac{\alpha_n}{s + p_n}.$$

Så hva kan vi bruke dette til? Jo, hvis $R(s) = 1$ (tilsvarer at $r(t)$ er en enhetsimpuls), så får vi

$$y(t) = \mathcal{L}^{-1}\{Y(s)\} = \sum_{i=1}^n \alpha_i e^{-p_i t}.$$

Hvordan håndtere komplekse eller repeterende poler kommer snart; se 3.3.1 i [Seborg et al., 2016].

B.2.2 Sluttverdi-teoremet

Sluttverdi-teoremet: $\lim_{t \rightarrow \infty} f(t) = \lim_{s \rightarrow 0} sF(s)$ gitt at 1) $f(0) = 0$; 2) Laplace-transformasjonene til $f(t)$ og $f'(t) = \frac{d}{dt}f(t)$ eksisterer; 3) grenseverdiene $\lim_{t \rightarrow \infty} f(t)$ og $\lim_{s \rightarrow 0} sF(s)$ eksisterer.

Bevis: $sF(s) = s\mathcal{L}\{f(t)\} = \int_0^\infty s f(t) e^{-st} dt = \int_0^\infty \left[\frac{d}{dt}(-f(t)e^{-st}) + f'(t)e^{-st} \right] dt = [-f(t)e^{-st}]_0^\infty + \int_0^\infty f'(t)e^{-st} dt = \int_0^\infty f'(t)e^{-st} dt \implies \lim_{s \rightarrow 0} sF(s) = \int_0^\infty f'(t) dt = \lim_{t \rightarrow \infty} f(t) - f(0) = \lim_{t \rightarrow \infty} f(t).$

B.2.3 Båndbredde

Alternative kilder: §J.5.2 i [Seborg et al., 2016]

Båndbredde: bredden på intervallet (båndet) av frekvenser hvor et filter eller regulator fungerer tilnærmet som ønsket.

For eksempel, båndbredden ω_{bb} til en reguleringsløyfe med overføringsfunksjon $Y(s)/R(s) = G_{LS}(s)$ er definert som frekvensen hvor $G_{LS}(j\omega_{bb}) = 1/\sqrt{2} \approx 0.707$. Båndbredden indikerer dermed frekvensområde hvor man har tilfredsstillende settpunktfølging. Spesielt er ω_{bb} den maksimale frekvensen hvor en sinusformet referanse ikke dempes med mer enn en faktor på $\approx 0,707$. Båndbredden er også relatert til hastigheten på responsen. Generelt er båndbredden (omtrent) omvendt proporsjonal med den lukkede sløyfens innstillingstid.

B.3. Lineær algebra

B.3.1 Matriser og vektorer

Invers av en 2×2 matrise:

Gitt en 2×2 matrise

$$M = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \tag{B.1}$$

så er dens invers (hvis den eksisterer) gitt ved

$$M^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}. \quad (\text{B.2})$$

Merk: M har en invers (M^{-1} eksisterer) hvis, og bare hvis, dens determinanter ikke er lik null, altså $\det(M) = ad - bc \neq 0$.

Lineært avhengige vektorer:

To vektorer v_1, v_2 er lineært avhengig hvis det eksisterer a_1 og a_2 begge ikke lik 0 slik at

$$a_1 v_1 + a_2 v_2 = 0.$$

Dette betyr at man kan uttrykke v_1 som en funksjon av v_2

$$v_1 = -\frac{a_2}{a_1} v_2.$$

B.3.2 Egenverdier og egenvektorer

Gitt en kvadratisk matrise $A \in \mathbb{R}^{n \times n}$, en (muligens kompleks) skalar λ og en (muligens kompleks) vektor $v \in \mathbb{C}^n$ med magnitudo større enn null ($\|v\| > 0$), sånn at

$$Av = \lambda v \iff (A - \lambda I)v = 0, \quad (\text{B.3})$$

så sier vi at λ og v er henholdsvis en **egenverdi** og en **egeneigenvektor** til A .

Betydningen av egenverdier og egenvektorer Det å bestemme egenvektorer og egenverdier til A er ekvivalent med å finne λ -verdier slik at matrisen $(A - \lambda I)$ har et ikke-trivielt nullrom. Dette tilsvarer å justere λ til man finner en løsning av (B.3) der $v \neq 0$. Matematisk utføres dette ved å beregne

$$\det(A - \lambda I) = |A - \lambda I| = 0$$

Hvis determinanten er 0 er matrisen singulær (nullrommet har rang større enn 0) og det eksisterer en ikke-triviell løsning på (B.3)

Singulære matriser og rang

Singulær matrise: En kvadratisk matrise $A \in \mathbb{R}^{n \times n}$ er *singulær* hvis minst én av dens egenverdier er lik null. Siden dette betyr at det finnes en $\lambda = 0$ sånn at

$$Av = 0, \quad v \neq 0, \quad (\text{B.4})$$

så har A altså en egenvektor som er en del av dets nullrom, og er dermed ikke inverterbar.

Merk at hvis A er singulær er $\det(A) = 0$, og dermed eksisterer ikke A^{-1} . Det er dermed ingen unik løsning på likningssettet

$$Ax = b.$$

Hvis det ikke er noen unik løsning, betyr det at likningssettet er underdefinert. Det igjen betyr at 2 eller flere av likningene er ekvivalente. Man kaller disse likningene (radvektorene i A) lineært avhengige.

Diagonalisering og Jordan-form

Diagonalisering: En ikke-singulær, kvadratisk matrise $M \in \mathbb{R}^{n \times n}$ med n ulike egenvektorer, $\lambda_1, \dots, \lambda_n$, kan skrives på formen

$$M = V\Lambda V^{-1} \tag{B.5}$$

der kolonnene til V består av egenvektorene til M , og D er en diagonalmatrise av de korresponderende egenverdiene til M .

Jordan-form: TODO

C. Ekstra ting

To come...

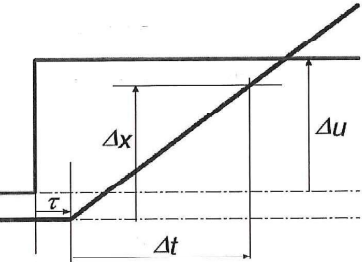
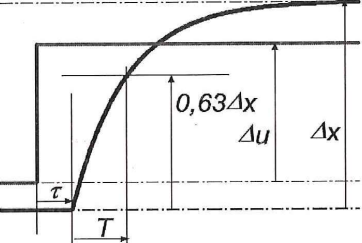
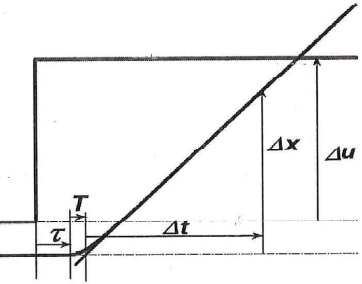
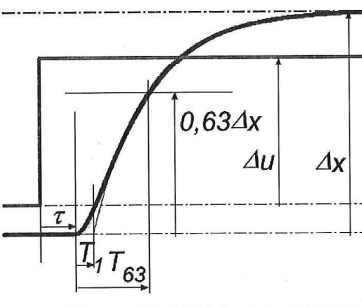
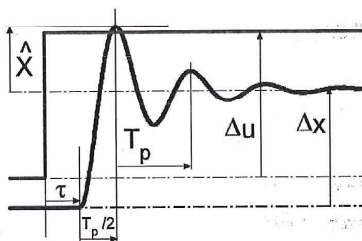


C.1. Tilstandsestimering

C.2. Systemidentifikasjon

C.3. Tilpassing av F-/A-OPTF-modell fra sprangresponser

Figur fra [Bjørvik and Hveem, 2014].

<p>Integrator (og tidsforsinkelse)</p>		$h_{ux}(s) = \frac{x(s)}{u(s)} = \frac{1}{T_i s} \cdot e^{-\tau s}$ $T_i = \frac{\Delta u}{\Delta x} \cdot \Delta t$
<p>1.ordens prosess (og tidsforsinkelse)</p>		$h_{ux}(s) = \frac{x(s)}{u(s)} = \frac{K}{1+Ts} \cdot e^{-\tau s}$ $K = \frac{\Delta x}{\Delta u}$
<p>2. ordens prosess som inneholder en integrator og en første ordens prosess (og tidsforsinkelse)</p>		$h_{ux}(s) = \frac{x(s)}{u(s)} = \frac{1}{T_i s (1+Ts)} \cdot e^{-\tau s}$ $T_i = \frac{\Delta u}{\Delta x} \cdot \Delta t$ <p>NB! Først når $t > \tau + 5T$ er stigningsforholdet til kurva lik stigningsforholdet til integratordelen.</p>
<p>2.ordens prosess med reelle poler (og tidsforsinkelse)</p>		$h_{ux}(s) = \frac{x(s)}{u(s)} = \frac{K}{(1+T_1 s)(1+T_2 s)} \cdot e^{-\tau s}$ $K = \frac{\Delta x}{\Delta u}, \quad T_{63} \approx T_1 + T_2$ $T_1 < T_2$ <p>NB! Først når $t > \tau + 5T_1$ forsvinner virkninga av den korteste tidskonstanten.</p>
<p>2.ordens prosess med kompleksskonjugerte poler (og tidsforsinkelse)</p>		$h_{ux}(s) = \frac{x(s)}{u(s)} = \frac{K}{\left(\frac{s}{\omega_0}\right)^2 + 2\zeta\left(\frac{s}{\omega_0}\right) + 1} \cdot e^{-\tau s}$ $K = \frac{\Delta x}{\Delta u}, \quad \delta = \frac{\hat{x}}{\Delta x}, \quad \zeta = -\frac{\ln \delta}{\sqrt{\pi^2 + (\ln \delta)^2}}$ $\omega_0 = \frac{2\pi}{T_p \sqrt{1 - \zeta^2}}$

Indeks

- Aktueringsgrad, 236
- Aliasing, *see* folding195
- Amplituderatio, 173
- Anti-windup, 149
- AOPTF, 77
- Auto-tuning, 118
- Avviksforholdet, 187

- Bernoulli-ligningen, 63
- Blokkdiagram, 15
- Bode-diagram, 176
- Butterworth-filter, 198
- Båndbredde, 238

- Condition number, *see* Kondisjonstall
- Cut-off-frekvens, 197

- Dekobling, 212
- Dekoblingsfilter, 212
- Desentralisert regulering, 210
- Direktesyntese, 121
- Direktevirking, 105
- Dreiemoment, 53
- Dynamikk, 39
- Dødbånd, 62

- Egenvektor, 239
- Egenverdi, 239
 - analyse, 221
- Energibalanse, 44
- Etterjustering av PID-regulatorer, 113
- Eulers bakover metode, 90
- Eulers fremover metode, 90

- Fasemargin, 185
- Fasevinkel, 173
- Folding, 195
 - Folding-filter, 195
- FOPTF, 76
- Forsterkningsmargin, 185

- Frihetsgrad, 236
- Friksjon, 61
- Følgeforholdet, 188

- Grensesvigninger, 62
- Greske bokstaver, 235

- Hysteresese, 62

- Ideell foroverkobling, 138
- Ikke-minimum-fase, 34
- initialverdiproblem, 87
- Inngang-utgangs-stabilitet, 35
- Instrumenteringsdiagram, 16
- Integrerende prosess, 77
- Intern-modell-kontrol, 125

- Kaskade-regulering, 158
- kirchhoffs spenningslov, 58
- kirchhoffs strømlov, 58
- Kjerneregelen , 237
- Komplementære sensitivitetsfunksjonen, 188
- Kondisjonstall, 224
- Kontrollvolum, 40
- Kovolusjon, *see* folding195
- Kritiske punktet, 179
- Krysskoblingsgrad, 212
- Kvotientregelen, 237

- Laplace-variabelen, 28
- Laplacetransformasjon, 28
- Lavpassfilter, 196
- Likevektspunkt, 25
- Limit cycle, *see* Grensesvigninger62
- Linearisering, 23
- Lineær algebra, 238
- LTI, 21, 22

- Massebalanse, 40
- Metning, 61
- MIMO, 206

- Minimum-fase, 34
Multivariable systemer, 206
Newtons kjølelov, 46
Newtons metode, 93
Nichols-diagram, 177
Nominelt pådrag, 104, 136
Nyquist-diagram, 175
Nyquist-frekvens, 194
ODE, 17, 20
PID-regulator, 102
 PD, 110
 PI, 110
pol-plassering, 121
Produktregelen, 237
Proper overføringsfunksjon, 31
Realiserbarhet, 34
Referanse-glatting, 146
Relative forsterkningsmatrisen, 214
Reversvirkning, 105
RFM-analyse, 214
RGA, *se* RFM-analyse
Robusthet, 10
Runge–Kutta, 89
Runge–Kutta-metoder, 95
Rykkfri overføring, 156
Sensitivitetsfunksjonen, 187
Sensivitetstoppen, 189
SI-enheter, 69
SIMC, 127
Singulærverdi-analyse, 221
Singulærverdi, 222
Singulærverdi-dekomposisjon, 223
SISO, 206
Skogestads havl-regel, 79
Stabilitet, 35
Stasjonær forsterkning, 71
Stiksjon, 62
Strengt proper overføringsfunksjon, 31
Strømlinje, 63
SVA, *se* Singulærverdi-analyse
SVD, *se* Singulærverdi-dekomposisjon
Tastefrekvens, 194
Tastetid, 194
Tidskonstant, 70
Tilbakekoblings-linearisering, 135
Tilstandsromform, 19
tracking, 154
Tregghetsmoment, 54
Tuning, 9
Ustabilitet, 36
Ventilligningen, 67
Ziegler-Nichols' metode, 114