HAL
open science

Time-continuous power-balanced simulation of nonlinear audio circuits : realtime processing framework and aliasing rejection

Rémy Müller

▶ To cite this version:

Rémy Müller. Time-continuous power-balanced simulation of nonlinear audio circuits : realtime processing framework and aliasing rejection. Sound [cs.SD]. Sorbonne Université, 2021. English. NNT : 2021SORUS453 . tel-03783502v2

HAL Id: tel-03783502
https://theses.hal.science/tel-03783502v2

Submitted on 22 Sep 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THÈSE DE DOCTORAT DE SORBONNE UNIVERSITÉ

Spécialité

Automatique et traitement du signal

École Doctorale Informatique, Télécommunications et Électronique (Paris)

Présentée par

Rémy MÜLLER

Pour obtenir le grade de

DOCTEUR de SORBONNE UNIVERSITÉ

Sujet de la thèse :

Time-continuous power-balanced simulation of nonlinear audio circuits : realtime processing framework and aliasing rejection

soutenue le 13 Juillet 2021

devant le jury composé de :

M. Stefan Bilbao	Rapporteur
M. Laurent Lefèvre	Rapporteur
Mme. Elena Celledoni	Examinatrice
M. Udo Zölzer	Examinateur
M. Bernhard Maschke	Examinateur
M. Benoit Fabre	Examinateur
M. Thomas Hélie	Directeur de thèse

Abstract

This work addresses the real-time simulation of nonlinear audio circuits. In this thesis, we use the port-Hamiltonian (PH) formalism to guarantee power balance and passivity. Moreover, we adopt a continuous-time functional framework to represent "virtual analog" signals and propose to approximate solutions by projection over time frames. As a main result, we establish a sufficient condition on projectors to obtain time-continuous power-balanced trajectories. Our goal is twofold: first, to manage frequency-bandwidth expansion due to nonlinearities, we consider numerical engines processing signals that are not bandlimited but, instead, have a "finite rate of innovation"; second, to get back to the bandlimited domain, we design "virtual analog-to-digital converters". Several numerical methods are built to be power-balanced, high-order accurate, with a controllable regularity order. Their properties are studied: existence and uniqueness, accuracy order and dispersion, but also, frequency resolution beyond the Nyquist frequency, aliasing rejection, reproducing and Peano kernels. This approach reveals bridges between numerical analysis, signal processing and generalised sampling theory, by relating accuracy, polynomial reproduction, bandwidth, Legendre filterbanks, etc. A systematic framework to transform schematics into equations and simulations is detailed. It is applied to representative audio circuits (for the UVI company), featuring both ordinary and differential-algebraic equations. Special work is devoted to PH modelling of operational amplifiers. Finally, we revisit PH modelling within the framework of Geometric Algebra, opening perspectives for structure encoding.

Acknowledgements

Je tiens à remercier Thomas Hélie, mon directeur de thèse, pour m'avoir permis de réaliser cette thèse et pour m'avoir guidé de manière à la fois exigeante et bienveillante tout au long de ce parcours. Je remercie infiniment mes employeurs, Alain et Jean-Bernard, dirigeants de la société UVI, d'avoir accepté que je dédie la moitié de mon temps de travail à cette thèse pendant 5 ans, sans financement extérieur. Je remercie également mes collègues d'UVI, pour leur patience pendant cette période, qui m'a rendu moins disponible. Je remercie l'IRCAM de m'avoir accueilli à nouveau. Je remercie les membres de l'équipe S3AM, en particulier Antoine, Tristan, Damien, Victor, Judy, Pierre, Joël, David et Marc pour nos discussions fertiles. Je remercie les rapporteurs et les examinateurs de me faire l'honneur de donner de leur temps en acceptant de relire ce manuscrit. Je remercie Olivier Verdier pour ses conseils bibliographiques qui m'ont guidé et ont ouvert mon horizon scientifique au début de cette thèse. Je remercie la communauté de la conférence DAFx, rigoureuse et pragmatique, pour laquelle j'ai toujours eu une curiosité et un attachement particuliers et sans laquelle je n'aurais probablement pas envisagé de faire cette thèse. Je remercie également l'équipe du projet INFIDHEM, pour l'organisation de l'école "Theory and applications of port-Hamiltonian systems", dans le cadre magnifique du lac Chiemsee. Cette formation m'a permis de de prendre du recul, de jeter des ponts, et de solidifier mes connaissances sur les systèmes Hamiltoniens à ports. Je remercie Vincent de m'avoir accompagné dans ce parcours en choisissant de s'engager (lui aussi et sans concertation préalable) dans le projet inhabituel d'une thèse tardive [Gou20] et d'un encadrement à distance. Je remercie également Jérome, Damien, Nicolas, Matthias ainsi que tous ceux que j'oublie qui ont suivi mon avancement de près ou de loin. Je remercie mon frère, mes parents et ma famille pour leur affection et pour m'avoir transmis le goût d'apprendre, de la dialectique et de l'esprit critique. Enfin je ne pourrai jamais assez remercier ma compagne et mon fils, qui m'ont à la fois soutenu et supporté durant ces 5 dernières années. Ils m'ont apporté l'équilibre familial indispensable pour pouvoir réaliser ce projet. Je vous suis éternellement redevable.

Contents

Introduction **1**

I Power-balanced modelling of electronic circuits **5**

1 Port-Hamiltonian Systems **7**
1.1 Reminder on dynamical systems and ODE . . . 8
1.2 Reminder on Differential Algebraic Equations (DAE) . . . 13
1.3 Introduction to port-Hamiltonian Systems . . . 17
1.4 From flow-effort to wave variables . . . 35

2 Revisiting circuit representations **43**
2.1 Kirchhoff laws . . . 45
2.2 From circuits to graphs . . . 46
2.3 Port-Hamiltonian representations of electronic circuits . . . 53
2.4 Bond Graphs and Wave Digital Filters . . . 64
2.5 Port-variable changes . . . 73

II Time-continuous power-balanced numerical methods **77**

3 Non-bandlimited signal representations, reconstruction and antialiasing **81**
3.1 Generalized-sampling theory and Finite Rate of Innovation . . . 83
3.2 Input reconstruction (Virtual DAC) . . . 89
3.3 Output antialiasing and sampling (Virtual ADC) . . . 93
3.4 Application: “virtual analog” resampler . . . 104

4 Power-balanced Adaptive collocation **107**
4.1 Satisfying the power-balance using adaptive collocation . . . 108
4.2 Method A: adaptive collocation . . . 109
4.3 Method B: symmetric adaptive collocation . . . 111
4.4 Increasing regularity: SPAC methods . . . 113

5 Power-balanced projection methods **117**
5.1 Regular Projection Methods for pH-ODE and pH-DAE . . . 119
5.2 Analysis of RPM for pH-ODE . . . 125
5.3 Analysis of RPM for pH-DAE . . . 135
5.4 Implementation choices . . . 140
5.5 Examples . . . 147

6 Power-balanced Exponential Integrators **159**
6.1 From functional Newton iteration to exponential integrators . . . 160
6.2 Exponential Average Vector Field method . . . 162
6.3 High-order energy-preserving exponential integrators . . . 168

III Applications **171**

7 Passive Operational Amplifier models **173**
7.1 A minimal passive model of the operational amplifier . . . 175
7.2 A passive fully differential amplifier model with infinite gain . . . 190
7.3 Towards a grey-box passive model of the OPA . . . 195

8 Circuits case studies **197**
8.1 Fuzz Face (NPN variant) . . . 199
8.2 Big Muff tone clipper . . . 204
8.3 Tube Screamer drive stage . . . 208
8.4 Korg MS-20 Filter . . . 212
8.5 FitzHugh–Nagumo relaxation oscillator . . . 217
8.6 Passive peaking equalizer (beyond the Nyquist frequency) . . . 221

IV Towards Geometric Algebra **235**

9 Geometric Algebra for PHS **237**
9.1 Introduction to Geometric Algebra . . . 240
9.2 Motivating examples and invariants . . . 249
9.3 Port-Hamiltonian systems using Geometric Algebra . . . 252
9.4 Representing Dirac structures with Geometric Algebra . . . 254
9.5 Exploring the geometry of R(n,n) with GA . . . 258
9.6 Rotor description of the flow-effort to wave variables change . . . 260

General Conclusion **263**

V Appendix **269**

A Relations: definitions and properties **271**

B Reminder on ODEs **275**
B.1 Runge–Kutta methods . . . 275
B.2 Numerical Stabillity . . . 276
B.3 Elementary differentials and B-series . . . 278

C Functional Analysis **281**
C.1 Definitions . . . 281
C.2 Banach, Hilbert and Sobolev spaces . . . 283
C.3 Strang–Fix conditions . . . 285
C.4 Shifted orthonormal Legendre polynomials . . . 286
C.5 Hermite polynomial splines . . . 287

D Proofs **289**
D.1 Exponential φ-functions: proofs and properties . . . 289
D.2 CSRK formulation of projected ODEs . . . 291
D.3 Proof of proposition 5.3 (CSRK order and Strang-Fix conditions) . . . 292
D.4 Proof of theorem 5.2 (existence and uniqueness of CSRK solutions) . . . 293
D.5 Proof of proposition 5.5 p.129 (nested projectors) . . . 294
D.6 Proof of theorem 5.7 (Legendre expansion) . . . 296
D.7 Stability function of L^2 projection methods . . . 297
D.8 Proof of Gauss-Legendre quadrature formula . . . 300
D.9 Proofs and appendix for section 7.1 (Minimal passive OPA) . . . 301
D.10 Z-domain response of Legendre projection filterbank (linear state-space system) 305

E Code listing (SPAC methods) **309**

F Geometric Algebra **313**
F.1 Algebra . . . 313
F.2 Calculus . . . 319
F.3 Maxwell equations (in empty space) . . . 322

G Articles **323**

Bibliography **356**

List of Figures

1.1 Lyapunov stability theorem . . . 12
1.2 Graphical description of a Port-Hamiltonian System. . . . 17
1.3 Composition of Dirac structures (Parallel composition). . . . 22
1.4 Block diagram of energy storing elements. . . . 24
1.5 Examples of adimensioned effort laws and their corresponding energies. . . . 26
1.6 Law of a linear resistor and its current and voltage power potentials. . . . 30
1.7 Law of a Shockley Diode and its current and voltage potentials . . . 30
1.8 Static characteristic of a tunnel diode . . . 31
1.9 (Bondspace) relation between the subspaces $\mathcal{D}, \mathcal{W}^-, \mathcal{W}^+$ . . . 37
1.10 Scattering of nonlinear storage structures . . . 41

2.1 Circuit modelling representations . . . 44
2.2 Diode clipper graph. . . . 47
2.3 (Graph Theory) loops . . . 48
2.4 Examples of spanning trees shown in black, with their cotree shown in dashed . . 48
2.5 (Graph Theory) prototyping boards and incidence matrices . . . 50
2.6 (Graph theory) cutsets . . . 51
2.7 (Graph theory) minimum spanning tree . . . 59
2.8 (Graph theory) LCLC circuit with implicit constraints . . . 59
2.9 Isothermal RLC embedding . . . 63
2.10 Automated Bondgraph modelling of the diode clipper circuit. . . . 67
2.11 Bi-partite bondgraph of the diode clipper circuit. . . . 68
2.12 Equivalence between circuit Bondgraph and WDF represention. . . . 69
2.13 Example of a circuit containing a rigid node . . . 71
2.14 Illustration of common-differential adaptation of a 2-port. . . . 74
2.15 Generalized n-port adapter. . . . 75
2.16 (continuous-time virtual analog signal processing) block-diagram of the approach 79

3.1 (non-bandlimited signals) common causes . . . 82
3.2 (continuous-time virtual analog signal processing) virtual DAC and ADC . . . 82
3.3 (B-splines) digital IIR prefiltering . . . 89
3.4 (B-splines) time domain plot . . . 90
3.5 (vDAC) shifted linear interpolation principle . . . 90
3.6 (vDAC) Time and frequency response of shifted linear interpolation . . . 91
3.7 (B-splines) comparison with cardinal interpolating B-splines . . . 92
3.8 (B-spline) impulse responses of cardinal interpolating pre-filters . . . 92
3.9 (Exact ARMA filtering) filtered polynomial basis . . . 95
3.10 (Exact ARMA filtering) first-order lowpass output . . . 96
3.11 (Exact ARMA filtering) Butterworth rejection of signal at the Nyquist frequency 96

3.12 (Exact ARMA filtering) Butterworth impulse and step responses 97
3.13 (vADC) Block diagram of causal Legendre to cubic B-spline projection filterbank. 98
3.14 (vADC) B-spline / Legendre conversion operators 99
3.15 (vADC) Barycentric overlap-add 101
3.16 (vADC) B-spline aproximation of square, saw and triangle 102
3.17 (vADC) B-spline sine reconstruction 103
3.18 (Virtual Analog resampler) block-diagram. 104
3.19 (Virtual Analog resampler) spectrum periodisation. 104

4.1 (PAC) optimal α as function of dissipation 110
4.2 (PAC) damped RLC orbits, mid-point vs PAC(1) 111
4.3 (SPAC) optimal β as function of dissipation 112
4.4 (SPAC) power-balanced regions in the complex plane 114
4.5 (SPAC) optimal collocation points as function of dissipation 114

5.1 (RPM) polynomial supplementary boundary functions 130
5.2 (RPM) comparison of operators $\mathcal{P}$ and $\mathcal{Q}$ 130
5.3 (RPM) convergence of approximation by operator $\mathcal{Q}$ 131
5.4 (RPM) Peano error kernels for $\mathcal{P}$ 133
5.5 (RPM) Peano error kernels for operator $\mathcal{Q}$ 134
5.6 (RPM) smoothing effect of the Average Discrete Gradient (ADG) 143
5.7 (RPM) convergence of Gauss–Legendre quadrature 145
5.8 (Nonlinear LC) orbits 148
5.9 (Nonlinear LC) trajectories 149
5.10 (Linear LC) orbits and frequeny warping 150
5.11 (Nonlinear LC) impact of projection order and smoothness on spectrum and aliasing 151
5.12 (Nonlinear LC) local energy error 152
5.13 (Nonlinear LC) long-term energy conservation 152
5.14 (Diode clipper) simulation 154
5.15 (Diode clipper) sinesweep spectrograms (linear scale) 155
5.16 (Diode clipper) sinesweep spectrograms (log scale) 156

6.1 Schematic description of the Exponential AVF method 163
6.2 (EAVF) visual proof 164
6.3 (EAVF) nonlinear potential function V 165
6.4 (EAVF) comparison of EAVF and AVF methods 166

7.1 (Passivity test) operational amplifier circuit. 174
7.2 (Passivity test) simulation results in LTSPICE 174
7.3 (OPA) circuit diagram 176
7.4 (OPA) adimensioned modulation factor 178
7.5 (OPA) non-inverting voltage amplifier circuit 179
7.6 (Voltage amplifier) input-output map 180
7.7 A single-rail voltage amplifier powered by a capacitor. 181
7.8 (single-rail amplifier) time-domain simulation 182
7.9 (single-rail amplifier) comparison of discharge rates 182
7.10 (Sallen-Key filter) Bode plot 183
7.11 (Sallen-Key filter) circuit to bond-graph to equations 184
7.12 (Sallen-Key filter) feedback law and its potential 185
7.13 (Sallen-Key filter) simulation 187

7.14 (Sallen-Key filter) comparison with LTSPICE . . . 188
7.15 (Sallen-Key filter) sine sweep spectrograms . . . 188
7.16 (FDA) Ideal non-energetic Fully Differential Amplifier 3-port. . . . 190
7.17 (FDA) Ideal law in the (v_I, v_0)-plane expressed as a multi-valued function. . . . 191
7.18 (FDA) Ideal law in (v_I, v_0, λ_1) coordinates . . . 191
7.19 (FDA) Dual adimensioned functions μ, μ^* . . . 193
7.20 (FDA) Ideal laws in (v_S, λ_1, v_O) and (v_S, λ_1, v_I) spaces . . . 193
7.21 (FDA) voltage buffer example . . . 193
7.22 (FDA) causal map in input-output Σ-Δ coordinates. . . . 194
7.23 (OPA, grey box model) Structure of the macro model . . . 195
7.24 (OPA, grey box model) building blocks candidates. . . . 195

8.1 (NPN Fuzz Face) schematic. . . . 199
8.2 (NPN Fuzz Face) simulation. . . . 203
8.3 (NPN Fuzz Face) overlay of simulations . . . 203
8.4 (BMP Tone clipper) schematic . . . 204
8.5 (BMP Tone Clipper) simulation. . . . 207
8.6 (Tube Screamer drive) schematic . . . 208
8.7 (Tube Screamer drive) linearised frequency response. . . . 209
8.8 (Tube Screamer drive) simulation . . . 211
8.9 (MS-20 filter) schematic . . . 212
8.10 (MS20 filter) overdrive amplifier . . . 214
8.11 (MS-20 filter) simulation . . . 215
8.12 (MS-20 filter) simulation with varying amplitudes . . . 216
8.13 (MS-20 filter) response to a 1V sawtooth signal . . . 216
8.14 (Fitzhugh–Nagumo oscillator) schematic . . . 217
8.15 (Fitzhugh–Nagumo oscillator) tunnel diode bias for multi-vibrator behaviour . . . 217
8.16 (Fitzhugh–Nagumo oscillator) simulation . . . 219
8.17 (Fitzhugh–Nagumo oscillator) phase plot . . . 220
8.18 (Peaking EQ) schematic . . . 221
8.19 (Peaking EQ) frequency response . . . 223
8.20 (Peaking EQ) interpretation of RPM as a continuous/discrete Legendre filterbank 225
8.21 (Peaking EQ) Legendre exponential approximation error in the frequency domain 227
8.22 (Peaking EQ) Legendre exponential approximation error in the Laplace domain . 228
8.23 (Peaking EQ) magnitude response of the projected system . . . 229
8.24 (Peaking EQ) phase response of the projected system . . . 230
8.25 (Peaking EQ) transfer function approximation: oversampling vs high-order . . . 232

9.1 Inner product of vectors. . . . 240
9.2 Exterior product of vectors. . . . 240
9.3 Projection and rejection. . . . 245
9.4 Reflection. . . . 246
9.5 Mirror in a line $\mathbf{x}$ and its dual hyperplane $\mathbf{x}^*$. . . . 246
9.6 Rotation. . . . 246
9.7 GA identity $|\mathbf{u} \wedge \mathbf{v}| = |\mathbf{u}||\mathbf{v}| \sin\theta$. . . . 247
9.8 Angular momentum of an harmonic oscillator . . . 249
9.9 Bernouilli's logarithmic spiral. . . . 251
9.10 Main involution of the bond space . . . 259

C.1 (Shifted orthonormal Legendre polynomials) . . . 287

C.2 (Shifted orthonormal Legendre polynomials) Fourier spectrum 288

D.1 (RPM method) Illustration of orthogonal and oblique projections $\mathcal{P}$, $\mathcal{Q}$, $\mathcal{R}$ 295
D.2 (Legendre projection) frequency warping 298
D.3 (Legendre projection) dissipative warping 298
D.4 (Legendre projection) conformal map of AVF/mid-point methods 299
D.5 (Push-pull) class-B amplifier. 302
D.6 (Push-pull) output functions 304

List of Tables

1.1 (power-conserving Dirac structures) common examples in electronics. 18
1.2 (energy storing components) examples in electronics. 24
1.3 (passive memoryless components) examples in electronics 27

3.1 (vADC) B-spline / Legendre conversion operators 99

4.1 (SPAC) numerical properties 113

5.1 (RPM) reproducing and Peano error kernel of Legendre projection 132
5.2 (RPM) principle of the time discretisation approach. 140

8.1 (Peaking EQ) Laplace transforms of Legendre polynomials restricted to $(0, 1)$ . . 226
8.2 (Peaking EQ) exponential approximation error in the Laplace domain 227
8.3 (Peaking EQ) transfer function error: oversampling vs high-order 231

9.1 (Geometric Algebra) canonical bases of $\mathbb{G}^2$, $\mathbb{G}^3$, $\mathbb{G}^4$ 242
9.2 (Geometric Algebra) Multiplication tables. 243

Notations

Common Spaces

$\mathbb{R}$	reals
$\mathbb{C}$	complex
$\mathbb{N}$	natural integers
$\mathbb{Z}$	signed integers
$\mathbb{P}^k$	space of polynomials of maximal degree k
$\mathcal{X}$	state space manifold
$\mathsf{T}\mathcal{X}$	tangent space
$\mathsf{T}^*\mathcal{X}$	co-tangent space

Common Variables

$\mathbf{u}$	system input
$\mathbf{x}$	system state
$\mathbf{y}$	system output
$\mathbf{e}$	effort
$\mathbf{f}$	flow
$\mathbf{v}$	tension
$\mathbf{i}$	current

Inner products and norms

$\langle f, g\rangle$	inner product
$\lVert f\rVert$	norm
$\langle f \mid g\rangle$	duality product
$\mid f\rangle$	vector (ket)
$\langle f \mid$	covector (bra)
$\langle f \mid g\rangle_A$	operator duality product ($\langle f \mid g\rangle_A = \langle f \mid A \mid g\rangle$)
$\lVert f\rVert_A$	induced norm by an operator A

Functional Notations

E^*	(algebraic) dual of the space E
E'	(topological / continuous) dual of the space E
A^*	adjoint of an operator A ($\langle f, Ag\rangle = \langle A^* f, g\rangle$)

$\mathcal{P}$	projection operator
$\sigma(A)$	spectrum of an operator A. $\sigma(A) = \mathrm{eig}(A)$
$\rho(A)$	spectral radius of an operator A. $\rho(A) = \sup\big\lvert\mathrm{eig}(A)\big\rvert$.
Ω	open domain
$\partial\Omega$	boundary of Ω
$\overline{\Omega}$	closure of Ω ($\overline{\Omega} = \Omega \cup \partial\Omega$)
$\mathcal{C}^k(\Omega)$	space of k times continuously derivable functions.
L^p	Lebesgue space of measurable functions in the p-norm over continuous domains
ℓ^p	Lebesgue space over discrete domains
L^2	Lebesgue space of square integrable functions (an Hilbert space)
$W^{k,p}$	Sobolev space of k-differentiable L^p functions
H^k	Sobolev space of k-differentiable L^2 functions (an Hilbert space)
$f\vert_\Omega$	the restriction of a function f to the domain Ω
$[f]_a^b$	difference of f over the interval $[a, b]$, $[f]_a^b = f(b) - f(a)$
$A \oplus B$	direct sum of two vector spaces.
$D(A)$	domain of an operator A
$R(A)$	range of an operator A
$N(A)$	nullspace of an operator A

Signals

$f(t)$	continuous time signal
$\delta(t)$	Dirac delta distribution
$\Theta(t)$	Heaviside function
$(t)_+$	$\max(0, t)$

Misc

$\sim$	similar, identifiable
$\approx$	approximately equal
δ_{ij}	Kronecker delta
$\mathbf{1}_A$	indicator function of a set A
$\lfloor x \rfloor$	largest integer smaller or equal to x (floor)
$\lceil x \rceil$	smallest integer greater or equal to x (ceil)

Integration and derivation

$f^{(n)}(t)$	n-th order derivative
$f^{[n]}(t)$	n-th order anti-derivative
$\dot{x}$	time derivative of x
$f'(x)$	derivative of f with respect to the free variable x.
∇	gradient operator
∂_x	partial derivative with respect to the variable x
$F'(x)(\cdot)$	Fréchet derivative of a function F at x (a linear operator)
$\mathcal{V}$	Volterra operator $(\mathcal{V}f)(x) = \int_0^x f(s)\,\mathrm{d}s$

Geometric Algebra

$\mathbf{u} \cdot \mathbf{v}$	inner product
$\mathbf{u} \wedge \mathbf{v}$	outer product
$\mathbf{uv}$	geometric product $\mathbf{uv} = \mathbf{u} \cdot \mathbf{v} + \mathbf{u} \wedge \mathbf{v}$

Acronyms

ADC	Analog to Digital Converter (p.79)
ADG	Average Discrete Gradient (p.143)
ARMA	Auto-Regressive Moving Average (filter) (p.82)
AVF	Average Vector Field [CGM+12]
BJT	Bipolar Junction Transistor (p.32)
BVP	Boundary Value Problem
CSRK	Continuous-Stage Runge–Kutta (methods) (p.126)
DAC	Digital to Analog Converter (p.79)
DAE	Differential-Algebraic Equations (p.13)
FEM	Finite Element Methods
FRI	Finite Rate of Innovation (p.86)
GA	Geometric Algebra (p.237)
GNI	Geometrical Numerical Integration [HLW06]
IVP	Initial Value Problem (p.8)
KCL	Kirchhoff Current Laws (p.45)
KVL	Kirchhoff Voltage Laws (p.45)
LTI	Linear time-invariant
LTV	Linear time-variant
MNA	Modified Nodal Analysis [HRB75]
MOSFET	Metal-Oxyde Semiconductor Field-Effect Transistor
ODE	Ordinary Differential Equation (p.8)
OPA	OPerational Amplifier (p.173)
PDE	Partial Differential Equations
PHS	Port Hamiltonian Systems (p.17)
RK	Runge–Kutta (methods) (p.275)
RPM	Regular Power-balance projection Method (p. 119)
(S)PAC	Symmetric Power-balanced Adaptive Collocation (method) (p.107)
SPICE	Simulation Program with Integrated Circuit Emphasis [AB90]
STA	Sparse Tableau Analysis [HBG71]
TFEM	Time-Finite Elements (Methods)
WDF	Wave Digital Filters (p.35, p.69)

Introduction

Context

This thesis is the result of a joint collaboration between UVI (my employer) and the S3AM [1] team of the STMS [2] laboratory at IRCAM [3]. It is unusual on two aspects: it happened as a late PhD, 12 years after the end of my studies, and it took place, for the last five years, as a part-time project, in parallel of my job at UVI. I am very grateful to my employers for this opportunity, their trust, their continuous support and for fully funding this PhD.

The UVI company UVI [4] is a french SME, founded in 1987 by Alain Etchart and Jean-Bernard Celier with head-quarters in Paris and offices in US and Japan. It is specialised in the creation of virtual instruments and digital audio effects for sound-design and music production. UVI's flagship product, called Falcon [5] (and the underlying UVI engine), is a multi-synthesis workstation with sixteen synthesis types and more than ninety audio effects. It integrates signal modelling (additive, subtractive, granular, FM, etc), physical modelling and algorithmic musical event processing within the same environment. The aim of this thesis for UVI is to broaden the range of audio systems that can be emulated in real-time by physical modelling of audio circuits.

The S3AM team Multi-physics audio acoustics and virtual analog modelling is an important thread of research in the S3AM team for which the port-Hamiltonian formalism [MV92, VDSJ14] constitutes an important backbone and a unifying language. This thesis is a followup on the work of Falaize [Fal16, FH16a] on PH audio circuit modelling (including the Wah-Wah [FH13], the Fender Rhodes[FH17], speaker modelling [FH20], etc) and (to a less extent) on the work of Lopes [Lop16] (in particular a conservative linearly-implicit method based on energy quadratisation [LHF15]). During that time, Falaize wrote a symbolic-numerical Python toolbox dedicated to PHS modelling and simulation called PyPHS [FH16b]. Earlier work in the team includes the work of Cohen and Usciati on audio circuit modelling (including triodes) [Coh12, Tar12]. Since then, ongoing work based on PHS have been dedicated to loudspeakers [LWH+20, LH20], the vocal tract [SHV19, WHS19, WHS20], Lie groups and (multi) symplectic integrators [CB17, CB19, BC19] active and finite-time control [JRH+17, JDT+17, WdNHR18, WdNF+19], the Ondes Martenot [NHRB20], PHS realisability [NHB+18] and magnetic hysteresis [NMHR20]. The team has been involved in two port-Hamiltonian research projects: the ANR projects Hamecmopsys [6] and the ANR-DFG project INFIDHEM [7], and is also actively working on Volterra series and identification methods [BHR18, Bou18, DHR19].

1. Sound Systems and Signals: Audio/Acoustics, InstruMents http://s3am.ircam.fr
2. Science and Technology of Music and Sound (UMR9912) https://www.stms-lab.fr
3. Institut de Recherche et Coordination Acoustique et Musique http://www.ircam.fr
4. https://www.uvi.net/about-us
5. https://www.uvi.net/falcon
6. https://hamecmopsys.ens2m.fr
7. https://websites.isae-supaero.fr/infidhem/

Objectives

To simulate nonlinear electronic audio circuits, we consider the class of open, power-balanced multi-physical systems. In this context, port-Hamiltonian systems (PHS) offer a structured representational framework capable of dealing with energetic, algebraic and dynamical properties. This thesis aims at designing a set of mathematical and computational methods that

1. accurately describe targeted systems in a modular way,
2. propose a systematic approach to automate modelling and real-time simulation of electronic audio circuits,
3. model dynamical systems as port-Hamiltonian Systems,
4. simulate PHS in the continuous-time domain,
5. numerically preserve the power-balance of the approximated PHS,
6. reproduce the regularity of continuous-time solutions.

Short literature overview

Virtual analog audio Modelling of (vintage) audio circuits is categorised in both academia and audio markets under the term *virtual analog* (VA) [DSS09, Sti05, VH06, VFSZ10, VBS+11, D'A14, Wer16, EGZ17, EPPB17a, BVS20] [8]. Motivations for VA modelling are multiples: 1) preserving the legacy of instruments and audio effects from obsolescence (old components are often fragile or discontinued), 2) capturing the pleasant (and sometime complex) behaviour of analog designs that is not easily reproducible by direct digital means, 3) simplifying the maintenance by replacing (heavy, expensive, fragile) hardware by software. Significant research has been devoted to the simulation of synthesiser filters [SS96, Huo04, Hél09, Pd13], equalisers [AB03a, SH11], guitar amplifiers [PY, DZ11b, DHZ11, Mac12a, Coh12], modulation effects [Huo05, EFHZ14, Mac16], distortion and saturation [HDZ11, EZ16, Hol16, HZ16], dynamic processors [AB03b, GMR12, GEZ17], analog delay and reverberation [BAC06, RS10, BP10, HP18]. Modelling approaches divides in black-box models (which aim at reproducing the input-output behaviour of systems disregarding their internal details) and white-box models (which by contrast decompose systems into networks of known elementary components). Black-box modelling approaches in audio include Volterra series and block models [BCD84, BTC83, EZ18, DHR19, EZ16], kernel methods [SW06, GE13] and neural networks [WDV19, PB19, MRBR20]. White-box approaches (which we consider in this thesis) can be categorised in two groups: state-space methods (based on Kirchhoff variables) [YAS10, DHZ10, HZ11, HZ15, FH16a] and Wave Digital Filters (based on wave variables) [Fet86, DSS09, Bil04, WBSS18]. Energy-conserving methods in audio have been considered in [Bil05, Bil08, THB14, CvW15a, CvW15b] and anti-derivative based anti-aliasing in [PZLB16, BEPV17, BEV17, MH17, Hol20, Alb20, Car20]. Note that VA audio often involves several physical domains within a single device (electric, magnetic, acoustic, mechanical, even optical). The port-Hamiltonian formalism is a natural candidate to deal with multi-physics: using power exchange as the common mean of interaction between physical domains.

Port-Hamiltonian Systems and Geometric Numerical integration The PH formalism [MV92, VDSJ14, VdS17] lies at the intersection of network modelling [Pay61], differential geometry [Olv00] and Geometrical Numerical Integration (GNI) [9]. The goal of GNI is to propose numerical integration methods (see [HNW93, HW96, BG08, Ise09]) which (in addition to numerical

8. An overview of VA (up to 2011) can be found in [PV11]
9. See [HLW06] and references therein for an overview of the domain.

accuracy) preserve geometric properties of the flow of differential equations such as symplecticity (see [Wei83]), first-integrals (such as the energy), time-reversibility, passivity (for dissipative systems) or group structure (in Lie group integrators [IMKNZ00, Cel03]). The preservation of geometric invariants leads to improved qualitative and quantitative solutions in particular over long time scales. Unconditionally energy-preserving (resp. dissipating) methods have been proposed in [Hai10, HL14, CMM+09, CGM+12, CMOQ10] (An automatic consequence of energy-preservation/passivity is the stability of simulated nonlinear systems). In particular, numerical methods for PHS have been considered in [KL19] (based on symplectic integration) and [CH17] (energy preserving/dissipating). In this thesis, our main geometric focus is on the power-balance of physical systems, i.e. exact energy preservation for conservative systems and monotonous energy decay for dissipative systems.

Thesis outline

This thesis is structured in 4 parts described below.

Power-balanced modelling of electronic audio circuits Starting from the netlist description of an electronic circuit, revisiting state of the art, methods are proposed to automatically generate different PHS representations (Kirchhoff–Dirac structure, Hybrid semi-implicit algebro-differential equations, input-state-output ordinary differential equations, thermodynamic embedding, etc). This part is meant as a guide for practitioners and implementers, where the PHS approach is favoured over classical circuit modelling approaches which are already well-documented such as modified nodal analysis. A particular attention is paid to the usefulness of each representation to derive efficient simulations. We also closely consider the sequence of transformations that are required to convert between these representations. Wave-variables formulations are recalled and a side by side comparison of network modelling using bond-graphs and Wave Digital Filters is proposed to highlight their striking and often unnoticed similarities.

Time-continuous power-balanced numerical methods In this thesis, high-order power-balanced numerical schemes are proposed. Their common ground and distinctive attribute is to exclusively consider continuous-time signal representations in functional spaces. The word *discretisation* is used in a generalised sense as the subspace representation of signals with a finite number of parameters per unit of time. This specific approach exhibits interesting connections between numerics, signal processing, generalised sampling theory, and physical modelling. A particular attention is paid to signal smoothness and rejection of spectral aliasing artefacts caused by system nonlinearities. The proposed approach relies on

1. piecewise parametric representation of *non-bandlimited signals* with a controllable regularity order and a finite rate of innovation,
2. appropriate choices of signal spaces and approximations preserving the continuous-time power-balance,
3. post-simulation *continuous-time anti-aliasing* filters and resampling.

An advantage of the proposed approach, is that the same functional discretisation methodology can be used to address both ordinary and differential-algebraic equations (which also applies to partial differential equations).

Electronic components and circuits: applications and results The proposed modelling framework and numerical discretisation methods are evaluated on a number of representative nonlinear audio circuits (covering both ODE and DAE) used by guitarists, synthesiser players and

sound-engineers. In particular, we consider the simulation of fuzz, overdrive and self-oscillating circuits. We also consider the simulation of (linear) systems having poles above the Nyquist frequency thanks to the extended generalised bandwidth of high-order methods. A chapter is dedicated to passive modelling of the operational amplifier. Indeed, the operational amplifier is a key component of analog electronics, but despite the amount of literature on the topic, we found that a simple passive model of the operational amplifier compatible with port-Hamiltonian modelling was still missing.

Towards Geometric Algebra The last part of this thesis is prospective. We explore the potentialities of Geometric Algebra (GA) in the context of port-Hamiltonian modelling. Geometric Algebra is an elegant graded algebra unifying the Euclidean inner product and Grassman exterior product into a single product called the geometric product. This unification has far reaching consequences since complex numbers, quaternions, octonions, spinors, exterior algebra, etc, can all be generated from simple axioms as sub-algebras of Geometric Algebra. Furthermore, since PH theory is deeply rooted in differential geometry and coordinate-free representations there is a natural match with GA. Given the scope of this thesis, we can only scratch the surface. In particular we consider intrinsic representations of linear transforms and Dirac structures using Geometric Algebra.

Publications

[MH17] Müller Rémy, Thomas Hélie, "Trajectory Anti-Aliasing on Guaranteed-Passive Simulation of Nonlinear Physical Systems", *20th International Conference on Digital Audio Effects (DAFx-17)*, 2017.

[MH18] Müller Rémy, Thomas Hélie, "Power-Balanced Modelling Of Circuits As Skew Gradient Systems", *20th International Conference on Digital Audio Effects (DAFx-18)*, 2018.

[MH19] Müller Rémy, Thomas Hélie, "A minimal passive model of the operational amplifier: application to Sallen–Key analog filters", *20th International Conference on Digital Audio Effects (DAFx-19)*, 2019.

[MH20] Müller Rémy, Thomas Hélie, "Fully-implicit algebro-differential parametrization of circuits", *20th International Conference on Digital Audio Effects (DAFx-20)*, 2020.

[NMHR20] Judy Najnudel, Rémy Muller, Thomas Hélie, David Roze, "A power-balanced dynamic model of ferromagnetic coils", *20th International Conference on Digital Audio Effects (DAFx-20)*, 2020.

Part I

Power-balanced modelling of electronic circuits

Chapter 1

Port-Hamiltonian Systems

Contents

1.1 Reminder on dynamical systems and ODE **8**
1.1.1 State-space representation, existence and uniqueness of solutions 8
1.1.2 Lyapunov stability and LaSalle invariance principle 9
1.1.3 Open systems and passivity 11
1.2 Reminder on Differential Algebraic Equations (DAE) **13**
1.2.1 DAE Indexes 13
1.2.2 Semi-explicit DAEs 14
1.2.3 Singular perturbations 15
1.2.4 Existence and uniqueness of solutions 16
1.3 Introduction to port-Hamiltonian Systems **17**
1.3.1 Power-conserving elements (Dirac structures) 18
1.3.2 Energy-storing elements 24
1.3.3 Passive memoryless elements 27
1.3.4 Input-State-Output Representation (PH-ODE) 33
1.3.5 Semi-explicit representation (PH-DAE) 34
1.4 From flow-effort to wave variables **35**
1.4.1 The classical wave variable change 35
1.4.2 Geometric viewpoint 36
1.4.3 Wave variables representation of Port-Hamiltonian Systems 38

The foundations of the Port-Hamiltonian formalism are recalled in this chapter. We restrict the presentation to the finite-dimensional settings which is sufficient to cover lumped electronic circuits. First, general results on existence, uniqueness and stability of state-space systems and Differential Algebraic Equations are recalled in section 1.1 and section 1.2, then the constitutive parts of port-Hamiltonian systems (power-balanced interconnections, energy-storing elements, passive algebraic components and external ports) are presented in section 1.3. Finally since the Wave Digital Filter (WDF) formalism [Fet86] is also an important modelling tool for physical modelling and virtual analog electronics, we try to bridge the gap between both formalisms by closing this chapter with section 1.4 on wave variables representations of port-Hamiltonian Systems.

1.1 Reminder on dynamical systems and ODE

This section recalls definitions and results on dynamical systems and stability (see [KG02]).

1.1.1 State-space representation, existence and uniqueness of solutions

We consider dynamical systems modelled by a finite number of coupled ordinary equations

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{f}(t, \boldsymbol{x}(t), \boldsymbol{u}(t)), \tag{1.1a}$$

where $\boldsymbol{f} : (t, \mathbf{x}, \mathbf{u}) \in \mathbb{R} \times \mathbb{R}^{n_\mathbf{x}} \times \mathbb{R}^{n_\mathbf{u}} \mapsto \boldsymbol{f}(t, \mathbf{x}, \mathbf{u}) \in \mathbb{R}^{n_\mathbf{x}}$ is the vector field function, $\dot{\boldsymbol{x}}$ denotes the time derivative [1] of the *state variable* $\boldsymbol{x}$ and $\boldsymbol{u}$ denotes the *input variable* of the system. The *state equation* (1.1a) is often associated with an *output equation*

$$\boldsymbol{y}(t) = \boldsymbol{h}(t, \boldsymbol{x}(t), \boldsymbol{u}(t)), \tag{1.1b}$$

where $\boldsymbol{h} : (t, \mathbf{x}, \mathbf{u}) \in \mathbb{R} \times \mathbb{R}^{n_\mathbf{x}} \times \mathbb{R}^{n_\mathbf{u}} \mapsto \boldsymbol{h}(t, \mathbf{x}, \mathbf{u}) \in \mathbb{R}^{n_\mathbf{y}}$ is the observation function.

Remark 1.1. If the input is known explicitely (e.g. an external source or a state feeback $\boldsymbol{u}(t) = \boldsymbol{g}(\boldsymbol{x}(t))$). Then, it is possible to rewrite (1.1a) to remove the dependence on $\mathbf{u}$ as

$$\dot{\boldsymbol{x}}(t) = \tilde{\boldsymbol{f}}(t, \boldsymbol{x}(t)), \qquad \text{with} \qquad \tilde{\boldsymbol{f}}(t, \mathbf{x}) = \boldsymbol{f}(t, \mathbf{x}, \boldsymbol{u}(t)).$$

Furthermore, by including time t into an extended state $\mathbf{z} = (t, \mathbf{x})$ and adding the differential equation $\dot{t} = 1$, it is always possible to obtain an *autonomous system*

$$\dot{\boldsymbol{z}}(t) = \breve{\boldsymbol{f}}(\boldsymbol{z}(t)), \qquad \text{with} \qquad \breve{\boldsymbol{f}}(\mathbf{z}) = \left[1, \tilde{\boldsymbol{f}}(t, \mathbf{x})\right]^\mathsf{T}.$$

To predict the future state of the system from its initial value $\mathbf{x}_0$ at time t_0, the following *Cauchy problem* must have a unique solution.

Definition 1.1 (Cauchy problem)**.** Let $\mathbb{T} = [t_0, t_1]$, $\mathbf{x}_0$ an initial condition in $\mathcal{X} \subset \mathbb{R}^n$ and $\boldsymbol{f} : \mathbb{T} \times \mathcal{X} \to \mathbb{R}^n$. The *Cauchy problem* is to find a unique function $\boldsymbol{x} : \mathbb{T} \to \mathcal{X}$ such that

$$\begin{cases} \dot{\boldsymbol{x}}(t) = \boldsymbol{f}(t, \boldsymbol{x}(t)), & \forall t \in \mathbb{T}, \\ \boldsymbol{x}(t_0) = \mathbf{x}_0, & t = t_0. \end{cases} \tag{1.2}$$

A key property to establish existence and uniqueness, is that $\boldsymbol{f}$ must satisfy a *Lipschitz condition*.

Theorem 1.1 (Local existence and uniqueness ([KG02] p.88))**.** *Let $\boldsymbol{f}(t, \mathbf{x})$ be piecewise continuous in t and satisfy the local Lipschitz condition*

$$\left\| \boldsymbol{f}(t, \mathbf{x}_1) - \boldsymbol{f}(t, \mathbf{x}_2) \right\| \leq L \|\mathbf{x}_1 - \mathbf{x}_2\|, \tag{1.3}$$

$\forall \mathbf{x}_1, \mathbf{x}_2 \in B = \left\{ \mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x} - \mathbf{x}_0\| \leq r \right\}$, $\forall t \in [t_0, t_1]$. *Then there exists some $h > 0$ such that the state equation (1.2) as a unique solution over $[t_0, t_0 + h]$.*

The previous theorem based on the Banach fixed point theorem only requires a simple Lipschitz condition but does not recover the maximal existence domain of solutions (even in the linear case). For stiff systems (when the step size h is bigger than some time constants of the system), the following theorem, based on Newton iteration, yields better estimates.

1. In this thesis, we use capital $\dot{\mathbf{x}} = \boldsymbol{f}(\mathbf{x})$ for *vectors* in $\mathbb{R}^{n_\mathbf{x}}$ and slanted $\dot{\boldsymbol{x}}(t) = \boldsymbol{f}(\boldsymbol{x}(t))$ for *functions* of time.

Theorem 1.2 (Stiff existence and uniqueness [Deu87]). *Let $\boldsymbol{f} \in \mathcal{C}^1(\mathcal{X})$, $\mathcal{X} \subseteq \mathbb{R}^n$. For the Jacobian $\mathbf{A} := \boldsymbol{f}'(\mathbf{x}_0)$, assume a one-sided Lispchitz condition*

$$\langle \mathbf{u}, \mathbf{A}\mathbf{u} \rangle \leq \mu \|\mathbf{u}\|^2 , \tag{1.4a}$$

where $\langle \cdot, \cdot \rangle$ denotes an inner product in $\mathbb{R}^n$, and $\|\cdot\|$ the associated norm. Assume that

$$\left\|\boldsymbol{f}(\mathbf{x})\right\| \leq L_0, \qquad \forall \mathbf{x} \in \mathcal{X}, \tag{1.4b}$$

$$\left\|\boldsymbol{f}'(\mathbf{u}) - \boldsymbol{f}'(\mathbf{v})\right\| \leq L_2 \|\mathbf{u} - \mathbf{v}\| , \qquad \forall \mathbf{u}, \mathbf{v} \in \mathcal{X}. \tag{1.4c}$$

Then, for $\mathcal{X}$ sufficiently large, existence and uniqueness of the solution of (1.2) holds for

$$h \text{ unbounded} \quad \text{if} \quad \mu\bar{h} < -1 \qquad \text{and} \qquad h \leq \bar{h}\Psi(\mu\bar{h}) \quad \text{if} \quad \mu\bar{h} > -1, \tag{1.4d}$$

$$\text{where} \quad \bar{h} := \frac{1}{\sqrt{2L_0L_2}}, \qquad \text{and} \qquad \Psi(x) := \begin{cases} \frac{1}{x}\ln(1+x) & x \neq 0, \\ 1 & x = 0. \end{cases} \tag{1.4e}$$

1.1.2 Lyapunov stability and LaSalle invariance principle

We recall results regarding Lyapunov stability for autonomous dynamical systems of the form

$$\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}), \tag{1.5}$$

about an equilibrium point $\bar{\mathbf{x}} \in \mathcal{X}$, where $\boldsymbol{f} : \mathcal{X} \to \mathbb{R}^n$ is locally Lipschitz. Without loss of generality, one can consider systems for which the equilibrium point is zero [2]. *Definitions and properties presented below are for systems whose equilibrium point is the origin.*

Definition 1.2 (Lyapunov stability ([KG02] p.112)). The equilibrium point $\bar{\mathbf{x}} = \mathbf{0}$ of (1.5) is

- *Stable* if, for all $\epsilon > 0$, there exists $\delta_\epsilon > 0$ such that

$$\left\|\boldsymbol{x}(0)\right\| < \delta_\epsilon \implies \left\|\boldsymbol{x}(t)\right\| < \epsilon, \quad \forall t \geq 0, \tag{1.6a}$$

- *Unstable* if it is not stable,

- *Locally Asymptotically Stable* (LAS) if it is stable and δ can be chosen such that

$$\left\|\boldsymbol{x}(0)\right\| < \delta \implies \lim_{t\to\infty} \boldsymbol{x}(t) = \mathbf{0}. \tag{1.6b}$$

- *Globally Asymptotically Stable* (GAS) if it is stable for $\mathcal{X} = \mathbb{R}^n$ and if

$$\lim_{t\to\infty} \boldsymbol{x}(t) = \mathbf{0}, \quad \forall \boldsymbol{x}(0) \in \mathbb{R}^n. \tag{1.6c}$$

As illustrated on figure 1.1, oscillatory solutions can be stable in the sense of Lyapunov. The stability of a system can be proved using a *Lyapunov function* (also called a storage function).

2. Indeed the variable change $\boldsymbol{z} = \boldsymbol{x} - \bar{\mathbf{x}}$, defines an equivalent system $\dot{\boldsymbol{z}} = \boldsymbol{g}(\boldsymbol{z})$ with $\dot{\boldsymbol{z}} = \dot{\boldsymbol{x}} = \boldsymbol{f}(\bar{\mathbf{x}} + \boldsymbol{z}) =: \boldsymbol{g}(\boldsymbol{z})$, and $\boldsymbol{g}(\mathbf{0}) = \boldsymbol{f}(\bar{\mathbf{x}}) = \mathbf{0}$.

Definition 1.3 (Lyapunov function)**.** Let $\mathcal{X}$ be an open subset of $\mathbb{R}^n$ containing the equilibrium point $\bar{\mathbf{x}} = \mathbf{0}$ for (1.5). The function $V : \mathcal{X} \to \mathbb{R}$ is called a *Lyapunov function* if

C1. V is of class $\mathcal{C}^1$ on $\mathcal{X}$,

C2. $V(\bar{\mathbf{x}}) = \mathbf{0}$ and $V(\mathbf{x}) > 0$ for all $\mathbf{x} \in \mathcal{X} \setminus \{\bar{\mathbf{x}}\}$,

C3. $\nabla V(\mathbf{x}) \cdot \boldsymbol{f}(\mathbf{x}) \leq 0$ for all $\mathbf{x} \in \mathcal{X}$.

If the inequality is strict on $\mathcal{X} \setminus \{\bar{\mathbf{x}}\}$. Then, the Lyapunov function is said to be strict.

Note that, along a given trajectory of the dynamical system, one has

$$\frac{\mathrm{d}}{\mathrm{d}t} V(\boldsymbol{x}(t)) = \nabla V(\boldsymbol{x}(t)) \cdot \boldsymbol{f}(\boldsymbol{x}(t)) \leq 0.$$

Therefore, if V is a Lyapunov function, then the value of V is nonincreasing along any trajectory.

Theorem 1.3 (Lyapunov stability theorem)**.** *If there exists a Lyapunov function V for (1.5). Then, the equilibrium point $\bar{\mathbf{x}} = \mathbf{0}$ is* stable*. Moreover, if V is strict. Then, $\bar{\mathbf{x}} = \mathbf{0}$ is LAS. And if V is proper*[a]*. Then, $\bar{\mathbf{x}}$ is GAS.*

a. V is said to be proper whenever $V^{-1}([0, L])$ is a compact subset of $\mathcal{X}$ for every $L \in V(\mathcal{X})$. When $\mathcal{X} = \mathbb{R}^n$, this is equivalent to $V(x) \to +\infty$ as $\|\mathbf{x}\| \to +\infty$

The Lyapunov theorem is illustrated on Figure 1.1 for the stable, asymptotic stable and unstable cases.

When a storage function V does not satisfy all hypotheses of the Lyapunov's theorem, LaSalle's invariance principle allows useful extensions, based on the following definitions.

Definition 1.4 (Invariant set)**.** A set $\mathcal{M}$ is said to be *invariant* for a trajectory $\boldsymbol{x}(t)$ of a dynamical system (1.5) if

$$\boldsymbol{x}(0) \in \mathcal{M} \implies \boldsymbol{x}(t) \in \mathcal{M}, \ \forall t \in \mathbb{R}. \tag{1.7a}$$

If a solution belongs to $\mathcal{M}$ at a given instant. Then, it belongs to $\mathcal{M}$ for all past and future instants. It is said to be *positively invariant* if

$$\boldsymbol{x}(0) \in \mathcal{M} \implies \boldsymbol{x}(t) \in \mathcal{M}, \ \forall t \in \mathbb{R}^+. \tag{1.7b}$$

If a solution belongs to $\mathcal{M}$ at a given instant. Then, it belongs to $\mathcal{M}$ for all future instants.

We say that $\boldsymbol{x}(t)$ approaches a $\mathcal{M}$ as t goes to infinity, if for all $\epsilon > 0$, there is $T > 0$ such that

$$\mathrm{dist}(\boldsymbol{x}(t), \mathcal{M}) < \epsilon, \quad \forall t > T, \tag{1.8}$$

where $\mathrm{dist}(\mathbf{p}, \mathcal{M})$ denotes the shortest distance from a point $\mathbf{p}$ to a set $\mathcal{M}$

$$\mathrm{dist}(\mathbf{p}, \mathcal{M}) := \inf_{\mathbf{x} \in \mathcal{M}} \|\mathbf{p} - \mathbf{x}\| . \tag{1.9}$$

Theorem 1.4 (LaSalle invariance principle ([KG02] p. 128))**.** *Let $\Omega \in \mathcal{X}$ be a compact set that is positively invariant with respect to (1.5). Let $V : \mathcal{X} \to \mathbb{R}$ be a continuously differentiable function such that $\nabla V(\mathbf{x}) \cdot \boldsymbol{f}(\mathbf{x}) \leq 0$ in Ω. Let E be the set of all points*

$$E = \left\{ \mathbf{x} \in \Omega \mid \nabla V(\mathbf{x}) \cdot \boldsymbol{f}(\mathbf{x}) = 0 \right\}. \tag{1.10}$$

Let $\mathcal{M}$ be the largest invariant set in E. Then, every solution starting in Ω approaches $\mathcal{M}$ as $t \to \infty$.

In this case, one does not talk about stability, but about convergence. The interest of this principle is that it remains valid for non positive definite functions V.

1.1.3 Open systems and passivity

In this thesis, we have a particular interest in nonlinear *open systems* with p control inputs and p outputs, which admit the state-space representation

$$\begin{cases} \dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{u}), \\ \boldsymbol{y} = \boldsymbol{h}(\boldsymbol{x}, \boldsymbol{u}). \end{cases} \tag{1.11}$$

where $\boldsymbol{x}(t) \in \mathbb{R}^n$, $\boldsymbol{u}(t) \in \mathbb{R}^p$, $\boldsymbol{y}(t) \in \mathbb{R}^p$ are respectively the state vector, the input and the output of the system. Unfortunately Lyapunov stability theorem rarely applies (e.g. constant inputs $\boldsymbol{u}$). The notation of *passivity* is a powerful tool for the analysis of nonlinear open systems.

Definition 1.5 (Passivity ([KG02] p. 236))**.** The system (1.11) is said to be *passive* if there exists a continuously differentiable positive semidefinite function $V(\mathbf{x})$ (called the storage function) such that

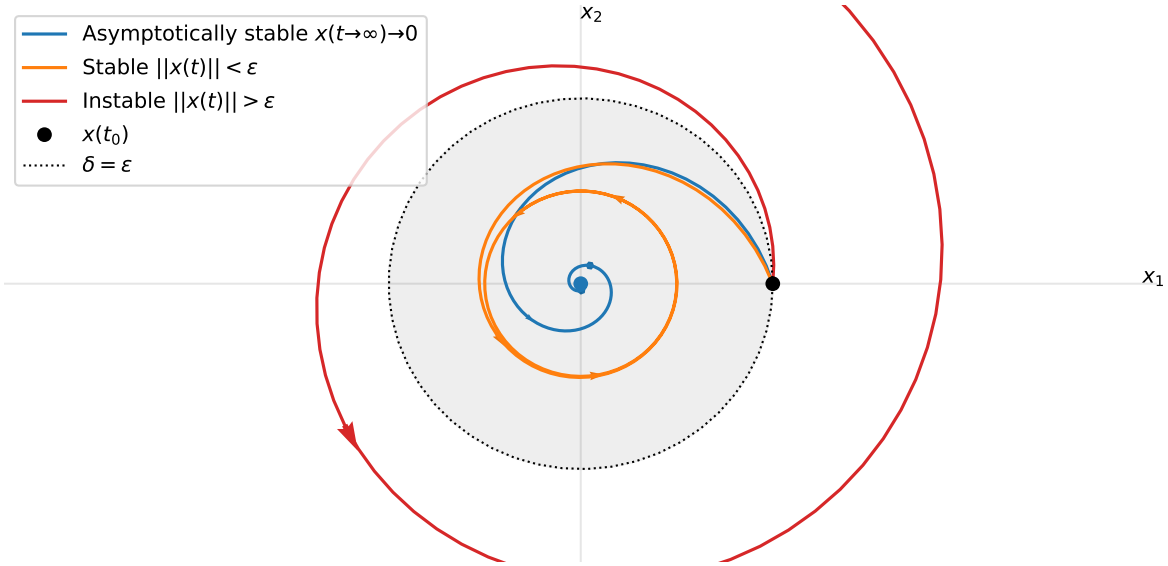
$$\langle \mathbf{u} \,|\, \mathbf{y} \rangle \geq \nabla V(\mathbf{x}) \cdot \boldsymbol{f}(\mathbf{x}, \mathbf{u}), \quad \forall (\mathbf{x}, \mathbf{u}) \in \mathbb{R}^n \times \mathbb{R}^p. \tag{1.12}$$

Moreover, it is said to be

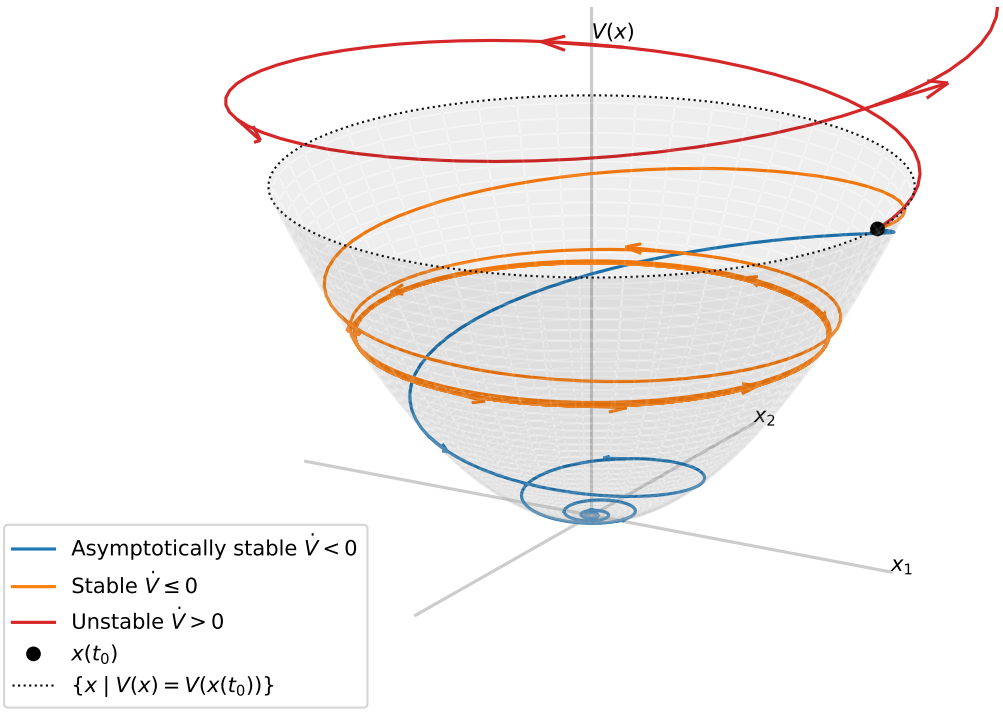
- *lossless* if $\langle \mathbf{u} \,|\, \mathbf{y} \rangle = \nabla V(\mathbf{x}) \cdot \boldsymbol{f}(\mathbf{x}, \mathbf{u})$,
- *strictly passive* if $\langle \mathbf{u} \,|\, \mathbf{y} \rangle \geq \nabla V(\mathbf{x}) \cdot \boldsymbol{f}(\mathbf{x}, \mathbf{u}) + \psi(\mathbf{x})$ for some positive definite function $\psi(\mathbf{x})$.

This definition shows that a passive system can only feed the function V through its input ports power $\langle \mathbf{u} \,|\, \mathbf{y} \rangle$. A natural candidate for this storage function is the (Hamiltonian) energy of the system under study: this vision is used throughout this thesis and is the cornerstone of Port-Hamiltonian systems.

Passivity can be related with Lyapunov stability. Indeed, when the input of a system is zero, the passivity condition implies that $\frac{\mathrm{d}}{\mathrm{d}t}(V \circ \boldsymbol{x}) \leq 0$. LaSalle invariance principle can be applied and proves that the system converges toward the largest invariant set where $\frac{\mathrm{d}}{\mathrm{d}t}(V \circ \boldsymbol{x}) = 0$. Moreover, the Lyapunov stability theorem ensures that the system is stable when V is positive definite.



(a) 2D orbits in the plane (x_1, x_2)



(b) 3D orbits in the plane $(x_1, x_2, V(\mathbf{x} = (x_1, x_2)))$

Figure 1.1 – (Lyapunov stability theorem) Stable orbits (orange), Asymptotically stable orbits (blue) and instable orbits (red). The stable orbit converges to a limit cycle for which $\frac{\mathrm{d}}{\mathrm{d}t} V(\boldsymbol{x}(t)) = 0$.

1.2 Reminder on Differential Algebraic Equations (DAE)

Results from this section are based on [Rhe90, Rei91, KM06, Hai11] and references therein. The most general form of a differential-algebraic equation over the reals is (for $m, n \in \mathbb{N}$)

$$\boldsymbol{F}(t, \boldsymbol{x}(t), \dot{\boldsymbol{x}}(t)) = \mathbf{0}, \tag{1.13a}$$

with $\boldsymbol{F} : \mathbb{I} \times \mathbb{D}_{\mathbf{x}} \times \mathbb{D}_{\dot{\mathbf{x}}} \to \mathbb{R}^m$, where $\mathbb{I} \subseteq \mathbb{R}$ is a closed interval and $\mathbb{D}_{\mathbf{x}}, \mathbb{D}_{\dot{\mathbf{x}}} \subset \mathbb{R}^n$ are open.

Existence and uniqueness of solutions are considered in the context of initial value problems, when we additionally require a solution to satisfy the condition

$$\boldsymbol{x}(t_0) = \mathbf{x}_0. \tag{1.13b}$$

Here, we recall general results about classical (continuously differentiable) and weak solutions (in the sense of distributions) of DAE of the form (1.13a) with initial condition (1.13b).

Classical solutions

Definition 1.6 (Classical DAE solution [KM06])**.** Let $\mathcal{C}^k(\mathbb{I}, \mathbb{R}^n)$ denote the vector space of all k-times continuously differentiable functions from the real interval $\mathbb{I}$ into the vector space $\mathbb{R}^n$.

1. A function $\boldsymbol{x} \in \mathcal{C}^1(\mathbb{I}, \mathbb{R}^n)$ is called a *solution* of (1.13a) if it satisfies (1.13a) pointwise.
2. The function $\boldsymbol{x} \in \mathcal{C}^1(\mathbb{I}, \mathbb{R}^n)$ is called a *solution of the initial value problem* if it additionnaly satisfies the initial condition (1.13b).
3. An initial condition (1.13b) is said to be *consistent* with $\boldsymbol{F}$, if the associated initial value problem has at least one solution.

A problem is called *solvable* if it has at least one solution.

Generalized solutions Many interesting aspects of DAEs (e.g. inconsistent initial values, impulsive solutions) can not be studied using classical solutions. Switched systems, ideal diodes, etc are common sources of non-differentiability which emphasise the need for generalised solutions beyond those of definition 1.6. To this end, consistency conditions and smoothness can be relaxed [Tre09] [3] by allowing generalized functions or distributions (with the difficulty that pointwise evaluation is not well-posed anymore, so that initial value problem cannot be formulated directly). A thorough study of *distributional DAE* is out of the scope of this thesis. We refer the reader to the references [KM06, AB08, Tre09]. However, we note that the DAE solutions of methods from chapter 5 can be interpreted as weak solutions arising from Galerkin projection in time.

1.2.1 DAE Indexes

The motivation to introduce an index is to classify different types of differential-algebraic equations with respect to the difficulty to solve them analytically as well as numerically. Several kind of DAE indexes have been introduced in the literature: differentiation index, strangeness index, perturbation index, tractability index, geometric index, structural index,etc. Their respective roles and definitions have been summarised in the overview paper [Meh12]. Here we only consider the differentiation and the perturbation indexes.

3. This reference is dedicated to *Distributional Differential Algebraic Equations* generalising the usage of weak solutions (commonly used to solve partial differential equations) to DAE.

Differentiation index

The differentiation index determines how far the differential-algebraic equation is from an ordinary differential equation (for which analysis and numerical methods are well-established).

Definition 1.7 (Differentiation Index ([Hai11] p.31))**.** Equation (1.13a) has *differentiation index* m if m is the minimal number of analytical differentiations

$$\boldsymbol{F}(t,\boldsymbol{x},\dot{\boldsymbol{x}}) = \boldsymbol{0}, \qquad \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{F}(t,\boldsymbol{x},\dot{\boldsymbol{x}}) = \boldsymbol{0}, \qquad \dots \qquad \frac{\mathrm{d}^m}{\mathrm{d}t^m}\boldsymbol{F}(t,\boldsymbol{x},\dot{\boldsymbol{x}}) = \boldsymbol{0} \tag{1.14}$$

such that equations (1.14) allows to extract by algebraic manipulations an explicit ordinary differential system $\dot{\boldsymbol{u}} = \boldsymbol{g}(\boldsymbol{u})$ (called the "underlying ODE").

Perturbation index

Complementary to the differential index, one can define the pertubation index.

Definition 1.8 (Perturbation index [TB99])**.** Equation 1.13a is said to have *perturbation index* m along a solution $\boldsymbol{x}(t)$ if m is the smallest integer such that, for all functions $\hat{\boldsymbol{x}}(t)$ having a defect $\boldsymbol{\epsilon}(t)$ given by

$$\boldsymbol{F}(t,\hat{\boldsymbol{x}},\dot{\hat{\boldsymbol{x}}}) = \boldsymbol{\epsilon}(t), \tag{1.15a}$$

there exists an estimate

$$\left\|\boldsymbol{x}(t) - \hat{\boldsymbol{x}}(t)\right\| = C\|\mathbf{x}_0 - \hat{\mathbf{x}}_0\| + \sum_{k=0}^{m-1}\sup_{\xi}\left\|\boldsymbol{\epsilon}^{(k)}(\xi)\right\|, \tag{1.15b}$$

for which the expression on the right hand side is sufficiently small and C is a constant that depends only on the function $\boldsymbol{F}$ and on the length of the time interval.

1.2.2 Semi-explicit DAEs

In this thesis, we consider semi-explicit DAE, that is systems admitting a semi-explicit form

Semi-explicit DAE with differential index-1

Consider differential-algebraic systems governed by equations of the form

$$\begin{cases} \dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x},\boldsymbol{z}), \\ \boldsymbol{0} = \boldsymbol{g}(\boldsymbol{x},\boldsymbol{z}), \end{cases} \tag{1.16}$$

with no occurrence of $\dot{\boldsymbol{z}}$. Differentiating the second equation of (1.16) with respect to time, if the matrix $\frac{\partial \boldsymbol{g}}{\partial \mathbf{z}}(\boldsymbol{x},\boldsymbol{z})$ is invertible in a neighbourhood of the solution, one obtains an ODE on $\boldsymbol{z}$.

$$\dot{\boldsymbol{z}} = -\left[\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{z}}(\boldsymbol{x},\boldsymbol{z})\right]^{-1}\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}}(\boldsymbol{x},\boldsymbol{z})\boldsymbol{f}(\boldsymbol{x},\boldsymbol{z}).$$

In practice it is not necessary to explicitly know the ODE on $\dot{\boldsymbol{z}}$: if consistent initial values satisfy $\boldsymbol{0} = \boldsymbol{g}(\mathbf{x}_0,\mathbf{z}_0)$ and if the matrix $\frac{\partial \boldsymbol{g}}{\partial \mathbf{z}}(\mathbf{x}_0,\mathbf{z}_0)$ is invertible, then the implicit function theorem

guarantees the local existence of a unique function $\boldsymbol{\zeta}(\mathbf{x})$ such that $\mathbf{0} = \boldsymbol{g}(\boldsymbol{x}, \boldsymbol{z} = \boldsymbol{\zeta}(\boldsymbol{x}))$. The problem then reduces locally to the ordinary differential equation

$$\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\zeta}(\boldsymbol{x})).$$

Existence and uniqueness of solutions can then be established using theorem 1.1 p.8.

Semi explicit DAE with differential index-2

Consider differential-algebraic systems governed by equations of the form

$$\begin{cases} \dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{z}), \\ \mathbf{0} = \boldsymbol{g}(\boldsymbol{x}). \end{cases} \tag{1.17a}$$

Here, differentiation of the second relation with respect to time leads to the hidden constraint

$$\mathbf{0} = \mathbf{g}'(\boldsymbol{x})\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{z}). \tag{1.17b}$$

If the matrix $\frac{\partial}{\partial \mathbf{z}}\left[\boldsymbol{g}'(\mathbf{x})\boldsymbol{f}(\mathbf{x}, \mathbf{z})\right]$ is invertible in a neighborhood of the solution, then $\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{z})$ and (1.17b) constitute an index 1 problem and differentiation of (1.17b) yields the missing differential equation for $\boldsymbol{z}$. If the initial values satisfy $\mathbf{0} = \boldsymbol{g}(\mathbf{x}_0)$ and $\mathbf{0} = \boldsymbol{g}'(\mathbf{x}_0)\boldsymbol{f}(\mathbf{x}_0, \mathbf{z}_0)$, we call them consistent. If in addition the matrix $\boldsymbol{g}'(\mathbf{x}_0)\frac{\partial \boldsymbol{f}}{\partial \mathbf{z}}(\mathbf{x}_0, \mathbf{z}_0)$ is invertible, the implicit function theorem implies the local existence of a function $\boldsymbol{\zeta}(\mathbf{x})$ satisfying $\mathbf{g}'(\boldsymbol{x})\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{z} = \boldsymbol{\zeta}(\boldsymbol{x})) = \mathbf{0}$ in a neighborhood of $\mathbf{x}_0$. We thus obtain a differential equation on a manifold, (see [Rhe90, Hai11])

$$\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\zeta}(\boldsymbol{x})), \qquad \text{where} \qquad \boldsymbol{x}(t) \in \mathcal{X} = \left\{\mathbf{x} \in \mathbb{R}^{n_{\mathbf{x}}} \mid \boldsymbol{g}(\mathbf{x}) = \mathbf{0}\right\}. \tag{1.17c}$$

Systems (1.17a) are called differential-algebraic equations in *Hessenberg form of index 2*.

Example 1.1 (Linear state space DAE)**.** Linear state space systems can be extended to state-space DAEs described by equations

$$\begin{cases} \mathbf{E}\,\dot{\boldsymbol{x}} = \mathbf{A}\boldsymbol{x} + \mathbf{B}\boldsymbol{u} \\ \boldsymbol{y} = \mathbf{C}\boldsymbol{x} + \mathbf{D}\boldsymbol{u} \end{cases} \tag{1.18}$$

where $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$ are matrices and $\mathbf{E}$ is a singular matrix. A typical example in electronics comes from the application of Modified Nodal Analysis [a] to VRLC circuits using node voltages as state variables. Many results are available for the class of Linear DAE stemming from the properties of the matrix pencil $(\mathbf{E}, \mathbf{A})$ (see [KM06, p. 13]).

a. The matrix $\mathbf{E}$ can be singular when the node voltages cannot all be expressed as a function of voltage sources and capacitor voltages.

1.2.3 Singular perturbations

Consider singularly perturbed systems governed by equations of the form

$$\begin{cases} \dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{z}), \\ \epsilon\,\dot{\boldsymbol{z}} = \boldsymbol{g}(\boldsymbol{x}, \boldsymbol{z}), \end{cases} \qquad \text{with} \qquad 0 < \epsilon \ll 1. \tag{1.19}$$

The limit case, $\epsilon \to 0$, yields an index one problem in semi-explicit form. This system may be proven to have an ϵ-expansion where the expansion coefficients are solution to the system of DAEs that we get in the limit of equation (1.19).

Example 1.2 (Autonomous Van der Pol oscillator [HW96])**.** The Van der Pol oscillator is governed by $\epsilon\ddot{z} + (z^2 - 1)\dot{z} + z = 0$ (in Lienard coordinates). Introducing the auxiliary variable $x := \epsilon\dot{z} + \frac{1}{3}(z^3 - z)$ yields a singular perturbation problem

$$\begin{cases} \dot{x} = -z \\ \epsilon\dot{z} = x - (\frac{1}{3}z^3 - z) \end{cases} \qquad \begin{array}{c} \text{with limit case as } \epsilon \to 0 \\ \text{the semi-explicit index-1 DAE} \end{array} \qquad \begin{cases} \dot{x} = -z \\ 0 = x - (\frac{1}{3}z^3 - z) \end{cases}$$

Differentiating the algebraic equation yields $\dot{x} = (z^2 - 1)\dot{z}$. Substituting $\dot{x} = -z$ yields a system of ODEs (where the ODE on z can be solved independently of x)

$$\dot{x} = -z, \qquad \dot{z} = -\frac{z}{z^2 - 1}.$$

1.2.4 Existence and uniqueness of solutions

A major difficulty to study existence and uniqueness of DAE is that not all of the analytical and numerical properties of differential-algebraic systems are completely understood. Several existence (and uniqueness) theories have been developed for classes of DAE with increasing levels of difficulty (and indexes). An overview can be found in [Gea71, Rhe90, Rei91, HW96, HLR06, KM06, Hai11]. General theorems for DAE of any index can be found in [KM06]. However pre-requisites are too numerous to be reproduced here.

Semi-explicit Index-1 DAE

In this thesis, we focus on semi-explicit hybrid circuit formulations (see section 2.3.3 p.57) with differential DAE index 1. This choice is motivated by the following excerpt from [dLVR13]:

Under passivity assumptions, the index of nodal models is known to be not greater than two, according to the results in [Tis98, EST00]. (...) By contrast, recent research has been focused on so-called hybrid models (...) their index does not exceed one in passive contexts [IT10, ITT12, TI10].

We have seen that for semi-explicit DAE of differential index 1 such as (1.16), one can use the implicit function theorem to establish the existence of an equivalent ODE. Then classical existence and uniqueness of DAE solutions can be obtained through the Lipschitz conditions of theorem 1.1 p.8.

Because of this, until the work of Gear [Gea71], implicit systems of the form (1.13a) were usually transformed into ordinary differential equations (1.5). However this approaches suffers from two drawbacks: 1) closed-form expression of function inverses can be either inexistent or inefficient; 2) classical existence and uniqueness theory is too restrictive on the simulation step size h for stiff ODE [4].

An alternative strategy, is to use theorem 1.2 [5] which recovers the full existence domain for linear ODE. However, as often with Newton iteration, practical conditions are not easy to obtain. It is now acknowledged that it is often preferable to develop methods that operate directly on the given differential-algebraic equations. Practical existence and uniqueness condition, exploiting particular forms of DAE, remains an important subject of research that we try to tackle in section 5.3 p.135.

4. Reduction of DAE to ODE can typically yield infinitely stiff ODE.

5. This theorem is based on functional Newton iteration rather than the fixed-point theorem. Note that in chapter 6 p.6 we use functional Newton iteration to show that exponential integrators arise as optimal Newton pre-conditioners for stiff ODE.

1.3 Introduction to port-Hamiltonian Systems

Let $\mathcal{F}$ denote spaces of flows (e.g. currents) and $\mathcal{E}$ the conjugated spaces of efforts (e.g. voltages) formally defined in definition C.19 p.284 below. From a network modelling perspective, lumped parameter physical systems are naturally described by [VdS17, p.149] (see fig. 1.2)

- *energy storing elements* described by a *storage structure* (see definition 1.18 p.25)

$$\mathcal{S} \subset \mathcal{F}_S \times \mathcal{E}_S \tag{1.20a}$$

- *memoryless passive elements* described by a *resistive structure* (def. 1.19 p.27)

$$\mathcal{R} \subset \mathcal{F}_R \times \mathcal{E}_R, \tag{1.20b}$$

- *power-conserving interconnections* formalised by a *Dirac structure* (def. 1.14 p.20)

$$\mathcal{D} \subset \mathcal{F}_S \times \mathcal{E}_S \times \mathcal{F}_R \times \mathcal{E}_R \times \mathcal{F}_P \times \mathcal{E}_P. \tag{1.20c}$$

- *external ports* to interact with their environment in the space

$$\mathcal{F}_P \times \mathcal{E}_P. \tag{1.20d}$$

A coordinate-free description of Port-Hamiltonian systems is given by the following definition.

Definition 1.9 (port-Hamiltonian System)**.** A *port-Hamiltonian System* Σ is defined by the composition (see see fig. 1.2 and definition 1.17 below)

$$\Sigma := (\mathcal{S} \parallel \mathcal{D} \parallel \mathcal{R}) \ \subset \ \mathcal{F}_P \times \mathcal{E}_P. \tag{1.21}$$

The constitutive parts of this modular framework are detailed below: Dirac structures are considered in section 1.3.1, energy storage structures in section 1.3.2, and passive memoryless elements in section 1.3.3. Finally, the PH ODE and DAE representations used in this thesis are detailed in sections 1.3.4 and 1.3.5.

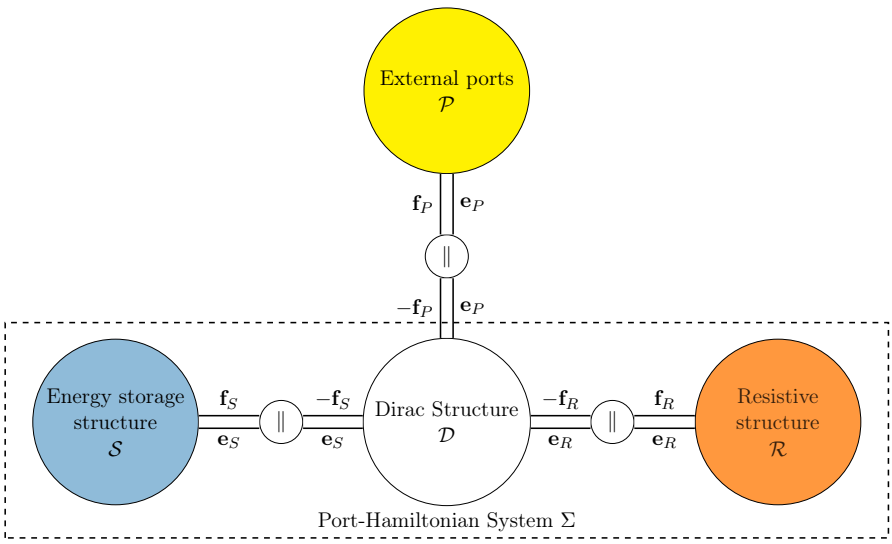


Figure 1.2 – Graphical description of a Port-Hamiltonian System.

1.3.1 Power-conserving elements (Dirac structures)

A foundation of PH modelling, is the notion of *power-conserving interconnections* which are mathematically formalised by *Dirac structures* (see electronic examples in table 1.1). The study of their mathematical formalisation and different *representations* is a key aspect in the port-Hamiltonian framework. After preliminary recalls from [VdS17, VDSJ14], we define Dirac structures, examine their matrix representations (to be used in this thesis) and their composability. Finally we comment and extend some of the examples in table 1.1.

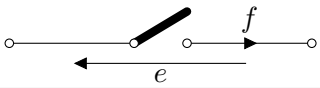
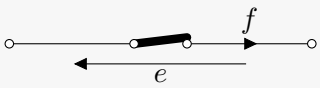
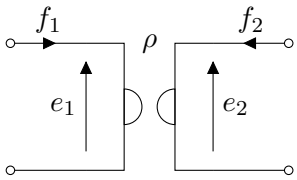
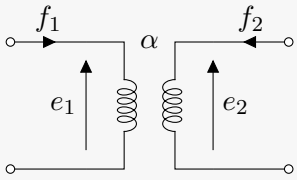
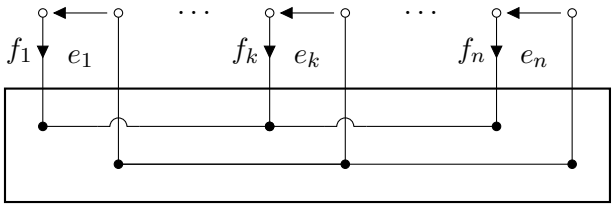
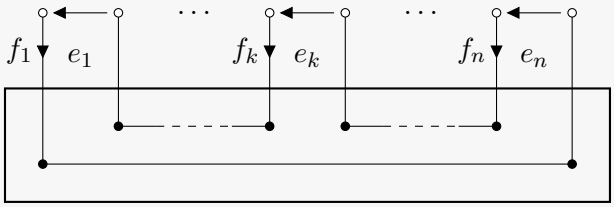
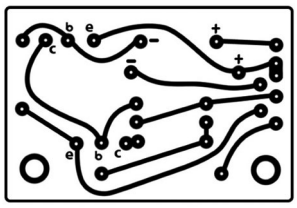
Name	Component	Equations
Open circuit		$f = 0,\ e \in \mathbb{R}$
Short circuit		$e = 0,\ f \in \mathbb{R}$
Gyrator		$\begin{bmatrix} e_1 \\ e_2 \end{bmatrix} = \begin{bmatrix} 0 & -\rho \\ \rho & 0 \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \end{bmatrix}$
Transformer		$\begin{bmatrix} e_1 \\ f_2 \end{bmatrix} = \begin{bmatrix} 0 & -\alpha \\ \alpha & 0 \end{bmatrix} \begin{bmatrix} f_1 \\ e_2 \end{bmatrix}$
Parallel connections		$e_1 = \ldots = e_n \in \mathbb{R}$ $f_1 + \ldots + f_n = 0$
Serial connections		$f_1 = \ldots = f_n \in \mathbb{R}$ $e_1 + \ldots + e_n = 0$
PCB		Kirchhoff Laws

Table 1.1 – (power-conserving Dirac structures) common examples in electronics.

Preliminary definitions

Interconnected physical systems interact through *power exchange*. Here we give definitions of flow, effort and power spaces to formalise power exchange in networked structures.

Definition 1.10 (flow and effort spaces)**.** Let $\mathcal{F}$ be a linear space (the space of flows). Its *dual space* is the set $\mathcal{E} = \mathcal{F}^*$ of linear functionals $\mathbf{e} : \mathcal{F} \to \mathbb{R}$ (the space of efforts).

Once the notion of dual flow and effort spaces is defined, one can define power as follows

Definition 1.11 (power)**.** Denote $\langle \cdot \,|\, \cdot \rangle : \mathcal{F}^* \times \mathcal{F} \to \mathbb{R}$ the *duality product* between $\mathcal{F}$ and $\mathcal{E} = \mathcal{F}^*$. The product space $\mathcal{B} := \mathcal{F} \times \mathcal{E}$ is called the space of *bonds* (or conjugated power variables), with *power* $P := \langle \mathbf{e} \,|\, \mathbf{f} \rangle$. This power is related to the quadratic form on $\mathcal{B}$

$$Q\left((\mathbf{f}, \mathbf{e})\right) := 2 \langle \mathbf{e} \,|\, \mathbf{f} \rangle, \qquad \forall (\mathbf{f}, \mathbf{e}) \in \mathcal{F} \times \mathcal{E}. \tag{1.22}$$

In this thesis, we only need $\mathcal{F} \sim \mathbb{R}^n$, (e.g. the space of currents) and its dual $\mathcal{E} \sim \mathbb{R}^n$ (e.g. the space of voltages) while $P = \langle \mathbf{e} \,|\, \mathbf{f} \rangle = \mathbf{e}^\mathsf{T} \mathbf{f}$ denotes electrical power [6].

Definition 1.12 (Canonical bilinear form)**.** The product space $\mathcal{B} = \mathcal{F} \times \mathcal{E}$, is equipped with a canonically defined symmetric bilinear form $\langle\langle \cdot, \cdot \rangle\rangle$ induced by the quadratic form Q

$$\langle\langle (\mathbf{f}_1, \mathbf{e}_1), (\mathbf{f}_2, \mathbf{e}_2) \rangle\rangle := \langle \mathbf{e}_1 \,|\, \mathbf{f}_2 \rangle + \langle \mathbf{e}_2 \,|\, \mathbf{f}_1 \rangle. \tag{1.23}$$

The bilinear form (1.23) is *indefinite* [a] but *non-degenerate* [b]. It gives $\mathcal{B}$ the structure of a *pseudo-euclidean space* (or Krein space, see C.14 p.283) equipped with $\langle \cdot, \cdot \rangle_{\mathcal{B}} := \langle\langle \cdot, \cdot \rangle\rangle$.

a. i.e. its metric matrix has both positive and negative eigenvalues (see section 1.4.2 p.36).
b. in finite dimension, this is equivalent to $\operatorname{rank}\left(\langle\langle \cdot, \cdot \rangle\rangle\right) = \dim \mathcal{B} = 2n$, (i.e. the metric is invertible).

Remark 1.2. The bilinear form $\langle\langle \cdot, \cdot \rangle\rangle$ arises from the polarization identity $\langle\langle \mathbf{u}, \mathbf{v} \rangle\rangle = \frac{1}{2}\left(Q(\mathbf{u}+\mathbf{v}) - Q(\mathbf{u}) - Q(\mathbf{v})\right)$. Indeed one easily proves using definition (1.22) that

$$\begin{aligned}\langle\langle (\mathbf{f}_1, \mathbf{e}_1), (\mathbf{f}_2, \mathbf{e}_2) \rangle\rangle &= \frac{1}{2}\Big(Q\left((\mathbf{f}_1, \mathbf{e}_1) + (\mathbf{f}_2, \mathbf{e}_2)\right) - Q\left((\mathbf{f}_1, \mathbf{e}_1)\right) - Q\left((\mathbf{f}_2, \mathbf{e}_2)\right)\Big) \\ &= \langle \mathbf{e}_1 + \mathbf{e}_2 \,|\, \mathbf{f}_1 + \mathbf{f}_2 \rangle - \langle \mathbf{e}_1 \,|\, \mathbf{f}_1 \rangle - \langle \mathbf{e}_2 \,|\, \mathbf{f}_2 \rangle = \langle \mathbf{e}_1 \,|\, \mathbf{f}_2 \rangle + \langle \mathbf{e}_2 \,|\, \mathbf{f}_1 \rangle.\end{aligned}$$

Definition 1.13 (Orthogonal complement)**.** Consider a subspace $\mathcal{D} \subset \mathcal{B} = \mathcal{F} \times \mathcal{E}$. Its orthogonal complement $\mathcal{D}^{\perp\!\!\!\perp}$ with respect to the inner product $\langle \cdot, \cdot \rangle_{\mathcal{B}}$ is defined by

$$\mathcal{D}^{\perp\!\!\!\perp} := \left\{ \mathbf{u} = (\mathbf{f}_\mathbf{u}, \mathbf{e}_\mathbf{u}) \in \mathcal{B} \;\middle|\; \langle \mathbf{u}, \mathbf{v} \rangle_{\mathcal{B}} = 0, \quad \forall \mathbf{v} = (\mathbf{f}_\mathbf{v}, \mathbf{e}_\mathbf{v}) \in \mathcal{B} \right\}. \tag{1.24}$$

Remark 1.3. If $\dim \mathcal{F} = n$, then $\dim \mathcal{B} = 2n$. Furthermore, as the bilinear form is non-degenerate, it follows that if $\dim \mathcal{D} = d$ then $\dim \mathcal{D}^{\perp\!\!\!\perp} = 2n - d$.

6. The PH framework also applies to more general spaces, possibly infinite-dimensional, to describe e.g. Partial Differential Equations (see [JZ12, DMSB09, VDSJ14]).

Dirac structures

Definition 1.14 (Dirac structure)**.** A subspace $\mathcal{D} \subset \mathcal{B} = \mathcal{F} \times \mathcal{E}$ is a (constant) *Dirac structure* if it is self-orthogonal with respect to $\langle \cdot, \cdot \rangle_{\mathcal{B}}$ (so that $\dim \mathcal{D} = \dim \mathcal{F} = \dim \mathcal{E}$.) i.e.

$$\mathcal{D} = \mathcal{D}^{\perp\!\!\!\perp}. \tag{1.25}$$

Corollary 1.1. *Let* $(\mathbf{f}, \mathbf{e}) \in \mathcal{D} = \mathcal{D}^{\perp\!\!\!\perp} \subset \mathcal{B}$*, then from equation* (1.24) *and equation* (1.23)*, a Dirac structure defines a* power conserving *relation between the variables* $(\mathbf{f}, \mathbf{e})$*, that is*

$$\langle\langle (\mathbf{f}, \mathbf{e}), (\mathbf{f}, \mathbf{e}) \rangle\rangle = 2 \langle \mathbf{e} \,|\, \mathbf{f} \rangle = 0. \tag{1.26}$$

Proposition 1.1. *A set* $\mathcal{D} \subset \mathcal{B} = \mathcal{F} \times \mathcal{E}$ *is a Dirac structure if and only if* $\langle \mathbf{e} \,|\, \mathbf{f} \rangle = 0$ *for all* $(\mathbf{f}, \mathbf{e}) \in \mathcal{D}$ *and* $\mathcal{D}$ *is a maximal subspace with this property. In particular, any subspace* $\mathcal{D} \subset \mathcal{B}$ *satisfying* $\langle \mathbf{e} \,|\, \mathbf{f} \rangle = 0$ *is a Dirac structure if and only if* $\dim \mathcal{D} = \dim \mathcal{F}$.

> **Remark 1.4.** The property $\dim \mathcal{D} = \dim \mathcal{F}$ translates that physical systems do not simultaneously impose both flow and efforts. This rules out the use of *singular network elements* in PH modelling such as *nullators* (both flow and effort are zero) and *norators* (both flow and effort are unconstrained) see references [Car64, Tel66] for more details. The *nullor* case (combination of a nullator and a norator) is considered in section 7.2 p.190.

Matrix representations

A Dirac structure $\mathcal{D} \subset \mathbb{R}^n \times \mathbb{R}^n$ can be represented in any of the following ways.

Definition 1.15 (kernel and image representations)**.** Let $\mathbf{E}, \mathbf{F} \in \mathbb{R}^{n \times n}$ satisfy

$$\mathbf{E}\mathbf{F}^\mathsf{T} + \mathbf{F}\mathbf{E}^\mathsf{T} = \mathbf{0}, \qquad \operatorname{rank} \begin{bmatrix} \mathbf{F} & \mathbf{E} \end{bmatrix} = n. \tag{1.27a}$$

- The *kernel representation* of the Dirac structure $\mathcal{D}$ is given by

$$\mathcal{D} = \left\{ (\mathbf{f}, \mathbf{e}) \in \mathbb{R}^n \times \mathbb{R}^n \;\middle|\; \mathbf{F}\mathbf{f} + \mathbf{E}\mathbf{e} = \mathbf{0} \right\} = \ker \begin{bmatrix} \mathbf{F} & \mathbf{E} \end{bmatrix}. \tag{1.27b}$$

- The *image representation*, (equivalent dual formulation) is given by

$$\mathcal{D} = \left\{ (\mathbf{f}, \mathbf{e}) \in \mathbb{R}^n \times \mathbb{R}^n \;\middle|\; \begin{bmatrix} \mathbf{f} \\ \mathbf{e} \end{bmatrix} = \begin{bmatrix} \mathbf{F}^\mathsf{T} \\ \mathbf{E}^\mathsf{T} \end{bmatrix} \boldsymbol{\lambda}, \quad \forall \boldsymbol{\lambda} \in \mathbb{R}^n \right\} = \operatorname{im} \begin{bmatrix} \mathbf{F} & \mathbf{E} \end{bmatrix}^\mathsf{T}. \tag{1.27c}$$

In short, $\mathcal{D} = \ker \begin{bmatrix} \mathbf{F} & \mathbf{E} \end{bmatrix} = \operatorname{im} \begin{bmatrix} \mathbf{F} & \mathbf{E} \end{bmatrix}^\mathsf{T}$.

Let $\mathcal{D}$ be given as in (1.27b) with $\operatorname{rank} \mathbf{F} = n_1 \leq n$. Select n_1 independent colums of $\mathbf{F}$ and partition $\mathbf{F}, \mathbf{E}, \mathbf{f}, \mathbf{e}$ into $\mathbf{F}_1, \mathbf{F}_2$, $\mathbf{E}_1, \mathbf{E}_2$, $\mathbf{f}_1, \mathbf{f}_2$, $\mathbf{e}_1, \mathbf{e}_2$ so that (1.27b) can be rewritten as

$$\begin{bmatrix} \mathbf{F}_1 & \mathbf{E}_2 \end{bmatrix} \begin{bmatrix} \mathbf{f}_1 \\ \mathbf{e}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{E}_1 & \mathbf{F}_2 \end{bmatrix} \begin{bmatrix} \mathbf{e}_1 \\ \mathbf{f}_2 \end{bmatrix} = \mathbf{0}.$$

It can be shown that $\begin{bmatrix}\mathbf{F}_1 & \mathbf{E}_2\end{bmatrix}$ is invertible so that $\mathcal{D}$ can be equivalently represented as the graph of a skew-symmetric matrix $\mathbf{J} = -\mathbf{J}^\top = -\begin{bmatrix}\mathbf{F}_1 & \mathbf{E}_2\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{E}_1 & \mathbf{F}_2\end{bmatrix}$. Conversely we have

Definition 1.16 (Hybrid skew-symmetric representation)**.** For any skew-symmetric matrix $\mathbf{J} \in \mathbb{R}^{n\times n}$, the subspace (1.28) with integers $n_1 + n_2 = n$ is a Dirac structure.

$$\mathcal{D} = \left\{ \big((\mathbf{f}_1, \mathbf{f}_2), (\mathbf{e}_1, \mathbf{e}_2)\big) \in \mathbb{R}^{n_1+n_2} \times \mathbb{R}^{n_1+n_2} \;\middle|\; \begin{bmatrix}\mathbf{f}_1\\ \mathbf{e}_2\end{bmatrix} = \mathbf{J}\begin{bmatrix}\mathbf{e}_1\\ \mathbf{f}_2\end{bmatrix} \right\}. \tag{1.28}$$

In this thesis, we use hybrid Dirac structures as our main representation (see definition 2.21 p.55).

Composition of Dirac structures

A key property of Dirac structures is their composability (see figure 1.3): the composition of two Dirac structures is again a Dirac structure so that the power-conserving interconnection of any number of Dirac structures is a Dirac structure.

Definition 1.17 (parallel/serial connection)**.** Let $\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3$ be flow spaces with dual effort spaces $\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3$. Let $\mathcal{D}_A$, $\mathcal{D}_B$ be two Dirac structures such that

$$\exists(\mathbf{f}_1, \mathbf{e}_1, \mathbf{f}_A, \mathbf{e}_A) \in \mathcal{D}_A \subset \mathcal{F}_1 \times \mathcal{E}_1 \times \mathcal{F}_2 \times \mathcal{E}_2,$$
$$\exists(\mathbf{f}_B, \mathbf{e}_B, \mathbf{f}_3, \mathbf{e}_3) \in \mathcal{D}_B \subset \mathcal{F}_2 \times \mathcal{E}_2 \times \mathcal{F}_3 \times \mathcal{E}_3,$$

with a shared space $\mathcal{F}_2 \times \mathcal{E}_2$ and a boundary space $\mathcal{F}_1 \times \mathcal{E}_1 \times \mathcal{F}_3 \times \mathcal{E}_3$. Then,

- The *parallel* connection $\mathcal{D}_A \parallel \mathcal{D}_B$ between $\mathcal{D}_A$ and $\mathcal{D}_B$ (common effort) is defined by

$$\mathbf{f}_A + \mathbf{f}_B = 0, \qquad \mathbf{e}_A = \mathbf{e}_B. \tag{1.29a}$$

- The *serial* connection $\mathcal{D}_A \circ \mathcal{D}_B$ between $\mathcal{D}_A$ and $\mathcal{D}_B$ (common flow) is defined by

$$\mathbf{e}_A + \mathbf{e}_B = 0, \qquad \mathbf{f}_A = \mathbf{f}_B. \tag{1.29b}$$

More formally,

$$\mathcal{D}_A \parallel \mathcal{D}_B := \left\{ \begin{array}{l} (\mathbf{f}_1, \mathbf{e}_1, \mathbf{f}_3, \mathbf{e}_3) \in \mathcal{F}_1 \times \mathcal{E}_1 \times \mathcal{F}_3 \times \mathcal{E}_3 \quad | \quad \exists(\mathbf{f}_2, \mathbf{e}_2) \in \mathcal{F}_2 \times \mathcal{E}_2 \\ s.t. \quad (\mathbf{f}_1, \mathbf{e}_1, \mathbf{f}_2, \mathbf{e}_2) \in \mathcal{D}_A, \quad (-\mathbf{f}_2, \mathbf{e}_2, \mathbf{f}_3, \mathbf{e}_3) \in \mathcal{D}_B \end{array} \right\},$$

$$\mathcal{D}_A \circ \mathcal{D}_B := \left\{ \begin{array}{l} (\mathbf{f}_1, \mathbf{e}_1, \mathbf{f}_3, \mathbf{e}_3) \in \mathcal{F}_1 \times \mathcal{E}_1 \times \mathcal{F}_3 \times \mathcal{E}_3 \quad | \quad \exists(\mathbf{f}_2, \mathbf{e}_2) \in \mathcal{F}_2 \times \mathcal{E}_2 \\ s.t. \quad (\mathbf{f}_1, \mathbf{e}_1, \mathbf{f}_2, \mathbf{e}_2) \in \mathcal{D}_A, \quad (\mathbf{f}_2, -\mathbf{e}_2, \mathbf{f}_3, \mathbf{e}_3) \in \mathcal{D}_B \end{array} \right\}.$$

For these definitions, we have the following result (see [VdS17])

Theorem 1.5 (Dirac structure composition)**.** *$\mathcal{D}_A \parallel \mathcal{D}_B$ and $\mathcal{D}_A \circ \mathcal{D}_B$ are Dirac structures.*

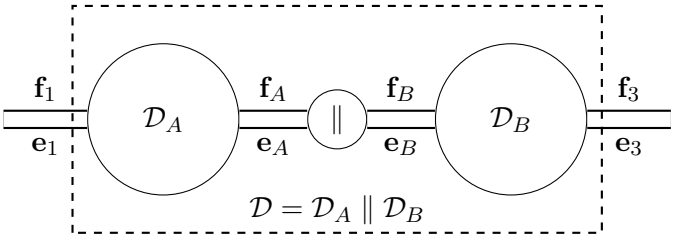


Figure 1.3 – Composition of Dirac structures (Parallel composition).

Remark 1.5. Equations (1.29a) and (1.29b) define a composition algebra so that an expression such as $(\mathcal{D}_A \parallel \mathcal{D}_B) \circ \mathcal{D}_C$ is well-defined. This key property is exploited in modular network representations such as Bondgraphs [Pay61] and Wave Digital Filters [Fet86].

Dirac structure examples

Example 1.3 (Ideal constraints)**.** Ideal flow or effort constraints such as

$$\mathcal{D}_{\mathbf{f}} = \left\{(\mathbf{f}, \mathbf{e}) \in \mathbb{R}^n \times \mathbb{R}^n \mid \mathbf{f} = \mathbf{0}\right\}, \qquad \text{or} \qquad \mathcal{D}_{\mathbf{e}} = \left\{(\mathbf{f}, \mathbf{e}) \in \mathbb{R}^n \times \mathbb{R}^n \mid \mathbf{e} = \mathbf{0}\right\}.$$

are trivial Dirac structures (in electronics: open circuits $i = 0$ or short circuits $v = 0$).

Example 1.4 ((Multi-dimensional) Transformer)**.** Transformers (see table 1.1) can be generalized to *multi-dimensional transformer* with a matrix-valued transformer ratio $\mathbf{A} \in \mathbb{R}^{n \times n}$ with flow and effort variables $(\mathbf{f}_1, \mathbf{f}_2, \mathbf{e}_1, \mathbf{e}_2) \in (\mathbb{R}^n \times \mathbb{R}^n) \times (\mathbb{R}^n \times \mathbb{R}^n)$ such that

$$\begin{bmatrix} \mathbf{e}_1 \\ \mathbf{f}_2 \end{bmatrix} = \begin{bmatrix} 0 & -\mathbf{A}^\mathsf{T} \\ \mathbf{A} & 0 \end{bmatrix} \begin{bmatrix} \mathbf{f}_1 \\ \mathbf{e}_2 \end{bmatrix}.$$

It is an instance of hybrid Dirac structure (see definition 1.16, see also [Bel68]).

Example 1.5 ((Multi-dimensional) Gyrator)**.** Similarly, a gyrator (see table 1.1) can be generalized as a *multi-dimensional gyrator* with gyration matrix $\mathbf{R} \in \mathbb{R}^{n_2 \times n_1}$ and flow and effort variables $(\mathbf{f}_1, \mathbf{f}_2, \mathbf{e}_1, \mathbf{e}_2) \in (\mathbb{R}^{n_1} \times \mathbb{R}^{n_2}) \times (\mathbb{R}^{n_1} \times \mathbb{R}^{n_2})$ such that

$$\begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \end{bmatrix} = \begin{bmatrix} 0 & -\mathbf{R}^\mathsf{T} \\ \mathbf{R} & 0 \end{bmatrix} \begin{bmatrix} \mathbf{f}_1 \\ \mathbf{f}_2 \end{bmatrix}.$$

Example 1.6 (Serial and Parallel junctions)**.** **0**-*junctions* (resp. **1**-junctions) (terminology from bond graph theory [Pay61, Bre86]), corresponds to a *parallel* (resp. serial) junctions in wave digital filters theory [Fet86]. They are defined by dual constraints: equality of efforts, and balance of flows (resp. equality of flows, and balance of efforts).

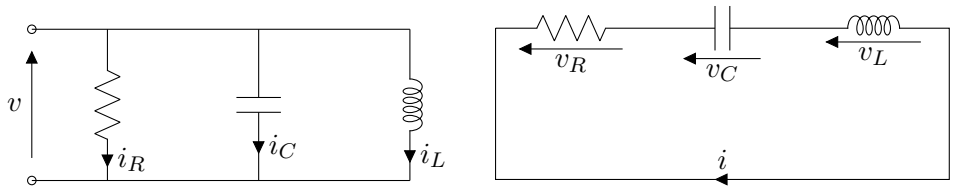
$$\text{Parallel:} \quad \mathcal{D}_{\mathbf{0}} = \left\{(\mathbf{f}, \mathbf{e}) \in \mathbb{R}^n \times \mathbb{R}^n \;\middle|\; e_1 = \ldots = e_n, \quad f_1 + \ldots + f_n = 0\right\}, \tag{1.30a}$$

$$\text{Serial:} \quad \mathcal{D}_{\mathbf{1}} = \left\{(\mathbf{f}, \mathbf{e}) \in \mathbb{R}^n \times \mathbb{R}^n \;\middle|\; f_1 = \ldots = f_n, \quad e_1 + \ldots + e_n = 0\right\}. \tag{1.30b}$$

Only one port $k \in \{1, \dots, n\}$ can be chosen to impose the *common effort* e_k (resp. flow f_k). Denoting $(\bar{\mathbf{f}}, \bar{\mathbf{e}})$ (for $i \in \{1, \dots, n\} \setminus \{k\}$) the remaining port variables, the following hybrid skew-symmetric matrix representations holds

$$\mathcal{D}_{\mathbf{0}} : \begin{bmatrix} f_k \\ \bar{e}_1 \\ \vdots \\ \bar{e}_{n-1} \end{bmatrix} = \begin{bmatrix} & -1 & \dots & -1 \\ 1 & & & \\ \vdots & & & \\ 1 & & & \end{bmatrix} \begin{bmatrix} e_k \\ \bar{f}_1 \\ \vdots \\ \bar{f}_{n-1} \end{bmatrix}, \quad \mathcal{D}_{\mathbf{1}} : \begin{bmatrix} e_k \\ \bar{f}_1 \\ \vdots \\ \bar{f}_{n-1} \end{bmatrix} = \begin{bmatrix} & -1 & \dots & -1 \\ 1 & & & \\ \vdots & & & \\ 1 & & & \end{bmatrix} \begin{bmatrix} f_k \\ \bar{e}_1 \\ \vdots \\ \bar{e}_{n-1} \end{bmatrix}.$$

In electronics, Kirchoff laws imply that for a parallel connection of components, voltages $v = v_R = v_C = v_L$ are equal (here efforts) and the current balance $i_R + i_C + i_L = 0$ of all branch currents is zero (conservation of charge). Dually, for a serial loop connection, dipoles share the same current $i = i_R = i_C = i_L$ and the oriented sum of branch voltages $v_R + v_C + v_L = 0$ must be zero.



1.3.2 Energy-storing elements

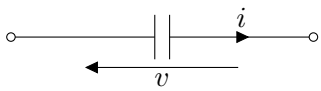
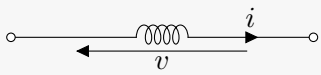
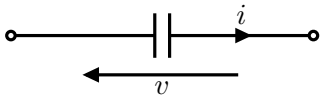
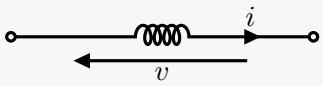
Name	Component	State	Energy	Equations
Linear Capacitor		q	$\frac{q^2}{2C}$	$i = \dot{q}, \quad v = \frac{q}{C}$
Linear Inductor		ϕ	$\frac{\phi^2}{2L}$	$v = \dot{\phi}, \quad i = \frac{\phi}{L}$
Non linear Capacitor		q	$H(q)$	$i = \dot{q}, \quad v = \nabla H(q)$
Non linear Inductor		ϕ	$H(\phi)$	$v = \dot{\phi}, \quad i = \nabla H(\phi)$

Table 1.2 – (energy storing components) examples in electronics.

In PHS (see figure 1.2 p.17), the structure $\mathcal{S}$ gathers all the energy-storing elements of the system (see examples in table 1.2). Its energy is defined on a *state space* $\mathcal{X}$ (a vector space or a manifold [7]) by a storage function called the *Hamiltonian*

$$H : \mathcal{X} \to \mathbb{R}.$$

Let $\boldsymbol{x}$ be a trajectory. For a given t, denote $\mathbf{x} = \boldsymbol{x}(t) \in \mathcal{X}$ a point along this trajectory with derivative $\dot{\mathbf{x}} = \dot{\boldsymbol{x}}(t)$. By convention, the incoming flow $\mathbf{f}_S$ and internal effort $\mathbf{e}_S$ are defined [8] by

$$\mathbf{f}_S := \dot{\mathbf{x}} \in \mathcal{F}_S := \mathsf{T}_{\mathbf{x}}\mathcal{X}, \qquad \text{and} \qquad \mathbf{e}_S := \frac{\partial H}{\partial \mathbf{x}}(\mathbf{x}) \in \mathcal{E}_S := \mathsf{T}^*_{\mathbf{x}}\mathcal{X}, \tag{1.31}$$

so that the time-variation of the stored energy is the received power

$$\frac{\mathrm{d}}{\mathrm{d}t} H(\boldsymbol{x}(t)) = \left\langle \nabla H(\mathbf{x}) \,\middle|\, \dot{\mathbf{x}} \right\rangle = \langle \mathbf{e}_S \,|\, \mathbf{f}_S \rangle, \tag{1.32}$$

where $\mathsf{T}_{\mathbf{x}}\mathcal{X}$ and $\mathsf{T}^*_{\mathbf{x}}\mathcal{X}$ denote the tangent space and co-tangent space at $\mathbf{x}$.

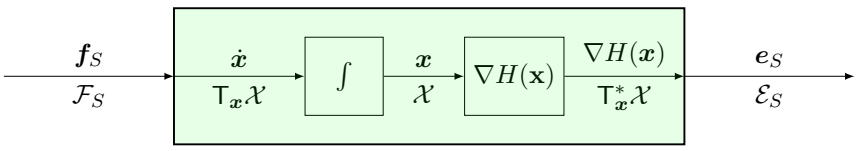


Figure 1.4 – Block diagram of energy storing elements.

7. In this manuscript the state space manifold is always $\mathcal{X} \sim \mathbb{R}^n$, so that $\mathsf{T}_{\mathbf{x}}\mathcal{X} \times \mathsf{T}^*_{\mathbf{x}}\mathcal{X} \sim \mathbb{R}^n \times \mathbb{R}^n$

8. Note that we use a different sign convention from [VDSJ14], here $(\mathbf{f}_S, \mathbf{e}_S)$ denotes the port variables of the storage structure $\mathcal{S}$ instead of the port variables of the Dirac structure $\mathcal{D}$ that are connected to storage ports.

One can sum up the above equations (see also figure 1.4) with the following definition

Definition 1.18 (Energy storage structure)**.** Let $\mathcal{X}$ be a state space (a linear space or a manifold) and $H : \mathcal{X} \to \mathbb{R}$ a Hamiltonian function. Flow and effort spaces are the tangent space $\mathcal{F}_S := \mathsf{T}_{\mathbf{x}}\mathcal{X}$ and co-tangent space $\mathcal{E}_S := \mathsf{T}^*_{\mathbf{x}}\mathcal{X}$. An *energy storage structure* is defined locally by

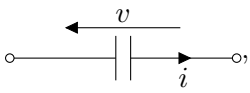
$$\mathcal{S}_{\mathbf{x}} := \left\{(\mathbf{f}_S, \mathbf{e}_S) \in \mathcal{F}_S \times \mathcal{E}_S \quad | \quad \mathbf{e}_S = \nabla H(\mathbf{x})\right\} \tag{1.33a}$$

where $\mathbf{x} \in \mathcal{X}$ denotes the current value of the trajectory

$$\boldsymbol{x}(t) = \int_{-\infty}^{t} \boldsymbol{f}_S(\tau)\,\mathrm{d}\tau. \tag{1.33b}$$

Remark 1.6 (Lagrangian submanifolds)**.** It is possible to generalise energy-storage structures using Lagrangian submanifolds (see reference [VdSM18] for the general theory and [GHVdSR20] for their use in circuit simulation). In this thesis, we do not use such generalisations, and thus skip their presentation.

Examples of storage structures

Example 1.7 (Capacitors)**.** For a capacitor 

, with energy storage function $H : \mathbb{R} \to \mathbb{R}$, the energy variable is the charge (see [CDK87, eq.1.2a])

$$q(t) := \int_{-\infty}^{t} i(\tau)\,\mathrm{d}\tau,$$

with the storage structure

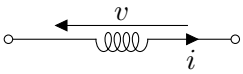
$$\mathcal{S}_q = \left\{(v, i) \in \mathbb{R}^2 \mid v = \nabla H(q)\right\}.$$

a) If the capacitor is linear, $v = \hat{v}(q) = \frac{q}{C}$, by integration we obtain the energy

$$H(q) = \int_0^q \hat{v}(x)\,\mathrm{d}x = \frac{q^2}{2C}.$$

b) If instead the capacitor is nonlinear, for example the saturating law $\hat{v}(q) = V_0 \operatorname{asinh}\left(\frac{q}{q_0}\right)$, then integrating the law we obtain the nonlinear energy

$$H(q) = \int_0^q \hat{v}(x)\,\mathrm{d}x = V_0 q_0 \left(1 + \frac{q}{q_0}\operatorname{asinh}\left(\frac{q}{q_0}\right) - \sqrt{1 + \left(\frac{q}{q_0}\right)^2}\right).$$

Example 1.8 (Inductors)**.** For an inductor 

, with energy storage function $H : \mathbb{R} \to \mathbb{R}$, the energy variable is the flux-linkage $\phi(t) := \int_{\infty}^{t} v(\tau)\,\mathrm{d}\tau$, (cf [CDK87, eq. 1.2b]) and the storage structure

$$\mathcal{S}_\phi = \left\{ (v, i) \in \mathbb{R}^2 \mid i = \nabla H(\phi) \right\}.$$

a) If the inductor is linear, $i = \hat{i}(\phi) = \frac{\phi}{L}$, then by integration, its energy is the quadratic potential

$$H(\phi) = \int_0^{\phi} \hat{i}(x)\,\mathrm{d}x = \frac{\phi^2}{2L}.$$

b) If the inductor is nonlinear, for example the anti-saturating law $\hat{i}(\phi) = I_0 \sinh\left(\frac{\phi}{\phi_0}\right)$, then by integration we obtain the non linear energy

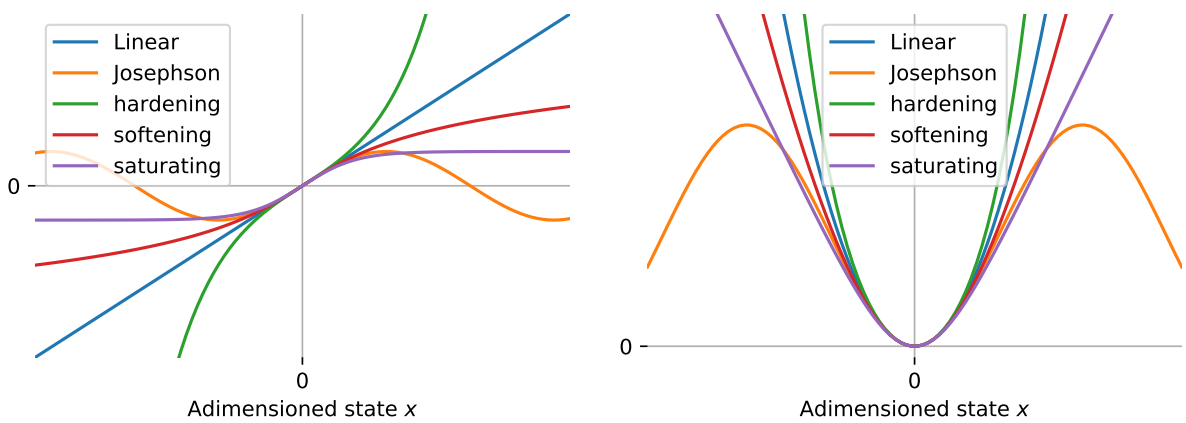
$$H(\phi) = \int_0^{\phi} \hat{i}(x)\,\mathrm{d}x = \phi_0 I_0 \left(\cosh\left(\frac{\phi}{\phi_0}\right) - 1 \right).$$

c) Another example of flux-controlled inductor with a nonlinear / non-bijective ϕ–i law is the Josephson junction [CDK87, ex.3b] governed by $\hat{i}(\phi) = I_0 \sin\left(\frac{\phi}{\phi_0}\right)$ with a positive but non-convex Hamiltonian [a].

$$H(\phi) = I_0 \phi_0 \left(1 - \cos\left(\frac{\phi}{\phi_0}\right) \right).$$

a. I_0 is a device parameter, $\phi_0 = \hbar/(2e)$, $\hbar = h/(2\pi)$ the reduced Planck constant and e the electron charge.

A typology of typical effort laws and their corresponding energy potentials (for linear, hardening, softening, saturating or oscillating laws), is shown on figure 1.5.



(a) Effort $e = \nabla H(x)$ **(b)** Energy $H(x)$

Figure 1.5 – Examples of adimensioned effort laws and their corresponding energies.

1.3.3 Passive memoryless elements

The second type of multi-port element $\mathcal{R}$ corresponds to energy dissipation (friction, resistance) or more generally to *passive memoryless elements* (examples shown in figure 1.3).

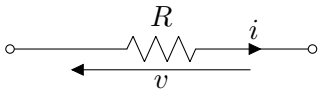
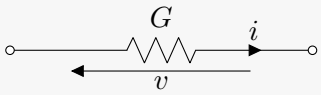
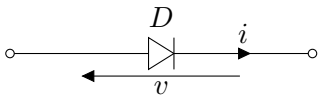
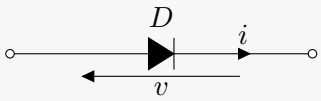
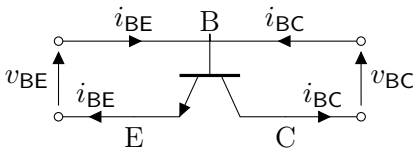
Name	Component	$\boldsymbol{w}$	Law $\boldsymbol{w}^* = \boldsymbol{z}(\boldsymbol{w})$
Resistor		i	$v = Ri$
Conductor		v	$i = Gv$
Shockley Diode		v	$i = \mathrm{pn}(v)$ (see (1.42))
Ideal Diode		v	$i \in \begin{cases} \{0\} & v \in \mathbb{R}^- \setminus \{0\} \\ \mathbb{R}^+ & v = 0 \end{cases}$
BJT		$\begin{bmatrix} v_{\mathsf{BC}} \\ v_{\mathsf{BE}} \end{bmatrix}$	$\begin{bmatrix} i_{\mathsf{BC}} \\ i_{\mathsf{BE}} \end{bmatrix} = \begin{bmatrix} \gamma_R & -1 \\ -1 & \gamma_F \end{bmatrix} \begin{bmatrix} \mathrm{pn}(v_{\mathsf{BC}}) \\ \mathrm{pn}(v_{\mathsf{BE}}) \end{bmatrix}$

Table 1.3 – (passive memoryless components) Examples in electronics. All components are dissipative except the ideal diode which is *non-energetic* (and multi-valued).

A *memoryless passive relation* (or a *resistive relation*) is given by the following definition

Definition 1.19 (Resistive relation)**.** Let $\mathcal{F}_R$ be a vector space with dual $\mathcal{E}_R = \mathcal{F}_R^*$. A *resistive relation* $\mathcal{R}$ is a subset $\mathcal{R} \subset \mathcal{F}_R \times \mathcal{E}_R$ defined by

$$\mathcal{R} := \left\{ (\mathbf{f}_R, \mathbf{e}_R) \in \mathcal{F}_R \times \mathcal{E}_R \mid \langle \mathbf{e}_R \,|\, \mathbf{f}_R \rangle \geq 0 \right\}. \tag{1.34}$$

with $\dim \mathcal{R} = \dim \mathcal{F}$.

Note that, it defines a *passive relation* that is neither over nor under determined, but can be multi-valued (see appendix A p.271). Following reference [RB16], we overload function notation and write $\mathcal{R}(\mathbf{f})$ to mean the set

$$\mathcal{R}(\mathbf{f}) = \left\{ \mathbf{e} \in \mathcal{E}_R \mid (\mathbf{f}, \mathbf{e}) \in \mathcal{R} \right\}. \tag{1.35}$$

We define the domain and image of a relation by $\mathrm{dom}\,\mathcal{R} := \left\{ \mathbf{f} \in \mathcal{F}_R \mid \mathcal{R}(\mathbf{f}) \neq \emptyset \right\}$, and $\mathrm{im}\,\mathcal{R} := \cup_{\mathbf{f} \in \mathrm{dom}\,\mathcal{R}} \mathcal{R}(\mathbf{f})$. Some important properties to describe relations are presented below.

Definition 1.20 (Relation properties)**.** A relation $\mathcal{R}$ (possibly multivalued) is said to be

- *passive* or *resistive* (resp. *strictly resistive*) if $\exists\, m \geq 0$ (resp. $m > 0$) such that

$$\left\langle \mathcal{R}(\mathbf{f}) \,\middle|\, \mathbf{f} \right\rangle \geq m, \quad \forall \mathbf{f} \in \operatorname{dom} \mathcal{R}, \tag{1.36a}$$

- *monotone* or *incrementally passive* when

$$\left\langle \mathcal{R}(\mathbf{f}_2) - \mathcal{R}(\mathbf{f}_1) \,\middle|\, \mathbf{f}_2 - \mathbf{f}_1 \right\rangle \geq 0, \quad \forall \mathbf{f}_1, \mathbf{f}_2 \in \operatorname{dom} \mathcal{R}, \tag{1.36b}$$

- *strongly monotone* or *coercive* when there exists $m > 0$ such that

$$\left\langle \mathcal{R}(\mathbf{f}_2) - \mathcal{R}(\mathbf{f}_1) \,\middle|\, \mathbf{f}_2 - \mathbf{f}_1 \right\rangle \geq m \|\mathbf{f}_2 - \mathbf{f}_1\|^2, \quad \forall \mathbf{f}_1, \mathbf{f}_2 \in \operatorname{dom} \mathcal{R}, \tag{1.36c}$$

- *one-sided Lipschitz* when there exist $L > 0$ such that

$$\left\langle \mathcal{R}(\mathbf{f}_1) - \mathcal{R}(\mathbf{f}_2) \,\middle|\, \mathbf{f}_2 - \mathbf{f}_1 \right\rangle \leq L \|\mathbf{f}_2 - \mathbf{f}_1\|^2, \quad \forall \mathbf{f}_1, \mathbf{f}_2 \in \operatorname{dom} \mathcal{R}, \tag{1.36d}$$

- *Lipschitz* when there exist $L > 0$ such that

$$\left\| \mathcal{R}(\mathbf{f}_2) - \mathcal{R}(\mathbf{f}_2) \right\| \leq L \|\mathbf{f}_2 - \mathbf{f}_1\|, \quad \forall \mathbf{f}_1, \mathbf{f}_2 \in \operatorname{dom} \mathcal{R}. \tag{1.36e}$$

Explicit mappings Let $(\mathcal{W}, \mathcal{W}^*)$ denote (possibly hybrid) flow-effort spaces induced by a suitable permutation among the coordinates of flow and effort spaces $(\mathcal{F}_R, \mathcal{E}_R)$. In the majority of cases, resistive relations can be defined by the graph of an explicit mapping $\boldsymbol{z} : \mathbf{w} \mapsto \mathbf{w}^* = \boldsymbol{z}(\mathbf{w})$ where $\boldsymbol{z}$ is a dissipative operator satisfying the power-balance.

$$\left\langle \boldsymbol{z}(\mathbf{w}) \,\middle|\, \mathbf{w} \right\rangle \geq 0. \tag{1.37}$$

Linear Resistive relations Linear resistive elements are characterized by linear mappings of the form $\boldsymbol{z}(\mathbf{w}) = \mathbf{A}\mathbf{w}$ with positive semi-definite matrix $\mathbf{A}$ (i.e. $\mathbf{A} + \mathbf{A}^\mathsf{T} \succeq 0$). For example, pure resistance ($\mathbf{v} = \mathbf{R}\mathbf{i}$) or conductance ($\mathbf{i} = \mathbf{G}\mathbf{v}$) relations are characterised by symmetric positive definite matrices ($\mathbf{R} = \mathbf{R}^\mathsf{T} \succ 0$, $\mathbf{G} = \mathbf{G}^\mathsf{T} \succ 0$).

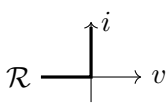
Implicit parametrisation Multi-valued or non monotone relations (e.g ideal or tunnel diodes) may be easier to describe using implicit parametrisations.

Definition 1.21 (Implicit resistive relation)**.** Denote $\Lambda = \mathbb{R}^n$ with $\mathcal{F}_R = \mathbb{R}^n = \mathcal{E}_R$ and let $\boldsymbol{E} : \Lambda \to \mathcal{E}_R$, $\boldsymbol{F} : \Lambda \to \mathcal{F}_R$ be two algebraic operators. If $\left\langle \boldsymbol{E}(\boldsymbol{\lambda}) \,\middle|\, \boldsymbol{F}(\boldsymbol{\lambda}) \right\rangle \geq 0$, for all $\boldsymbol{\lambda} \in \Lambda$, the set $\mathcal{R}$ is called an *implicit resistive structure* in image parametrisation, where

$$\mathcal{R} = \left\{ (\boldsymbol{F}(\boldsymbol{\lambda}), \boldsymbol{E}(\boldsymbol{\lambda})) \in \mathcal{F}_R \times \mathcal{E}_R \quad \forall \boldsymbol{\lambda} \in \Lambda \right\}. \tag{1.38}$$

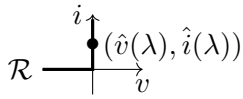
To illustrate this, consider the set-valued relation of the ideal diode from table 1.3

$$\mathcal{R} = \left\{ (v, i) \in \mathbb{R} \times \mathbb{R} \;\middle|\; i \in \begin{cases} \{0\} & v \in \mathbb{R}^- \setminus \{0\} \\ \mathbb{R}^+ & v \in \{0\} \end{cases} \right\}.$$



with $\operatorname{dom}\mathcal{R} = \mathbb{R}^-$, $\operatorname{im}\mathcal{R} = \mathbb{R}^+$. Equivalently, it can be implicitely parametrized by

$$\mathcal{R} = \left\{ (v,i) \in \mathbb{R}\times\mathbb{R} \;\middle|\; \begin{bmatrix} v \\ i \end{bmatrix} = \begin{bmatrix} -V_0\,\mathbf{1}_{\mathbb{R}^-}(\lambda) \\ I_0\,\mathbf{1}_{\mathbb{R}^+}(\lambda) \end{bmatrix}, \quad \forall\lambda\in\mathbb{R} \right\}. \tag{1.39}$$



where $\mathbf{1}_S$ denote the indicator function of a set S and V_0, I_0 can be any positive normalisation constants. $\mathcal{R}$ clearly defines a one-dimensional manifold in $\mathbb{R}\times\mathbb{R}$ that cannot be represented as a single-valued function in the (v,i) plane. The implicit parametrisation has the advantage of making the one dimensional constraint explicit, and uses *continuous single-valued functions*. This last fact is useful numerically. It has been exploited by the author in the article [MH20].

Dissipative potentials

In this thesis, we use the results from [Mil51, Che51] about dissipative potentials for simulation purposes [9]. As effort laws derive from the gradient of the Hamiltonian for storage components. In a similar manner, dissipative laws can be regarded as arising from the gradient of a "power potential" (this is related to Brayton–Moser mixed-potential theory [BM64a, BM64b, JS03]). To this end, consider the power differential

$$\mathrm{d}(\mathbf{e}\cdot\mathbf{f}) = \mathbf{e}\cdot\mathrm{d}\mathbf{f} + \mathbf{f}\cdot\mathrm{d}\mathbf{e}.$$

For an integrable resistive relation $\mathcal{R}$, define potential functions $D : \mathcal{F}_R \to \mathbb{R}$ and $D^* : \mathcal{E}_R \to \mathbb{R}$ respectively called content and co-content [10] by the line integrals

$$D(\mathbf{f}_R) := \int_{\mathbf{0}}^{\mathbf{f}_R} \boldsymbol{E}(\mathbf{f})\cdot\mathrm{d}\mathbf{f}, \qquad D^*(\mathbf{e}_R) := \int_{\mathbf{0}}^{\mathbf{e}_R} \boldsymbol{F}(\mathbf{e})\cdot\mathrm{d}\mathbf{e}, \tag{1.40}$$

so that for all $(\mathbf{f}_R, \mathbf{e}_R) \in \mathcal{R}$, integrating the differential $\mathrm{d}(\mathbf{e}\cdot\mathbf{f})$ along the path $\gamma : (\mathbf{0},\mathbf{0}) \to (\mathbf{f}_R, \mathbf{e}_R) \in \mathcal{R}$, the power is equal to the sum of content and co-content potentials

$$\mathbf{e}_R\cdot\mathbf{f}_R = D(\mathbf{f}_R) + D^*(\mathbf{e}_R), \qquad \forall(\mathbf{f}_R,\mathbf{e}_R)\in\mathcal{R}. \tag{1.41}$$

Differentiating (1.41) with respect to $(\mathbf{e}_R, \mathbf{f}_R)$ it follows from the definition that we can indeed recover efforts or flows respectively from the gradient of the content and co-content potentials.

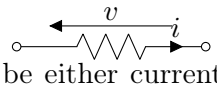
$$\mathbf{e}_R = \nabla D(\mathbf{f}_R), \qquad \text{and} \qquad \mathbf{f}_R = \nabla D^*(\mathbf{e}_R).$$

Equation (1.41) is illustrated visually on figures 1.6 and 1.7 below.

> **Remark 1.7** (Legendre transformation)**.** Content and co-content potential D and D^* are dual to each other (see figures 1.6, 1.7) and represent the same information. In the case of convex potentials, they are respectively equal to the Legendre transformation of each other
>
> $$D(\mathbf{f}_R) = \mathbf{e}_R\cdot\mathbf{f}_R - D^*(\mathbf{e}_R), \qquad D^*(\mathbf{e}_R) = \mathbf{e}_R\cdot\mathbf{f}_R - D(\mathbf{f}_R).$$
>
> Note that this is just a reformulation of (1.41). See [ZRM09] for a detailed introduction to the Legendre and Legendre–Fenchel transformations.

9. Our motivation is that in subsection 5.4.1 p.140, antiderivatives allow closed-form computation of projection coefficients. They are also useful for anti-aliasing and discrete gradient can be generalised to dissipative potentials.
10. These potential are sometimes called Rayleigh dissipation functions or current and voltage potentials

Examples of resistive structures

Linear resistor For a linear resistor, 

the resistive structure is bijective. It can be either current or voltage controlled

$$\left\{(v,i) \in \mathbb{R}^2 \,|\, v = \hat{v}(i) = Ri\right\} = \mathcal{R} = \left\{(v,i) \in \mathbb{R}^2 \,|\, i = \hat{i}(v) = \frac{v}{R}\right\}.$$

Its current and voltage potentials are respectively given by

$$D(i) = \int_0^i \hat{v}(\iota)\,\mathrm{d}\iota = \frac{Ri^2}{2}, \quad D^*(v) = \int_0^v \hat{i}(\nu)\,\mathrm{d}\nu = \frac{v^2}{2R}.$$

This is shown on figure 1.6. In this particular case (because of linearity), we have $D(i) = D^*(\hat{v}(i)) = Ri^2$, but this result should not be extrapolated as the next example shows.

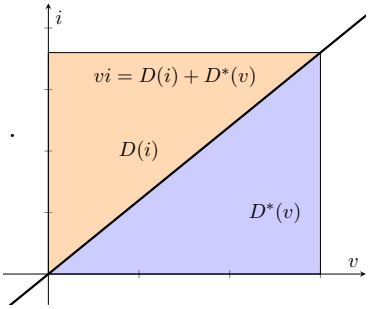
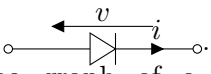


Figure 1.6 – Law of a linear resistor and its current and voltage power potentials.

PN Diode Consider the voltage controlled Shockley diode model [Sho49] 

. The resistive structure is given by the graph of a PN junction $\mathcal{R} = \left\{(v,i) \in \mathbb{R}^2 \mid i = \mathrm{pn}(v)\right\}$ with

$$\mathrm{pn}(v) := I_S\left(\exp\left(\frac{v}{nV_T}\right) - 1\right). \tag{1.42}$$

where I_S is the saturation current, n the ideality factor, $V_T = \frac{kT}{q_e}$ the thermal voltage with k the Boltzmann constant, T the temperature in Kelvin and q_e the charge of the electron. By integration, its voltage potential is given by

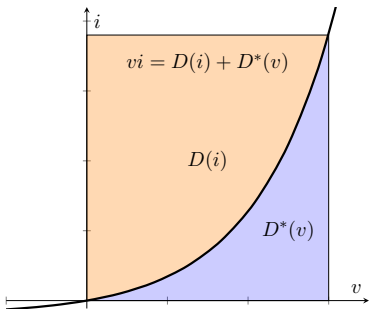


Figure 1.7 – Law of a Shockley Diode and its power potentials.

$$D^*(v) = \int_0^v \hat{i}(\nu)\,\mathrm{d}\nu = nV_T I_S\left(\exp\left(\frac{v}{nV_T}\right) - \frac{v}{nV_T} - 1\right). \tag{1.43}$$

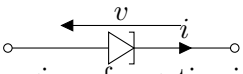
Using bijectivity, we can express the current potential indirectly by using the inverse map

$$v = \mathrm{pn}^{-1}(i) = nV_T \ln\left(1 + \frac{i}{I_S}\right), \quad i > -I_S, \tag{1.44}$$

and the Legendre transformation $D(i) = \left[vi - D^*(v)\right]_{v=\hat{i}^{-1}(i)}$ to obtain

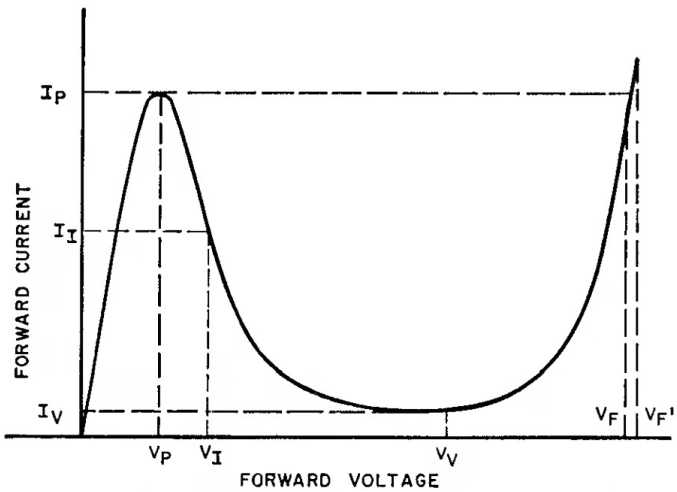
$$D(i) = nV_T I_S\left(\left(1 + \frac{i}{I_S}\right)\ln\left(1 + \frac{i}{I_S}\right) - \frac{i}{I_S}\right) \tag{1.45}$$

Using the above definitions, the current and voltage potentials being known, for simulations purposes, the component can be either flow or effort-driven (according to the constraints of circuit interconnections). On figure 1.7, the areas filled by the diode power $P(v,i)$ and the current and voltage potentials $D(i)$ and $D^*(v)$ are shown in the (v,i) plane for $I_S = 1$, $nV_T = 1$. It is geometrically clear that the current and voltage potentials are complimentary and their sum equals the power vi. It is also clear that in the nonlinear case $D(i) \neq D^*(\hat{v}(i))$.

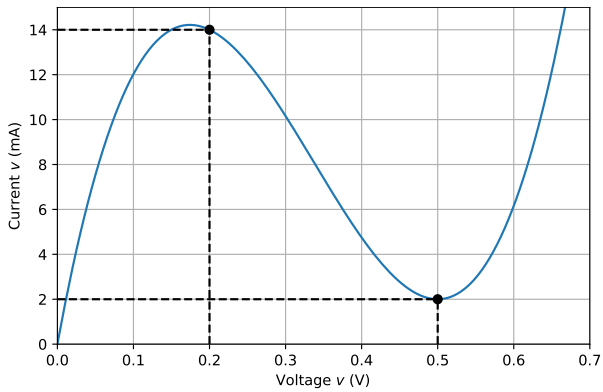
Example 1.9 (Tunnel diode)**.** A tunnel diode 

is a passive device, but its characteristic is not monotonous. It exhibits a region of negative incremental resistance. the resistive structure is given by $\mathcal{R} = \{(v, i) \in \mathbb{R}^2 \mid i = g(v)\}$ where the nonlinear characteristic g is shown on figure 1.8 with V_P the peak voltage, V_I, the inflection voltage and V_V the valley voltage. Common modelling approaches uses cubic ([NAY62, HDF+10]) or quintic ([CDK87, p.409]) polynomials. More physical approaches (see [Ng06]) use the standard PN diode model in parallel with additional terms to model the tunnel effect, the simplest being (see figure 1.8)

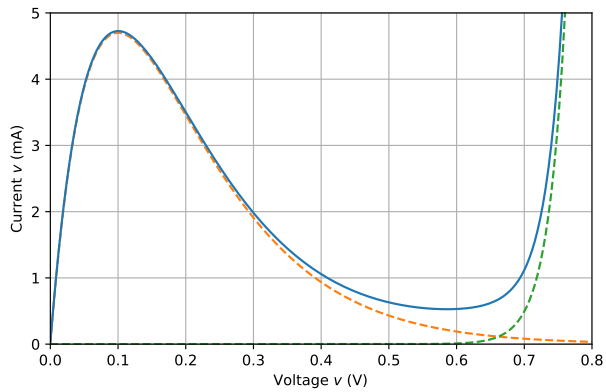
$$g(v) = \underbrace{I_S\left(e^{\frac{v}{V_T}} - 1\right)}_{\text{PN diode}} + \underbrace{I_P\left(\frac{v}{V_P}\right)e^{-\frac{v-V_P}{V_P}}}_{\text{peak current}}. \tag{1.46}$$



(a) RCA tunnel diode



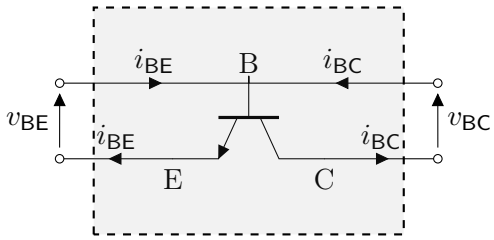
(b) Cubic approximation



(c) Exponential approximation

Figure 1.8 – Static characteristic of a tunnel diode. (a) tunnel diode plot from the RCA tunnel diode manual [RCA63]). (b) cubic approximation as used in Van der Pol oscillators, (c) exponential model.

Example 1.10 (BJT)**.** An important electronic component is the Bipolar junction transistor.



The Ebers-Moll model of a NPN Bipolar Junction Transistor, which is equivalent to two coupled PN diodes [a], can be written compactly (see Gummel–Poon article [GP70, Eq.3]) as

$$\mathcal{R}_{\mathsf{BJT}} = \left\{ \left(\begin{bmatrix} i_{\mathsf{BC}} \\ i_{\mathsf{BE}} \end{bmatrix}, \begin{bmatrix} v_{\mathsf{BC}} \\ v_{\mathsf{BE}} \end{bmatrix} \right) \in \mathbb{R}^2 \times \mathbb{R}^2 \;\middle|\; \begin{bmatrix} i_{\mathsf{BC}} \\ i_{\mathsf{BE}} \end{bmatrix} = \begin{bmatrix} \gamma_R & -1 \\ -1 & \gamma_F \end{bmatrix} \begin{bmatrix} \mathrm{pn}(v_{\mathsf{BC}}) \\ \mathrm{pn}(v_{\mathsf{BE}}) \end{bmatrix} \right\}. \tag{1.47}$$

where the parameters β_F, β_R (usually $\beta_F \approx 100$, $\beta_R \approx 20$) are respectively the forward and reverse common-emmiter curent gains. The derived parameters γ_F, γ_R are given by $\gamma = 1 + 1/\beta > 1$. Since the PHS formalism is all about explicitly formalising passive power exchange, it is important to verify before using a model that it is energetically well-posed. An original proof of passivity (not commonly found in the literature) is proposed below [b].

Proof. To prove passivity of the Ebers–Moll model, notice that function pn (see eq. (1.42)), is both passive ($\mathrm{pn}(v) \cdot v \geq 0$) and incrementally passive ($(\mathrm{pn}(v_1) - \mathrm{pn}(v_2)) \cdot (v_1 - v_2) \geq 0$). Finally, decompose the power as a sum of non-negative terms

$$\begin{aligned} \begin{bmatrix} v_{\mathsf{BC}} & v_{\mathsf{BE}} \end{bmatrix} \begin{bmatrix} i_{\mathsf{BC}} \\ i_{\mathsf{BE}} \end{bmatrix} &= \begin{bmatrix} v_{\mathsf{BC}} & v_{\mathsf{BE}} \end{bmatrix} \left(\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} + \begin{bmatrix} \gamma_R - 1 & 0 \\ 0 & \gamma_F - 1 \end{bmatrix} \right) \begin{bmatrix} \mathrm{pn}(v_{\mathsf{BC}}) \\ \mathrm{pn}(v_{\mathsf{BE}}) \end{bmatrix} \\ &= \underbrace{\big(\mathrm{pn}(v_{\mathsf{BC}}) - \mathrm{pn}(v_{\mathsf{BE}})\big)\,(v_{\mathsf{BC}} - v_{\mathsf{BE}})}_{\geq 0} + \underbrace{(\gamma_F - 1)}_{\geq 0}\underbrace{v_{\mathsf{BC}}\,\mathrm{pn}(v_{\mathsf{BC}})}_{\geq 0} + \underbrace{(\gamma_R - 1)}_{\geq 0}\underbrace{v_{\mathsf{BE}}\,\mathrm{pn}(v_{\mathsf{BE}})}_{\geq 0} \geq 0. \end{aligned}$$

□

a. see equation (1.42) for the definition of the pn function.

b. Note that this proof assumes incremental passivity with both PN junctions having the same process parameters. SPICE modelling is more flexible than that: different saturation currents and ideality factors can be used, but then proving (local) passivity becomes dependent on the particular choice of parameters.

1.3.4 Input-State-Output Representation (PH-ODE)

An important class of port-Hamiltonian systems is the structured state-space representation.

Definition 1.22 (Input-State-Output PHS [VdS17] p.113)**.** An *input-state-output port-Hamiltonian system* with n_S-dimensional state-space manifold $\mathcal{X}$, n_P-dimensional input and output spaces $\mathcal{U} \sim \mathcal{Y} = \mathbb{R}^{n_P}$, and Hamiltonian $H : \mathcal{X} \to \mathbb{R}$, is given by

$$\begin{cases} \dot{\mathbf{x}} = \left[\boldsymbol{J}(\mathbf{x}) - \boldsymbol{R}(\mathbf{x})\right] \nabla H(\mathbf{x}) + \boldsymbol{G}(\mathbf{x})\mathbf{u} \\ \mathbf{y} = \boldsymbol{G}^{\mathsf{T}}(\mathbf{x}) \nabla H(\mathbf{x}) \end{cases} \tag{1.48}$$

where matrix functions $\boldsymbol{J}(\mathbf{x})$, $\boldsymbol{R}(\mathbf{x}) \in \mathbb{R}^{n_S \times n_S}$ satisfy $\boldsymbol{J} = -\boldsymbol{J}^{\mathsf{T}}$ and $\boldsymbol{R} = \boldsymbol{R}^{\mathsf{T}} \succeq 0$.

It follows that it structurally satisfies the following passive power balance (see definition 1.5)

$$\frac{\mathrm{d}}{\mathrm{d}t}(H \circ \boldsymbol{x}) = \underbrace{\left\langle \nabla H(\boldsymbol{x}) \,\middle|\, \dot{\boldsymbol{x}} \right\rangle}_{P_S} = -\underbrace{\left\langle \nabla H(\boldsymbol{x}) \,\middle|\, \boldsymbol{R}(\boldsymbol{x}) \,\middle|\, \nabla H(\boldsymbol{x}) \right\rangle}_{P_R \geq 0} + \underbrace{\langle \boldsymbol{y} \,|\, \boldsymbol{u} \rangle}_{P_P} \leq \langle \boldsymbol{y} \,|\, \boldsymbol{u} \rangle\,, \tag{1.49}$$

meaning that storing components <u>receive</u> the power P_S, dissipative components <u>receive</u> (and dissipate) P_R and external sources <u>supply</u> P_P in a balanced manner.

> **Remark 1.8** (Receiver convention)**.** Exceptionnaly, in order to make the connection with state-space system theory easier, the power $\langle \boldsymbol{u} \,|\, \boldsymbol{y} \rangle$ in (1.49) uses the <u>emitter convention</u>. From now on (and throughout this document), we uniformly use the <u>receiver convention</u> *for all components including external ports / sources* so that power balances can be written under the canonical form
>
> $$\sum_i \langle \mathbf{e}_i \,|\, \mathbf{f}_i \rangle = 0.$$
>
> This choice is made to simplify sign conventions in automated modelling and is very common in electronics (Tellegen theorem). However it requires special care with input/outputs when using results from state-space and bond-graph theory where the emitter convention is often implied for input-output ports.

An extension of definition 1.22 for systems with direct feed-through is given by

Definition 1.23 (input-state-output PHS with feedthrough ([VdS17] p.114))**.** An *input-state-output port-Hamiltonian system with feed through* with n-dimensional state-space manifold $\mathcal{X}$, input and output spaces $\mathcal{U} \sim \mathcal{Y} = \mathbb{R}^{n_P}$, and Hamiltonian $H : \mathcal{X} \to \mathbb{R}$, is given as

$$\begin{bmatrix} \dot{\mathbf{x}} \\ \mathbf{y} \end{bmatrix} = \left[\boldsymbol{J}(\mathbf{x}) - \boldsymbol{R}(\mathbf{x})\right] \begin{bmatrix} \nabla H(\mathbf{x}) \\ \mathbf{u} \end{bmatrix}, \tag{1.50}$$

where matrix functions $\boldsymbol{J}(\mathbf{x})$, $\boldsymbol{R}(\mathbf{x}) \in \mathbb{R}^{(n_S+n_P)\times(n_S+n_P)}$ satisfy $\boldsymbol{J} = -\boldsymbol{J}^{\mathsf{T}}$ and $\boldsymbol{R} = \boldsymbol{R}^{\mathsf{T}} \succeq 0$.

likewise it satisfies the passive power balance (P_P now denotes the power <u>received</u> by sources)

$$\underbrace{\left\langle \nabla H(\mathbf{x}) \,\middle|\, \dot{\mathbf{x}} \right\rangle}_{P_S} + \underbrace{\langle \mathbf{u} \,|\, \mathbf{y} \rangle}_{P_P} = -\underbrace{\left\langle \begin{bmatrix} \nabla H(\mathbf{x}) \\ \mathbf{u} \end{bmatrix} \middle| \boldsymbol{R}(\mathbf{x}) \middle| \begin{bmatrix} \nabla H(\mathbf{x}) \\ \mathbf{u} \end{bmatrix} \right\rangle}_{P_R \geq 0} \leq 0. \tag{1.51}$$

1.3.5 Semi-explicit representation (PH-DAE)

A PHS does not always admit an explicit input-state-output representation. Moreover it is not always desirable (or may be computationally difficult) to find one. Another important representation of PHS, which is used in this thesis, is the following semi-explicit PH-DAE.

Definition 1.24 (semi-explicit PH-DAE)**.** A *semi-explicit port-Hamiltonian DAE* with n_S-dimensional state-space manifold $\mathcal{X}$ and Hamiltonian $H : \mathcal{X} \to \mathbb{R}$, resistive structure $\mathcal{R} \subset W \times W^*$ given by an explicit map $\boldsymbol{z} : W \to W^*$ with $W \sim W^* = \mathbb{R}^{n_R}$, and input output spaces $\mathcal{U} \sim \mathcal{Y} = \mathbb{R}^{n_P}$, is given by

$$\begin{bmatrix} \dot{\boldsymbol{x}} \\ \boldsymbol{w} \\ \boldsymbol{y} \end{bmatrix} = \mathbf{J} \begin{bmatrix} \nabla H(\boldsymbol{x}) \\ \boldsymbol{z}(\boldsymbol{w}) \\ \boldsymbol{u} \end{bmatrix}, \qquad \text{where} \qquad \mathbf{J} = \begin{bmatrix} \mathbf{J_{xx}} & * & * \\ \mathbf{J_{wx}} & \mathbf{J_{ww}} & * \\ \mathbf{J_{ux}} & \mathbf{J_{uw}} & \mathbf{J_{uu}} \end{bmatrix}, \tag{1.52}$$

and the $(n_S + n_R + n_P) \times (n_S + n_R + n_P)$ matrix $\mathbf{J} = -\mathbf{J}^\mathsf{T}$ (possibly depending on $\boldsymbol{x}$).

In this case, the power-balance writes as follows.

Property 1.1 (Power balance)**.** By skew-symmetry, the PH-DAE has the structured instantaneous *power balance*

$$\underbrace{\langle \nabla H(\boldsymbol{x}) \,|\, \dot{\boldsymbol{x}} \rangle}_{\text{stored power } P_S} + \underbrace{\langle \boldsymbol{z}(\boldsymbol{w}) \,|\, \boldsymbol{w} \rangle}_{\text{dissipated power } P_R \geq 0} + \underbrace{\langle \boldsymbol{u} \,|\, \boldsymbol{y} \rangle}_{\text{external power } P_P} = 0. \tag{1.53a}$$

Integrating over a time step $[t_0, t_1]$ this yields the *energy balance*

$$\Big[H\left(\boldsymbol{x}(t)\right) \Big]_{t_0}^{t_1} + \int_{t_0}^{t_1} \underbrace{P_R(t)}_{\geq 0} \,\mathrm{d}t + \int_{t_0}^{t_1} P_P(t) \,\mathrm{d}t = 0. \tag{1.53b}$$

Finally in the absence of external input, this reduces to the *passivity* relation

$$H\left(\boldsymbol{x}(t_1)\right) \leq H\left(\boldsymbol{x}(t_0)\right). \tag{1.53c}$$

Equation (1.52) can be rewritten as a semi-explicit state-space DAE (see section 1.2.2)

$$\begin{cases} \dot{\mathbf{x}} = \boldsymbol{f}(\mathbf{x}, \mathbf{w}, \mathbf{u}) & = & \mathbf{J_{xx}} \nabla H(\mathbf{x}) - \mathbf{J_{wx}^\mathsf{T}} \boldsymbol{z}(\mathbf{w}) - \mathbf{J_{ux}^\mathsf{T}} \mathbf{u} \\ \mathbf{0} = \boldsymbol{g}(\mathbf{x}, \mathbf{w}, \mathbf{u}) & = & \mathbf{w} - \left(\mathbf{J_{wx}} \nabla H(\mathbf{x}) + \mathbf{J_{ww}} \boldsymbol{z}(\mathbf{w}) - \mathbf{J_{uw}^\mathsf{T}} \mathbf{u} \right) \\ \mathbf{y} = \boldsymbol{h}(\mathbf{x}, \mathbf{w}, \mathbf{u}) & = & \mathbf{J_{ux}} \nabla H(\mathbf{x}) + \mathbf{J_{uw}} \boldsymbol{z}(\mathbf{w}) + \mathbf{J_{uu}} \mathbf{u} \end{cases}. \tag{1.54}$$

Remark 1.9 (Index-1)**.** According to definition 1.7, the DAE has differential index-1 if $\boldsymbol{g}$ is solvable for $\mathbf{w}$, i.e. if matrix $\frac{\partial \boldsymbol{g}}{\partial \mathbf{w}} = \mathbf{I} - \mathbf{J_{ww}} \boldsymbol{z}'(\mathbf{w})$ is invertible. A case that frequently arises in applications is when either $\mathbf{J_{ww}} = \mathbf{0}$ or $\boldsymbol{z}'(\mathbf{w})$ is positive definite. Then the DAE is automatically of index-1. This will be addressed for circuits in section 2.3.4 p.60.

For more details such as representation of PHS in canonical coordinates, or constrained PHS using Lagrange multipliers, we refer to [VDSJ14].

1.4 From flow-effort to wave variables

In this section, we show that flow-effort variables, can be equivalently represented by *incoming* and *outgoing* wave variables. In the Bondgraph litterature, wave variable representations of circuits have been pioneered by Paynter ([Pay61] p.268) and Breedveld ([Bre85] p.6) where they constitute an alternate choice of variables (see [SVDSMM02, SSvdSF05]). By contrast, in Wave Digital Filters (see Fettweiss [Fet86]), which is still an active research field [Bil04, WNSA15, WBSS18, BS17] in audio, wave variables are a defining feature of the formalism. A distinguishing feature of WDF is to use impedance adaptation to obtain a majority of explicit or reflection-free ports, which considerably reduces the cost of numerical simulations [11].

We first present the classical wave variable change (defined locally for each port), then we provide an alternative geometric viewpoint to show that the wave variable change naturally arise from a splitting of the bondspace $\mathcal{B}$ into an euclidean space $\mathcal{W}^+$ for incident waves and an anti-euclidean space $\mathcal{W}^-$ for outgoing waves both induced by the indefinite metric.

1.4.1 The classical wave variable change

Classically (see [Fet86]), for each port, incoming and outgoing waves $(w^+, w-)$ are introduced with a reference "resistance" R (and possibly a reference voltage V_0 for adimensionalisation [12]) by the variable change $(e, f) \leftrightarrow (w^+, w-)$

$$\begin{cases} w^+ = \dfrac{e + Rf}{V_0} \\ w^- = \dfrac{e - Rf}{V_0} \end{cases} \qquad \Longleftrightarrow \qquad \begin{cases} f = \dfrac{V_0}{R}\left(\dfrac{w^+ - w^-}{2}\right) \\ e = V_0\left(\dfrac{w^+ + w^-}{2}\right) \end{cases} \tag{1.55}$$

Multiplying e and f yields that the instantaneous power P is proportional to the difference between incoming power $\left|w^+\right|^2/2$ and the outgoing powers $\left|w^-\right|^2/2$

$$P = ef = \frac{V_0^2}{2R}\left(\frac{\left|w^+\right|^2 - \left|w^-\right|^2}{2}\right).$$

Classical choices for V_0 are:

- $V_0 = 1$ which yields the definition of *effort wave variables.*
- $V_0 = R$ which yields the definition of *flow wave variables.*
- $V_0 = \sqrt{2R}$ which yields the definition of *power wave variables.*

Note that, for the last choice, the variable change is a sequence of two power-conserving *unitary transforms*: an *hyperbolic squeeze* (with hyperbolic angle $\varphi = \ln(\sqrt{R})$) followed by a *rotation* (by angle $\theta = -\pi/4$)

$$\begin{bmatrix} w^+ \\ w^- \end{bmatrix} = \underbrace{\frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}}_{\text{rotation}} \underbrace{\begin{bmatrix} \sqrt{R} & 0 \\ 0 & 1/\sqrt{R} \end{bmatrix}}_{\text{hyperbolic rotation}} \begin{bmatrix} f \\ e \end{bmatrix}. \tag{1.56}$$

11. One can show, that in the linear case, port-adaptation automatically and structurally performs on the fly matrix inversion. This is closely related to QR decomposition using sequences of Householder reflections.

12. Note that bi-parametric waves (introduced in [BS17]) also makes use of two degrees of freedom.

1.4.2 Geometric viewpoint

We now adopt a top-down geometric viewpoint. Considering the bond space $\mathcal{B}$ equipped with the indefinite bilinear form $\langle \cdot, \cdot \rangle_{\mathcal{B}}$ (see definition 1.12), we show that it naturally splits into a positive euclidean space (for incoming waves) and a negative anti-euclidean space (for outgoing waves): wave variables emerges as a consequence of the indefinite metric (see definition C.14 p.283) induced by the duality pairing .

Following [VdS17, SVDSMM02], let $\mathcal{F}$ be a linear vector space, $\mathcal{E} := \mathcal{F}^*$ its dual output space and $\mathcal{B} = \mathcal{F} \times \mathcal{E}$ the product space where $(\mathbf{f}, \mathbf{e})$ have already been normalized. The bilinear form $\langle \cdot, \cdot \rangle_{\mathcal{B}}$ has the matrix representation (using the notation $\langle \mathbf{u} \,|\, \mathbf{A} \,|\, \mathbf{v} \rangle = \mathbf{u}^\mathsf{T} \mathbf{A} \mathbf{v}$)

$$\left\langle \begin{bmatrix} \mathbf{f}_1 \\ \mathbf{e}_1 \end{bmatrix}, \begin{bmatrix} \mathbf{f}_2 \\ \mathbf{e}_2 \end{bmatrix} \right\rangle_{\mathcal{B}} = \left\langle \begin{bmatrix} \mathbf{f}_1 \\ \mathbf{e}_1 \end{bmatrix} \middle| \begin{bmatrix} \mathbf{0}_m & \mathbf{I}_m \\ \mathbf{I}_m & \mathbf{0}_m \end{bmatrix} \middle| \begin{bmatrix} \mathbf{f}_2 \\ \mathbf{e}_2 \end{bmatrix} \right\rangle.$$

It immediately follows using the eigenvalue decomposition that

$$\begin{bmatrix} \mathbf{0}_m & \mathbf{I}_m \\ \mathbf{I}_m & \mathbf{0}_m \end{bmatrix} = \underbrace{\frac{1}{\sqrt{2}} \begin{bmatrix} \mathbf{I}_m & -\mathbf{I}_m \\ \mathbf{I}_m & \mathbf{I}_m \end{bmatrix}}_{\mathbf{U}} \underbrace{\begin{bmatrix} \mathbf{I}_m & \mathbf{0}_m \\ \mathbf{0}_m & -\mathbf{I}_m \end{bmatrix}}_{\mathbf{\Lambda}} \underbrace{\frac{1}{\sqrt{2}} \begin{bmatrix} \mathbf{I}_m & \mathbf{I}_m \\ -\mathbf{I}_m & \mathbf{I}_m \end{bmatrix}}_{\mathbf{U}^\mathsf{T}}.$$

It has m eigenvalues $+1$ and m eigenvalues -1 and thus defines an *indefinite* inner product. As in (1.56), the change of basis from flow-effort to waves is given by the rotation matrix $\mathbf{U}^\mathsf{T}$

$$\begin{bmatrix} \mathbf{w}^+ \\ \mathbf{w}^- \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} \mathbf{I}_m & \mathbf{I}_m \\ -\mathbf{I}_m & \mathbf{I}_m \end{bmatrix} \begin{bmatrix} \mathbf{f} \\ \mathbf{e} \end{bmatrix} \qquad \Longleftrightarrow \qquad \begin{bmatrix} \mathbf{f} \\ \mathbf{e} \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} \mathbf{I}_m & -\mathbf{I}_m \\ \mathbf{I}_m & \mathbf{I}_m \end{bmatrix} \begin{bmatrix} \mathbf{w}^+ \\ \mathbf{w}^- \end{bmatrix}. \tag{1.57}$$

The *scattering* representation consists in decomposing the vector $(\mathbf{f}, \mathbf{e}) \in \mathcal{F} \times \mathcal{E}$ according to the positive and negative eigenvalues. It defines respectively a *positive euclidean subspace* $\mathcal{W}^+ \sim \mathbb{R}^{m,0}$ and a *negative anti-euclidean subspace*[13] $\mathcal{W}^- \sim \mathbb{R}^{0,m}$ so that $\mathcal{W}^+ \oplus \mathcal{W}^- \sim \mathbb{R}^{m,m}$.

Definition 1.25 (Scattering subspaces ([VdS17] p. 27))**.** Any pair $(\mathcal{W}^+, \mathcal{W}^-)$ of subspaces $\mathcal{W}^+, \mathcal{W}^- \subset \mathcal{B} = \mathcal{F} \times \mathcal{E}$ is called a pair of scattering subspaces if

1. $\mathcal{W}^+ \oplus \mathcal{W}^- = \mathcal{F} \times \mathcal{E}$,
2. $\left\langle \mathbf{w}_1^+, \mathbf{w}_2^+ \right\rangle_{\mathcal{B}} > 0, \quad \forall\, \mathbf{w}_1^+, \mathbf{w}_2^+ \in \mathcal{W}^+ \setminus \mathbf{0}$,
3. $\left\langle \mathbf{w}_1^-, \mathbf{w}_2^- \right\rangle_{\mathcal{B}} < 0, \quad \forall\, \mathbf{w}_1^-, \mathbf{w}_2^- \in \mathcal{W}^- \setminus \mathbf{0}$,
4. $\langle \mathbf{w}^+, \mathbf{w}^- \rangle_{\mathcal{B}} = 0, \quad \forall\, (\mathbf{w}^+, \mathbf{w}^-) \in \mathcal{W}^+ \oplus \mathcal{W}^-$.

Any vector $(\mathbf{f}, \mathbf{e}) \in \mathcal{F} \times \mathcal{E}$ can be represented as a pair $\mathbf{w}^+ \oplus \mathbf{w}^- \in \mathcal{W}^+ \oplus \mathcal{W}^-$. The representation $(\mathbf{f}, \mathbf{e}) = \mathbf{w}^+ \oplus \mathbf{w}^-$ is called a *scattering representation* and $\mathbf{w}^\pm$ are called the *wave vectors* of the combined vector $(\mathbf{f}, \mathbf{e})$. It follows that for all $(\mathbf{f}_i, \mathbf{e}_i) = \mathbf{w}_i^+ \oplus \mathbf{w}_i^-$, $i = 1, 2$

$$\left\langle \begin{bmatrix} \mathbf{f}_1 \\ \mathbf{e}_1 \end{bmatrix}, \begin{bmatrix} \mathbf{f}_2 \\ \mathbf{e}_2 \end{bmatrix} \right\rangle_{\mathcal{B}} = \left\langle \begin{bmatrix} \mathbf{w}_1^+ \\ \mathbf{w}_1^- \end{bmatrix} \middle| \begin{bmatrix} \mathbf{I}_m & \mathbf{0}_m \\ \mathbf{0}_m & -\mathbf{I}_m \end{bmatrix} \middle| \begin{bmatrix} \mathbf{w}_2^+ \\ \mathbf{w}_2^- \end{bmatrix} \right\rangle = \left\langle \mathbf{w}_1^+, \mathbf{w}_2^+ \right\rangle_{\mathbb{R}^m} - \left\langle \mathbf{w}_1^-, \mathbf{w}_2^- \right\rangle_{\mathbb{R}^m}. \tag{1.58}$$

13. $\mathbb{R}^{p,q}$ denotes the pseudo-euclidean space with metric signature $\underbrace{1, \ldots 1}_{p}, \underbrace{-1, \ldots, -1}_{q}$

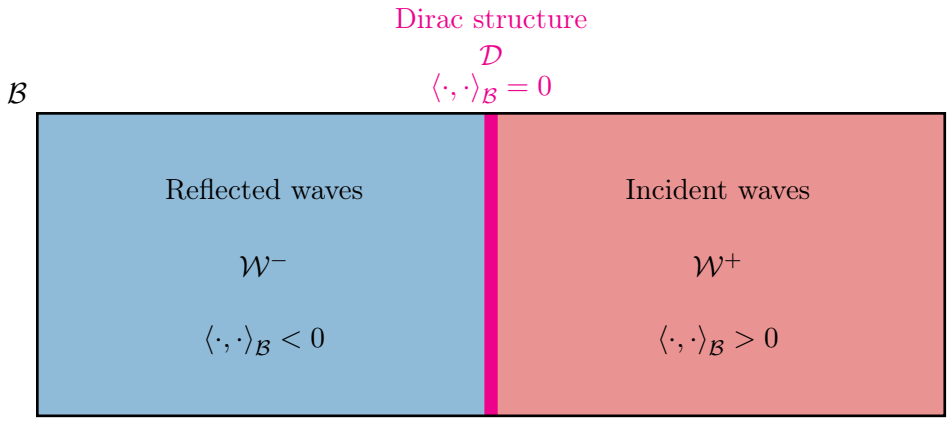


Figure 1.9 – Abstract illustration of the splitting of the (indefinite inner product) space $\mathcal{B}$ into a *positive space* $\mathcal{W}^+$, a *negative space* $\mathcal{W}^-$ and a *null space* $\mathcal{D}$.

so that (for $(\mathbf{e}_1, \mathbf{f}_1) = (\mathbf{e}_2, \mathbf{f}_2) = (\mathbf{e}, \mathbf{f}) = \mathbf{w}^+ \oplus \mathbf{w}^-$), the power writes

$$P = \langle \mathbf{e} \,|\, \mathbf{f} \rangle = \frac{1}{2} \left\langle \begin{bmatrix} \mathbf{f} \\ \mathbf{e} \end{bmatrix}, \begin{bmatrix} \mathbf{f} \\ \mathbf{e} \end{bmatrix} \right\rangle_{\mathcal{B}} = \frac{1}{2} \left(\left\| \mathbf{w}^+ \right\|^2_{\mathbb{R}^m} - \left\| \mathbf{w}^- \right\|^2_{\mathbb{R}^m} \right). \tag{1.59}$$

Remark 1.10 (Physical Units)**.** In the previous development, it is assumed that flow and effort variables $(\mathbf{f}, \mathbf{e})$ have already been scaled *to the same physical unit* so that linear combinations make sense physically. Since we also use $P = \langle \mathbf{e} \,|\, \mathbf{f} \rangle$ to denote power, for $(\tilde{\mathbf{f}}, \tilde{\mathbf{e}})$ expressed in power-conjugated natural units (e.g. Ampere and Volts), it is necessary to use a power-preserving variable change $\rho : (\tilde{\mathbf{f}}, \tilde{\mathbf{e}}) \mapsto (\mathbf{f}, \mathbf{e})$ (expressed in square root of Watt).

$$\rho : \begin{bmatrix} \mathbf{f} \\ \mathbf{e} \end{bmatrix} = \begin{bmatrix} \mathbf{R}^{1/2} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}^{-1/2} \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{f}} \\ \tilde{\mathbf{e}} \end{bmatrix}, \quad \mathbf{R} = \operatorname{diag}(R_1, \dots R_m) > 0.$$

where $R_1, \dots, R_m$ can be chosen as arbitrary scaling constants [a]. Since the variable change is power preserving, we can verify that the scaling ρ also preserves the inner product

$$\left\langle \rho \begin{bmatrix} \mathbf{f}_1 \\ \mathbf{e}_1 \end{bmatrix}, \rho \begin{bmatrix} \mathbf{f}_2 \\ \mathbf{e}_2 \end{bmatrix} \right\rangle_{\mathcal{B}} = \left\langle \begin{bmatrix} \mathbf{f}_1 \\ \mathbf{e}_1 \end{bmatrix}, \begin{bmatrix} \mathbf{f}_2 \\ \mathbf{e}_2 \end{bmatrix} \right\rangle_{\mathcal{B}}.$$

Combining variable changes, we obtain the unitary power-wave transform $(\tilde{\mathbf{f}}, \tilde{\mathbf{e}}) \mapsto (\mathbf{w}^+, \mathbf{w}^-)$.

$$\begin{bmatrix} \mathbf{w}^+ \\ \mathbf{w}^- \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} \mathbf{I} & \mathbf{I} \\ -\mathbf{I} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{R}^{1/2} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}^{-1/2} \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{f}} \\ \tilde{\mathbf{e}} \end{bmatrix}. \tag{1.60}$$

a. Recently, in reference [BMS20], the authors have proposed a vector definition of waves of the form $\mathbf{w}^\pm = \mathbf{e} \pm \mathbf{R}\mathbf{f}$ where $\mathbf{R}$ can be any invertible real matrix (not necessarily symmetric positive definite), including "across ports" linear combinations. We investigate this topic independently in section 2.5 p.73.

In section 9.4 p.254, thanks to Geometric Algebra, we revisit flow-effort and wave representation using simpler notations.

1.4.3 Wave variables representation of Port-Hamiltonian Systems

We consider the scattering representation of Dirac, storage and dissipative structures considered as causal maps $\boldsymbol{w}^+ \mapsto \boldsymbol{w}^-$. This section (mostly formal) is a step towards establishing deeper links between PHS and WDF.

Dirac structures

A Dirac structure $\mathcal{D}$ can be represented by the graph of an invertible linear map $\mathbf{S} : \mathcal{W}^+ \to \mathcal{W}^-$. This is related to the standard results from Carlin [Car64, Car67]: normal linear passive networks always possesses a scattering representation. This is summarized by the following definition.

Definition 1.26 (Scattering representation [VdS17] p.164)**.** Let $\mathcal{D} \subset \mathcal{F} \times \mathcal{E}$ be a Dirac structure, and $(\mathcal{W}^+, \mathcal{W}^-)$ scattering subspaces. The linear map $\mathbf{S} : \mathcal{W}^+ \to \mathcal{W}^-$ satisfying

$$\mathcal{D} = \left\{ (\mathbf{f}, \mathbf{e}) = \mathbf{w}^+ \oplus \mathbf{w}^- \mid \mathbf{w}^- = \mathbf{S}\mathbf{w}^+ \right\} \tag{1.61}$$

is called the *scattering representation* of $\mathcal{D}$.

For a skew-symmetric Dirac structure, we have the following proposition.

Proposition 1.2 (Scattering of skew-symmetric Dirac structure)**.** *For* (1.60)*, the scattering representation of a Dirac structure $\mathcal{D}$ given by $\mathbf{f} = \mathbf{J}\mathbf{e}$ with $\mathbf{J} = -\mathbf{J}^\mathsf{T}$, is the matrix*

$$\mathbf{S}_\mathcal{D} = (\mathbf{I} - \mathbf{J}_\mathbf{R})(\mathbf{I} + \mathbf{J}_\mathbf{R})^{-1}, \qquad \textit{where} \qquad \mathbf{J}_\mathbf{R} := \mathbf{R}^{1/2}\mathbf{J}\mathbf{R}^{1/2} = -\mathbf{J}_\mathbf{R}^\mathsf{T} \tag{1.62}$$

$\mathbf{S}_\mathcal{R}$ (the Cayley transform of $\mathbf{J}_\mathbf{R}$) is orthonormal, so that $\left\|\mathbf{S}_\mathcal{D}\mathbf{w}^+\right\|_{\mathbb{R}^n} = \left\|\mathbf{w}^+\right\|_{\mathbb{R}^n}$.

Proof. Substituting $\mathbf{f} = \mathbf{J}\mathbf{e}$ in (1.60) and factoring $\mathbf{R}^{-1/2}$ on the right, we obtain

$$\mathbf{w}^+ = \frac{1}{\sqrt{2}}\left(\mathbf{I} + \mathbf{R}^{1/2}\mathbf{J}\mathbf{R}^{1/2}\right)\mathbf{R}^{-1/2}\mathbf{e}, \qquad \mathbf{w}^- = \frac{1}{\sqrt{2}}\left(\mathbf{I} - \mathbf{R}^{1/2}\mathbf{J}\mathbf{R}^{1/2}\right)\mathbf{R}^{-1/2}\mathbf{e}.$$

Defining $\mathbf{J}_\mathbf{R} = \mathbf{R}^{1/2}\mathbf{J}\mathbf{R}^{1/2}$ and solving for $\mathbf{e}$ we obtain the map $\mathbf{w}^- = (\mathbf{I} - \mathbf{J}_\mathbf{R})(\mathbf{I} + \mathbf{J}_\mathbf{R})^{-1}\mathbf{w}^+$. □

Linear resistive relations

Using the same argument as above, we obtain

Proposition 1.3 (Scattering of linear resistive structures)**.** *For a linear resistive structure,*

$$\mathcal{R}_{lin} = \left\{ (\mathbf{e}, \mathbf{f}) \in \mathbb{R}^n \times \mathbb{R}^n \mid \mathbf{f} = \mathbf{A}\mathbf{e} \right\} \quad \textit{where} \quad \mathbf{A} \succeq m\mathbf{I},\ m \geq 0. \tag{1.63}$$

and the wave variable change (1.60)*, the scattering representation of* (1.63) *is the matrix*

$$\mathbf{S}_\mathcal{R} = (\mathbf{I} - \mathbf{A}_\mathbf{R})(\mathbf{I} + \mathbf{A}_\mathbf{R})^{-1}, \qquad \textit{where} \qquad \mathbf{A}_\mathbf{R} := \mathbf{R}^{1/2}\mathbf{A}\mathbf{R}^{1/2} \succ 0, \tag{1.64}$$

By properties of the Cayley transform, $\mathbf{S}_\mathcal{R}$ is non expansive, so that

$$\|\mathbf{S}_\mathcal{R}\mathbf{w}\|_{\mathbb{R}^n} \leq \alpha\|\mathbf{w}\|_{\mathbb{R}^n}, \quad \textit{with} \quad \alpha = |1 - m| / (1 + m). \tag{1.65}$$

Note that, when $\mathbf{A}$ is diagonal (i.e. a multiport constituted of indepedent resistors), choosing $\mathbf{R} = \mathbf{A}^{-1}$, it is possible to make the structure *reflection-free*. In this case $\mathbf{S}_\mathcal{R} = \mathbf{0}$.

Non linear multi-valued resistive relations

Following reference [RB16], it is possible to generalise the Cayley transform to nonlinear multi-valued relations. First we recall the following results [14]. Let $\boldsymbol{A}$ be a relation and $\boldsymbol{I}$ the identity relation, then for $\alpha \in \mathbb{R}$, the *resolvent* of $\boldsymbol{A}$ is $R_{\boldsymbol{A},\alpha} = (\boldsymbol{I}+\alpha\boldsymbol{A})^{-1}$ and its *Cayley operator* (see equation (A.1) p.273) is $C_{\boldsymbol{A},\alpha} = 2R_{\boldsymbol{A},\alpha} - \boldsymbol{I}$. When $\boldsymbol{A}$ is maximal and single-valued, then

$$C_{\boldsymbol{A},\alpha} = (\boldsymbol{I} - \alpha\boldsymbol{A})(\boldsymbol{I} + \alpha\boldsymbol{A})^{-1}, \quad \forall \alpha \geq 0. \tag{1.66}$$

When $\boldsymbol{A}$ is maximal monotone but not necessarily single-valued, then $C_{\boldsymbol{A}}$ satisfies

$$C_{\boldsymbol{A},\alpha}(\boldsymbol{I} + \alpha\boldsymbol{A}) = (\boldsymbol{I} - \alpha\boldsymbol{A}), \quad \forall \alpha > 0. \tag{1.67}$$

Proposition 1.4 (Scattering of resistive relations)**.** *For a resistive relation*

$$\mathcal{G} = \left\{(\mathbf{e},\mathbf{f}) \in \mathbb{R}^n \times \mathbb{R}^n \quad | \quad \mathbf{f} \in \boldsymbol{G}(\mathbf{e}), \quad \langle \mathbf{f} \,|\, \mathbf{e} \rangle \geq 0\right\}. \tag{1.68}$$

and the wave variable change (1.60)*, then its scattering representation is the Cayley operator*

$$\boldsymbol{S}_{\mathcal{G}} = 2(\boldsymbol{I} - \boldsymbol{G}_{\mathbf{R}})^{-1} - \boldsymbol{I} \qquad \textit{where} \qquad \boldsymbol{G}_{\mathbf{R}} = \mathbf{R}^{1/2}\boldsymbol{G}\mathbf{R}^{1/2} \tag{1.69}$$

According to [RB16], if $\boldsymbol{G}$ is monotone, then $\boldsymbol{S}_{\mathcal{G}}$ is nonexpansive, and if $\boldsymbol{G}$ is strongly monotone with parameter m and Lipschitz with constant L (see definition 1.20 p.28), then $\boldsymbol{S}_{\mathcal{G}}$ is a contraction with parameter

$$L_{\boldsymbol{S}_{\mathcal{G}}} = \sqrt{1 - \frac{4m}{(1+L)^2}}. \tag{1.70}$$

Storage structures

We pursue the same approach to characterise the scattering operators of storage structures. for flows and efforts evolving in the Lebesgue spaces $\mathcal{F}_S \sim \mathcal{E}_S \sim L^2(\Omega, \mathbb{R}^n)$ (over time steps Ω).

Proposition 1.5 (scattering of linear storage structure)**.** *For a linear storage structure*

$$\mathcal{S} = \left\{(\boldsymbol{f},\boldsymbol{e}) \in \mathcal{F}_S \times \mathcal{E}_S \mid \exists \boldsymbol{x} \in H^1(\Omega,\mathbb{R}^n),\ \boldsymbol{f} = \dot{\boldsymbol{x}},\ \boldsymbol{e} = \mathbf{Q}\boldsymbol{x},\ \mathbf{Q} = \mathbf{Q}^{\mathsf{T}} \succ 0\right\}, \tag{1.71}$$

the scattering representation of $\mathcal{S}$ *through* (1.60) *is the formal differential operator*

$$\boldsymbol{S}_{\mathcal{S}} = -(\mathcal{D} - \mathbf{Q}_{\mathbf{R}})(\mathcal{D} + \mathbf{Q}_{\mathbf{R}})^{-1}, \qquad \textit{where} \qquad \mathbf{Q}_{\mathbf{R}} = \mathbf{R}^{-1/2}\mathbf{Q}\mathbf{R}^{-1/2} \textit{ and } \mathcal{D} = \tfrac{\mathrm{d}}{\mathrm{d}t}. \tag{1.72}$$

Proof. Substituting the constitutive relation in (1.60) and factoring $\mathbf{R}^{1/2}$ on the right, we obtain

$$\boldsymbol{w}^+ = \frac{1}{\sqrt{2}}\left(\mathbf{R}^{-1/2}\mathbf{Q}\mathbf{R}^{-1/2} + \mathcal{D}\right)\mathbf{R}^{1/2}\boldsymbol{x}, \qquad \boldsymbol{w}^- = \frac{1}{\sqrt{2}}\left(\mathbf{R}^{-1/2}\mathbf{Q}\mathbf{R}^{-1/2} - \mathcal{D}\right)\mathbf{R}^{1/2}\boldsymbol{x}.$$

Defining $\mathbf{Q}_{\mathbf{R}} = \mathbf{R}^{-1/2}\mathbf{Q}\mathbf{R}^{-1/2}$ and solving for $\boldsymbol{x}$ we obtain $\boldsymbol{w}^- = -(\mathcal{D} - \mathbf{Q}_{\mathbf{R}})(\mathcal{D} + \mathbf{Q}_{\mathbf{R}})^{-1}\,\boldsymbol{w}^+$. □

14. For more details regarding relations, their inverse, resolvant and Cayley operator, please refer to reference [RB16] whose main results are recalled in appendix A p.271.

Note that for scalar components (e.g. for a capacitor $\mathbf{Q} = 1/C$, $\mathbf{Q}_R = 1/RC$), the Laplace transform of (1.72) (see definition C.10 p.282) yields the familiar *allpass operator*

$$H_C(s) = \mathcal{L}(\boldsymbol{S}_C) = -\frac{s - q_R}{s + q_R} = \frac{1 - sRC}{1 + sRC}, \qquad \text{so that} \qquad \left|H_C(s)\right| = 1,\ \forall s \in i\mathbb{R}.$$

In Wave Digital Filters, the Laplace variable is usually substituted by the finite difference approximation $s \approx (2/h) \cdot (1 - z^{-1})/(1 + z^{-1})$, where $z = e^{hs}$ denotes the time-shift operator, so that after substitution and using impedance adaption $R = h/2C$, we get the causal map

$$\widetilde{H}_C(z) = \frac{1 - \left(\frac{2}{h}\frac{1-z^{-1}}{1+z^{-1}}\right) RC}{1 + \left(\frac{2}{h}\frac{1-z^{-1}}{1+z^{-1}}\right) RC} = \frac{(1 + z^{-1}) - (1 - z^{-1})}{(1 + z^{-1}) + (1 - z^{-1})} = z^{-1}. \tag{1.73}$$

Numerically, this means that reflected waves $w^-[n]$ only depend on *previous incoming waves* $w^+[n-1]$, so that the numerical scheme is *explicit*.

Nonlinear storage structures

Finally, for nonlinear storage structures, we have the following formal result

Proposition 1.6 (scattering of nonlinear storage structure)**.** *For a storage structure*

$$\mathcal{S} = \left\{(\boldsymbol{f}, \boldsymbol{e}) \in \mathcal{F}_S \times \mathcal{E}_S \quad | \quad \exists \boldsymbol{x} \in H^1(\Omega, \mathbb{R}^n), \quad \boldsymbol{f} = \dot{\boldsymbol{x}}, \quad \boldsymbol{e} = \nabla H(\mathbf{x})\right\} \tag{1.74}$$

the wave variable change (1.60) *yields a scattering representation given by the formal operator*

$$\boldsymbol{S}_{\mathcal{S}} = \frac{1}{\sqrt{2}}(\mathcal{D} - \nabla H_{\mathbf{R}}) \circ (\mathcal{D} + \nabla H_{\mathbf{R}})^{-1}\sqrt{2} \quad \textit{where} \quad \nabla H_{\mathbf{R}} = \mathbf{R}^{-1/2} \circ \nabla H \circ \mathbf{R}^{-1/2}. \tag{1.75}$$

Proof. Let $\boldsymbol{f} = \dot{\boldsymbol{x}}$, $\boldsymbol{e} = \nabla H(\boldsymbol{x})$ in the wave variable change (1.60), we get

$$\frac{1}{\sqrt{2}}\left(\mathbf{R}^{1/2}\dot{\boldsymbol{x}} + \mathbf{R}^{-1/2}\nabla H(\boldsymbol{x})\right) = \boldsymbol{w}^+, \qquad \frac{1}{\sqrt{2}}\left(\mathbf{R}^{1/2}\dot{\boldsymbol{x}} - \mathbf{R}^{-1/2}\nabla H(\boldsymbol{x})\right) = \boldsymbol{w}^-.$$

Introducing $\boldsymbol{z} = \mathbf{R}^{1/2}\boldsymbol{x}$, and $\nabla H_R(\boldsymbol{z}) = \mathbf{R}^{-1/2}\nabla H(\mathbf{R}^{-1/2}\boldsymbol{z})$ yields the state-space ODE

$$\begin{cases} \dot{\boldsymbol{z}} = -\nabla H_{\mathbf{R}}(\boldsymbol{z}) + \sqrt{2}\boldsymbol{w}^+, \\ \boldsymbol{w}^- = \dfrac{1}{\sqrt{2}}(\dot{\boldsymbol{z}} - \nabla H_{\mathbf{R}}(\boldsymbol{z})). \end{cases} \iff \begin{cases} (\mathcal{D} + \nabla H_{\mathbf{R}})(\boldsymbol{z}) = \sqrt{2}\boldsymbol{w}^+, \\ \boldsymbol{w}^- = \dfrac{1}{\sqrt{2}}(\mathcal{D} - \nabla H_{\mathbf{R}})(\boldsymbol{z}). \end{cases} \tag{1.76}$$

The output equation can be further refined (by eliminating $\dot{\boldsymbol{z}}$) as $\boldsymbol{w}^- = -\sqrt{2}\nabla H_{\mathbf{R}}(\boldsymbol{z}) + \boldsymbol{w}^+$. □

A first difficulty to simulate nonlinear PHS directly from wave variables is being able to compute the inverse operator $(\mathcal{D} + \nabla H_{\mathbf{R}})^{-1}$, i.e. solving the system (1.76). To this end, numerical integration methods such the ones in chapters 4, 5, 6 can be applied (but are usually iterative, nonlinear and implicit), see also references [SVDSMM02, SSvdSF05].

A second difficulty, is making the mapping $\boldsymbol{w}^+ \mapsto \boldsymbol{w}^-$ explicit in time after discretisation as in (1.73) (which is the whole purpose of WDF). Impedance matching for PHS is also discussed in [SVDSMM02]. This non-trivial task is still an open subject for research. For this reason, in the remainder of this thesis we focus on the flow-effort representation for simulation.

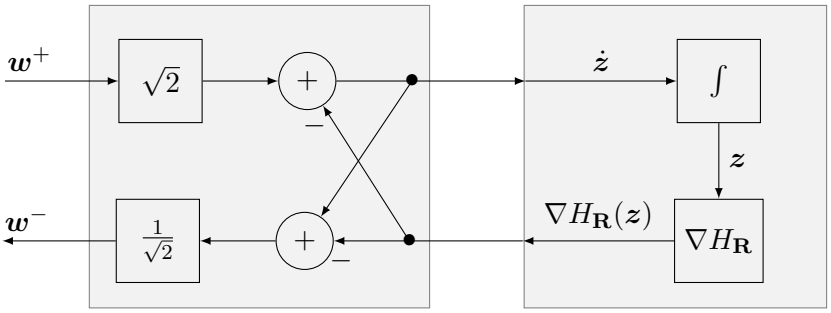


Figure 1.10 – Scattering of nonlinear storage structures (see (1.76)).

Conclusion

In this chapter, we have reviewed fundamental results about ODE, state-space systems and DAE. In particular existence and uniqueness theorems, DAE indexes, stability, Lyapunov functions and power balance. The foundations of Port-Hamiltonian Systems (Dirac structure, storage structures and resistive structures) which are required to model electronic circuits were recalled. In particular, in Part II, input-state-output PH-ODE and semi-explicit PH-DAE are the main representations used to construct numerical methods which preserves the energy balance in Equation 1.53b .

An introduction to flow-effort and wave variables representations of PHS has been detailed (in order to establishing deeper links between PHS and WDF) with an emphasis on the geometric structure of the indefinite metric bond space $\mathcal{B} \sim \mathbb{R}^{n,n}$ and its positive and negative wave polarisations $\mathcal{W}^+ \sim \mathbb{R}^{n,0}$ and $\mathcal{W}^+ \sim \mathbb{R}^{0,n}$. Special care has also been paid to include impedance-adaption (to yield causal explicit numerical schemes) and to formalise the scattering representation of Dirac structures, resistive structures and storage structures. In this context, the central tool is the Cayley transform (and its generalisation to relations and maximal monotone operators). Finding explicit time-stepping schemes through port-adaptation for nonlinear relations and storage structures is an interesting opportunity for future research.

A number of electronic components have already been presented as illustrational examples. However, we did not explain yet how to obtain PH-DAE and PH-ODE from circuit schematics. This topic is precisely the object of chapter 2 below.

Chapter 2

Revisiting circuit representations

Contents

2.1 Kirchhoff laws **45**
2.2 From circuits to graphs **46**
2.2.1 Elements of graph theory 47
2.3 Port-Hamiltonian representations of electronic circuits **53**
2.3.1 Kirchhoff-Dirac structure 53
2.3.2 Reduced Hybrid Dirac structure 54
2.3.3 From hybrid Dirac structures to semi-explicit pH-DAE 57
2.3.4 Reduction to Input-State-Output pH-ODE 60
2.3.5 Dissipative pH-DAE to conservative pH-ODE embedding 61
2.4 Bond Graphs and Wave Digital Filters **64**
2.4.1 Bondgraphs 64
2.4.2 Wave Digital Filters (WDF) 69
2.5 Port-variable changes **73**
2.5.1 Conversion to common and differential ports 73
2.5.2 Generalized linear port variables adapters 74

With numerical simulation in mind, we present the steps that are required to convert between circuit representations (see Figure 2.1): from the most general (netlists) to the most specific (PH-ODE and semi-explicit PH-DAE). For each formulation, we establish a systematic link with the underlying Dirac structures and the power balance. We quickly recall Kirchhoff laws and the lumped circuit hypothesis in section 2.1, elements of graph theory are recalled in section 2.2, PHS formulation of circuits are detailed in section 2.3, followed by a side by side comparison of bond-graphs and wave digital filters in section 2.4, finally we conclude by power-preserving port variable changes in section 2.5 that we use to preserve topological circuit symmetries (e.g. common and differential modes). Along the way, causal computations are addressed in subsection 2.3.3, PH-DAE to PH-ODE reduction in subsection 2.3.4. We also present in subsection 2.3.5 an alternative PH-DAE to (modulated) PH-ODE conversion such that the total energy (Hamiltonian+heat) is an explicit invariant (which can be exploited in numerical simulation).

state of the art For space reason, we focus on Port-Hamiltonian representations. We do not present classical circuit formulations that are already well covered in the literature, namely Modified Nodal Analysis (MNA) [HRB75] (the foundation of SPICE [Nag75]), Sparse Tableau Analysis (STA) [HBG71], Hybrid Analysis (HA) [CC76] and State Space formulation [KR65] (including the K/DK-methods [BDPR00, YAS10] and [HZ15]). The Brayton-Moser approach

[BM64a, BM64b] based on mixed-potential an co-energy variables is also skipped (we refer to [JS03] for their dual relation with PHS). In contrast, Wave Digital Filters (WDF) [Fet86], which are common in audio electronics, are shortly presented together with bond-graphs [Pay61] to highlight their similarities. Finally, we note that recent formulations of circuits as PHS [GHVdSR20, GBJR20] have been published during the redaction of this manuscript. The approach presented here is close to the first reference while the second one considers the PHS equivalent of charge-flux oriented MNA (which is not explored in this thesis).

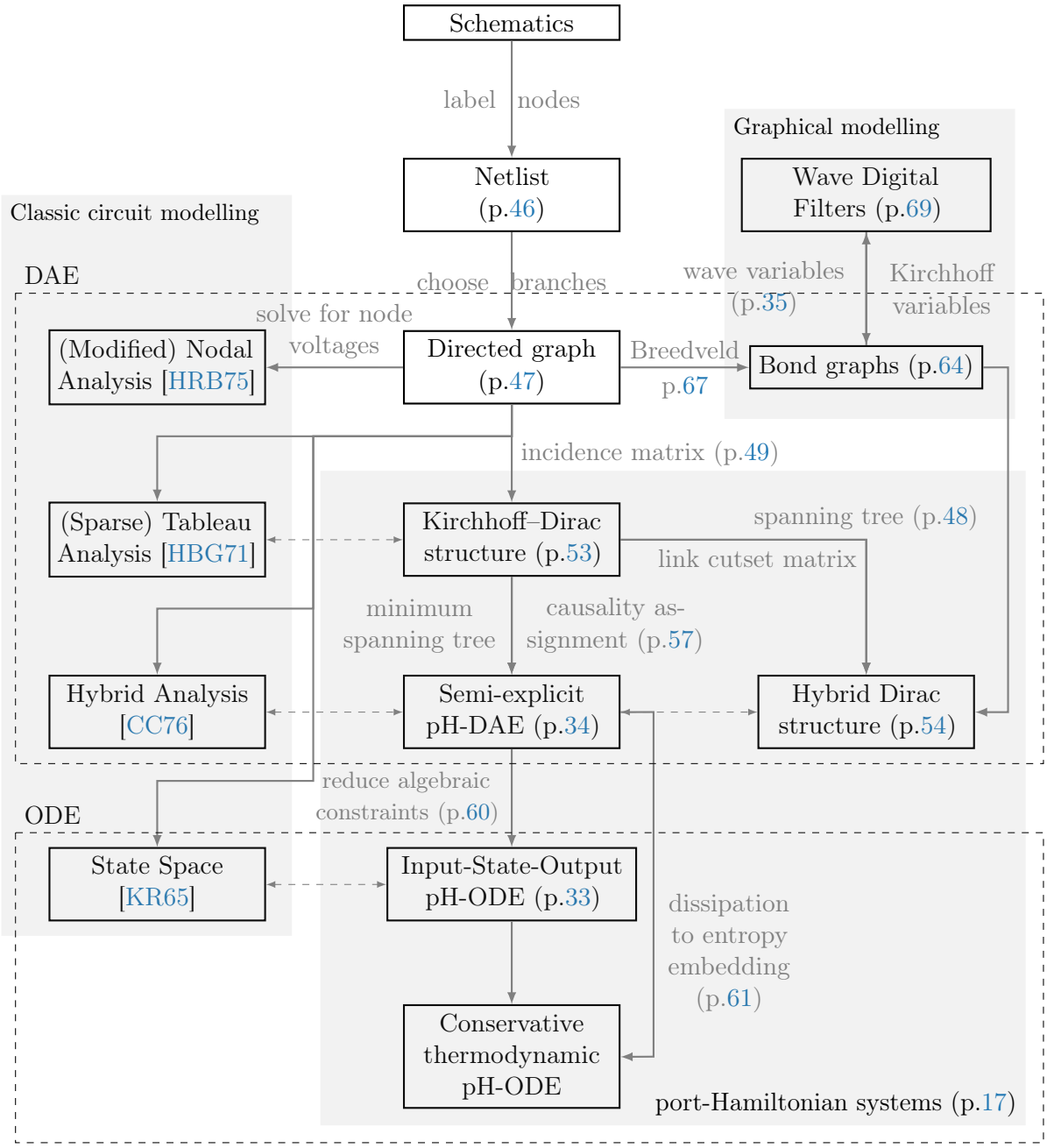


Figure 2.1 – Map of state of the art circuit modelling: representations, transformation diagram and relations with port-Hamiltonian formulations.

2.1 Kirchhoff laws

In this manuscript, we only consider *lumped circuits* in the context of audio applications with ideal conducting wires. To reduce a circuit to a lumped representation, for a given time scale, its physical dimension must be small enough so that the propagation speed of electromagnetic waves can be considered instantaneous [1].

Hypothesis 2.1 (Lumped circuit)**.** The lumped circuit hypothesis assumes that the circuit's characteristic length L_c is much smaller than the circuit's operating wavelength λ such that electro-magnetic steady-state is assumed, i.e.

- The change of the magnetic flux in time *outside a conductor* is zero. $\dfrac{\partial \phi_B}{\partial t} = 0$
- The change of the charge in time *inside conducting elements* is zero. $\dfrac{\partial q}{\partial t} = 0$

When this condition is satisfied, the current $i(t)$ through any branch, and the voltage $v(t)$ difference between any pair of nodes are well defined [FAC63]. The behaviour of the circuit becomes independent of the physical location of each component, only its topological interconnections becomes relevant [2]. Kirchhoff laws are a direct consequence of the lumped circuit hypothesis 2.1 and the assumption of ideal connections.

Kirchhoff Voltage Laws For any connected circuits with n nodes, since the electric potential is jauge-invariant, one can choose arbitrarily one *reference node* with respect to which one can measure $n-1$ node *voltages* $\{e_i\}_{i=1}^{n-1}$ and by definition $e_0 = 0$.

Definition 2.1 (Kirchhoff Voltage Laws (KVL))**.** The following are equivalent and defined for all lumped connected circuits, for all times, for all choices of reference node

- (closed node sequences) For all closed node sequences, the algebraic sum of all node-to-node voltages around the chosen closed node sequence is equal to zero.
- (Loop) The directed sum of the potential differences (voltages) around any closed loop is zero.
- (branch) For all pairs of nodes j, k, the branch voltage v_{kj} is equal to the difference of the node voltages $v_{kj}(t) = e_k(t) - e_j(t)$.

Kirchhoff Current Laws Kirchhoff Current Law (KCL) is an expression the electric charge conservation law. The fundamental concept to express KCL, is the notion of a *gaussian surface*.

Definition 2.2 (Gaussian surface)**.** A *gaussian surface* S is a two-sided closed surface in three-dimensional space enclosing a volume V through which the flux of a vector field is calculated. $S = \partial V$.

1. For audio circuits, the characteristic length L_c of a standard mounted rack is $L_c = 19'' \approx 48.26$ cm and the upper limit of the human auditory system is about $f = 20$ kHz. This corresponds to an electromagnetic wavelength $\lambda = c/f = 15$ km: that is *four orders of magnitude* higher than d. This justifies the lumping condition $L_c \ll \lambda$.
2. This is analog to the lumping of rigid-body mass-spring systems using point-masses.

Charge conservation, which was proved by Faraday in 1843, implies that the change in the amount of electric charge in any volume of space is exactly equal to the amount of charge flowing into the volume minus the amount of charge flowing out of the volume.

Definition 2.3 (Kirchhoff Current Laws (KCL))**.** Kirchhoff current laws, which are valid for all lumped circuits, for all times t, can be expressed equivalently

- (Gaussian surface law) The algebraic sum of the currents *entering* a Gaussian surface is equal to zero.
- (Node Law) The algebraic sum of the currents *entering* any node is equal to zero.
- (Cutset law) The algebraic sum of the currents associated with any cutset is equal to zero.

Proof. Les S be a gaussian surface enclosing a volume V, q the quantity of charges within the volume and $\boldsymbol{J}$ (A/m^2) the current density. By 1) definition of the current entering a gaussian surface, 2) the Stokes/divergence theorem, 3) charge conservation, 4) the lumped circuit hypothesis one obtains

$$I \overset{1}{=} -\oiint_{S=\partial V} \boldsymbol{J} \cdot \mathrm{d}\boldsymbol{S} \overset{2}{=} -\iiint_V (\nabla \cdot \boldsymbol{J})\,\mathrm{d}V \overset{3}{=} \frac{\partial q}{\partial t} \overset{4}{=} 0.$$

□

Remark 2.1. To every node corresponds a gaussian surface enclosing the node which cuts every edges connected to the node, and to every cutset corresponds a gaussian surface which cuts exactly the same branches.

A direct consequence of Kirchhoff laws is the power-balance of electronic circuits.

Theorem 2.1 (Tellegen theorem [Tel52])**.** *For all lumped circuits, for all times t, the sum of power over all circuit's branches is zero.*

2.2 From circuits to graphs

Any lumped circuit can be splitted into two independant parts: component laws which exist independently of the context in which components are used, and Kirchhoff Laws which are algebraic constraints on branch voltages and currents arising from the interconnection of components. Network topology deals with the properties of lumped networks solely determined by the interconnection of components. This modelling step is standard and common to all circuit modelling methods [Chu75, CDK87] (for PHS in audio circuits see [Fal16, FH16a]).

Netlist The standardized description of a circuit for electronic simulations is through a *netlist*. For our current purpose, it is enough to say that each line of a netlist stands for a *component* structured as follows

```
<type><label> <list of nodes> <parameters>; <comments>
```

For complete netlist specifications, please refer to SPICE documentation [Vla94].

Example 2.1. The netlist of a diode clipper circuit in figure 2.1 is given by

```
R1 1 2 1k     ; Resistor
D1 2 0 1N914  ; Shockley Diode
C1 2 0 1u     ; Linear capacitor
V1 1 0 1V     ; Voltage source
```

The knowledge of this netlist is then sufficient to one obtain the directed graph on the right

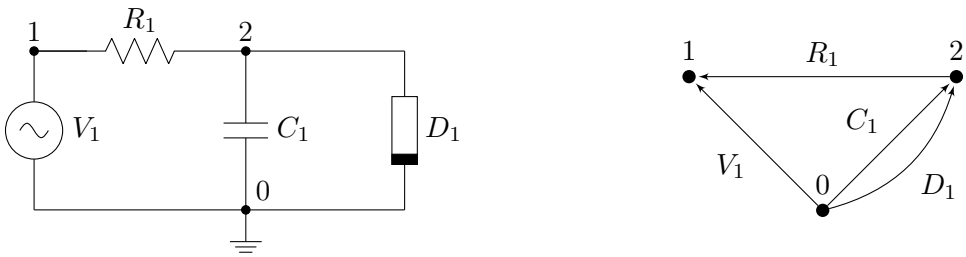


Figure 2.2 – Diode clipper graph.

2.2.1 Elements of graph theory

In order to automate the description and manipulation of Kirchhoff laws for any circuit, it is necessary to first recall some important results from graph theory that will be needed thereafter. We rely on references [Chu75, Deo17], and [Sle12, Sma00].

Definition 2.4 (Graph)**.** A graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ consists of two sets: a finite set of nodes (vertices) $\mathcal{N} = \{\eta_1, \ldots, \eta_n\}$ and a finite set of edges (branches, links) $\mathcal{E} = \{\epsilon_1, \ldots \epsilon_b\}$. Each edge is identified with a pair of vertices which can be ordered (directed graph) or non ordered (undirected graph)

Definition 2.5 (Path)**.** A set of edges $p = \{\epsilon_1, \ldots, \epsilon_n\}$ in a graph $\mathcal{G}$ is called a *path* between two nodes η_j, η_k if

1. consecutive branches ϵ_i, ϵ_{i+1} have a common node,
2. No node of $\mathcal{G}$ is contained in more that two edges of the set p,
3. η_j and η_k belong to exactly one edge in p.

Definition 2.6 (Connected Graph)**.** A graph $\mathcal{G}$ is said to be *connected* if there exists a path between any two nodes of the graph.

Definition 2.7 (Loop)**.** A subgraph $\mathcal{G}_s$ of a graph $\mathcal{G}$ is called a *loop* (or cycle) if

1. $\mathcal{G}_s$ is connected,
2. every node of $\mathcal{G}_s$ has exactly two edges of $\mathcal{G}_s$ incident at it.

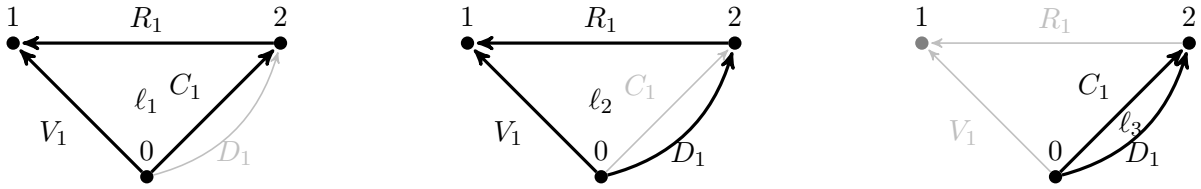


Figure 2.3 – Examples of loops shown in black.

Definition 2.8 (Tree)**.** A subgraph $\mathcal{G}_s$ of a connected graph $\mathcal{G}$ is called a *tree* if

1. $\mathcal{G}_s$ is connected,
2. $\mathcal{G}_s$ has no loop.

Definition 2.9 (Spanning Tree)**.** A subgraph $\mathcal{G}_s$ of a connected graph $\mathcal{G}$ is called a *spanning tree* if it is a tree that contains all nodes of $\mathcal{G}$.

Edges that belong to a spanning tree T are called *tree edges*, and those which do not belong to a spanning tree T are called *links*. All the links of a spanning tree T form a *cotree* $\overline{T}$ such that $T \cup \overline{T} \sim \mathcal{G}$. For a connected Graph $\mathcal{G}$ with n nodes, any spanning tree has exactly $n-1$ tree edges.

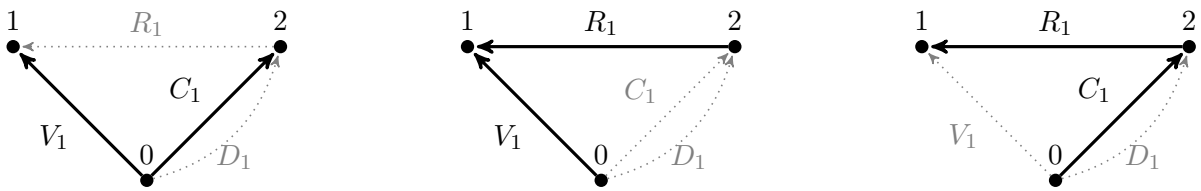


Figure 2.4 – Examples of spanning trees shown in black, with their cotree shown in dashed

Definition 2.10 (Cutset)**.** A set of edges C of a connected graph $\mathcal{G}$ is said to be a *cutset* if

1. The removal of edges C (not their nodes) results in a graph that is not connected,
2. after the removal of the edges, the restoration of any one edge from the set, will result in a connected graph.

To each cutset corresponds a partion of nodes $\mathcal{N}$ into two disjoint sets $(\mathcal{N}_1, \mathcal{N}_2)$ which can be oriented or non-oriented.

Definition 2.11 (Fundamental Loop and Cutset)**.** Let T be a spanning tree of a connected digraph $\mathcal{G}$ with cotree $\overline{T}$.

- For each branch $b \in \overline{T}$, the loop $L_b := \text{loop}(b \cup T)$ is said to be a *fundamental loop*
- For each branch $b \in T$, the cutset $C_b := b \cup \overline{T}$ is said to be a *fundamental cutset*.

These concepts are important to express Kirchhoff Laws in matrix form. In particular the notion of a (minimum) spanning tree, is required for automated generation of hybrid Dirac structures p.55 and in causality assignment p.57.

Incidence matrix

Definition 2.12 (Incidence Matrix)**.** For a directed graph $\mathcal{G}$ with n nodes and b branches, the *(node-edge) incidence matrix* of the graph is the $n \times b$ matrix defined by

$$\mathbf{A} := \left[a_{ij}\right]_{n\times b}, \quad a_{ij} = \begin{cases} 1 & \text{if branch } j \text{ enters node } i, \\ -1 & \text{if branch } j \text{ leaves node } i, \\ 0 & \text{otherwise} \end{cases} \tag{2.1}$$

Example 2.2 (Diode clipper incidence matrix)**.** For the circuit shown on Figure 2.1, the incidence matrix is

$$\mathbf{A} = \begin{array}{c} \\ \eta_0 \\ \eta_1 \\ \eta_2 \end{array} \begin{array}{c} \begin{array}{cccc} R_1 & D_1 & C_1 & V_1 \end{array} \\ \begin{bmatrix} 0 & -1 & -1 & -1 \\ 1 & 0 & 0 & 1 \\ -1 & 1 & 1 & 0 \end{bmatrix} \end{array}.$$

Definition 2.13 (Reduced Incidence matrix)**.** Any $(n-1) \times b$ submatrix $\mathbf{A}_f$ of an incidence matrix $\mathbf{A}(G)$ obtained by removing the row corresponding to a chosen reference node is called a *reduced incidence matrix.*

Example 2.3 (Diode clipper reduced incidence matrix)**.** Choosing node η_0 as reference node, one obtains the reduced incidence matrix

$$\mathbf{A}_f = \begin{array}{c} \\ \eta_1 \\ \eta_2 \end{array} \begin{array}{c} \begin{array}{cccc} R_1 & D_1 & C_1 & V_1 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 1 \\ -1 & 1 & 1 & 0 \end{bmatrix} \end{array}.$$

Definition 2.14 (Co-incidence Matrix)**.** For a directed graph $\mathcal{G}$ with n nodes and b branches, the co-incidence matrix of the graph is the $b \times n$ matrix defined by $\mathbf{D} = \mathbf{A}^\mathsf{T}$.

An important result to obtain a hybrid Dirac structure (p.55) from Kirchhoff laws is given in the following theorem and its corollary

Theorem 2.2 ([Deo17] thm 7.3)**.** *Let $\mathbf{A}$ be the incidence matrix of a connected graph $\mathcal{G}$ with n vertices. An $(n-1) \times (n-1)$ submatrix of $\mathbf{A}$ is non-singular if and only if the $n-1$ edges corresponding to the $n-1$ columns of this matrix constitutes a spanning tree in $\mathcal{G}$.*

Corollary 2.1. *For a spanning tree T, $\mathbf{A}$ can be partitionned into an $(n-1) \times (n-1)$ tree incidence matrix $\mathbf{A}_T$ and an $(n-1) \times (b-n+1)$ link incidence matrix $\mathbf{A}_L$ such that $\mathbf{A} = \begin{bmatrix} \mathbf{A}_T & \mathbf{A}_L \end{bmatrix}$, then $\mathbf{A}_T$ is invertible.*

Example 2.4. For the spanning tree $T = \{V_1, C_1\}$, with cotree $\overline{T} = L = \{R_1, D_1\}$

$$\mathbf{A}_f = \begin{array}{c} \\ \eta_1 \\ \eta_2 \end{array}\!\!\begin{array}{c} \begin{array}{cccc} V_1 & C_1 & R_1 & D_1 \end{array} \\ \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & -1 & 1 \end{bmatrix} \end{array}, \quad \mathbf{A}_T = \begin{array}{c} \\ \eta_1 \\ \eta_2 \end{array}\!\!\begin{array}{c} \begin{array}{cc} V_1 & C_1 \end{array} \\ \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \end{array}, \quad \mathbf{A}_L = \begin{array}{c} \\ \eta_1 \\ \eta_2 \end{array}\!\!\begin{array}{c} \begin{array}{cc} R_1 & D_1 \end{array} \\ \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix} \end{array}.$$

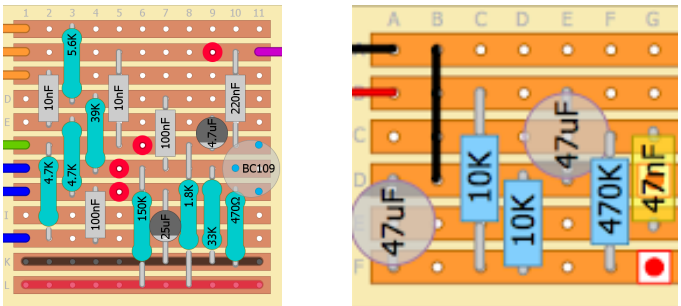


Figure 2.5 – Prototyping boards, a close physical analogy of a graph incidence matrix.

Loop matrix

Definition 2.15 (Loop incidence matrix)**.** For a directed graph $\mathcal{G}$ with ℓ oriented loops and b branches, the *loop incidence matrix* of the graph is the $\ell \times b$ matrix defined by

$$\mathbf{B} := \left[b_{ij}\right]_{\ell \times b}, \; b_{ij} = \begin{cases} 1 & \text{if branch } j \text{ is in loop } i \text{ with the same orientation,} \\ -1 & \text{if branch } j \text{ is in loop } i \text{ with the opposite orientation} \\ 0 & \text{otherwise.} \end{cases} \tag{2.2}$$

Example 2.5 (Diode clipper loop matrix)**.** For the Diode clipper circuit, one obtains the loop matrix

$$\mathbf{B} = \begin{array}{c} \\ \ell_1 \\ \ell_2 \\ \ell_3 \end{array}\!\!\begin{array}{c} \begin{array}{cccc} R_1 & C_1 & V_1 & D_1 \end{array} \\ \begin{bmatrix} 1 & 1 & -1 & 0 \\ 1 & 0 & -1 & 1 \\ 0 & 0 & -1 & 1 \end{bmatrix} \end{array}. \tag{2.3}$$

Theorem 2.3. *If $\mathcal{G}$ is a graph without self-loops, with incidence matrix $\mathbf{A}$ and loop matrix $\mathbf{B}$ whose columns are arranged using the same order of edges, then every row of $\mathbf{B}$ is orthogonal to every row of $\mathbf{A}$, that is $\mathbf{A}\mathbf{B}^\mathsf{T} = \mathbf{B}\mathbf{A}^\mathsf{T} = \mathbf{0}$.*

Definition 2.16 (Fundamental Loop matrix)**.** Any $b-n+1 \times b$ submatrix $\mathbf{B}_f$ of a loop matrix $\mathbf{B}$ in which all rows correspond to a set of fundamental loops (with respect to a spanning tree T) is called a *fundamental loop matrix*.

Property 2.1. A Fundamental loop matrix can be partitionned as $\mathbf{B}_f = [\mathbf{B}_T \ \mathbf{I}_L]$.

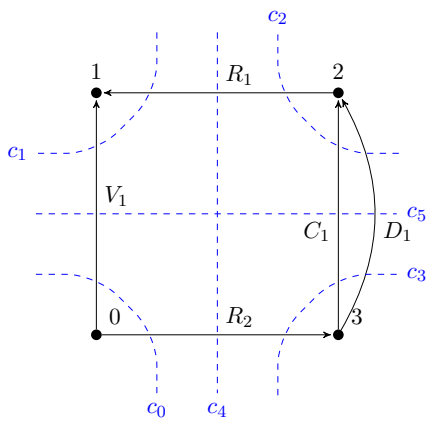
Example 2.6 (Diode clipper fundamental loop matrix)**.** The fundamental loop matrix for the tree $T = \{V_1, C_1\}$ with cotree $\overline{T} = \{R_1, D_1\}$ is obtained by removing the loop ℓ_2 (using the rule of only one cotree link per fundamental loop) and reordering columns into tree branches $\{C_1, V_1\}$ and cotree branches $\{R_1, D_1\}$

$$\mathbf{B}_f = \begin{array}{c} \\ \ell_1 \\ \ell_3 \end{array}\!\!\begin{array}{c} \begin{array}{cccc} C_1 & V_1 & R_1 & D_1 \end{array} \\ \begin{bmatrix} 1 & -1 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{bmatrix} \end{array}, \quad \mathbf{B}_T = \begin{array}{c} \\ \ell_1 \\ \ell_3 \end{array}\!\!\begin{array}{c} \begin{array}{cc} C_1 & V_1 \end{array} \\ \begin{bmatrix} 1 & -1 \\ 0 & -1 \end{bmatrix} \end{array}, \quad \mathbf{B}_L = \begin{array}{c} \\ \ell_1 \\ \ell_3 \end{array}\!\!\begin{array}{c} \begin{array}{cc} R_1 & D_1 \end{array} \\ \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \end{array}.$$

Cutsets matrix

Definition 2.17 (Cutset incidence matrix)**.** For a directed graph $\mathcal{G}$ with n_c oriented cutsets and n_b branches, the *cutset incidence matrix* of the graph is the $n_c \times n_b$ matrix defined by

$$\mathbf{C} := \left[c_{ij}\right]_{n_c \times n_b}, \ c_{ij} = \begin{cases} 1 & \text{if branch } j \text{ is in cutset } i \text{ with the same orientation,} \\ -1 & \text{if branch } j \text{ is in cutset } i \text{ with the opposite orientation,} \\ 0 & \text{otherwise.} \end{cases} \tag{2.4}$$



$$\mathbf{C} = \begin{array}{c} \\ c_0 \\ c_1 \\ c_2 \\ c_3 \\ c_4 \\ c_5 \end{array}\!\!\begin{array}{c} \begin{array}{ccccc} R_1 & R_2 & C_1 & V_1 & D_1 \end{array} \\ \begin{bmatrix} 0 & 1 & 0 & 1 & 0 \\ -1 & 0 & 0 & -1 & 0 \\ 1 & 0 & -1 & 0 & -1 \\ 0 & 1 & -1 & 0 & -1 \\ -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix} \end{array}$$

Figure 2.6 – Graph with cutsets and its cutset matrix

Theorem 2.4. *If $\mathcal{G}$ is a connected graph, then the rank of a cut-set matrix $\mathbf{C}(\mathcal{G})$ is equal to the rank of incidence matrix $\mathbf{A}(\mathcal{G})$, which equals the rank of graph $\mathcal{G}$.*

Definition 2.18 (Fundamental cutset matrix)**.** Let $\mathcal{G}$ be a connected graph with n nodes and b branches. The fundamental cut-set matrix $\mathbf{C}_f$ of $\mathcal{G}$ is an $(n-1) \times b$ submatrix of $\mathbf{C}$ such that the rows correspond to the set of fundamental cut-sets with respect to some spanning tree T.

Property 2.2. A Fundamental cutset matrix can be partitioned into a diagonal tree cutset matrix and a link cutset matrix as $\mathbf{C}_f = [\mathbf{I}_T \ \mathbf{C}_L]$.

Example 2.7. For a tree $T = \{R_1, R_2, V_1\}$, and its cotree $\overline{T} = \{C_1, D_1\}$, reordering columns, and removing cutsets c_0, c_1, c_4 corresponding to tree edges R_1, R_2, V_1 (i.e. $c_0 \cup R_1 = T$, $c_1 \cup R_2 = T$, $c_4 \cup V_1 = T$) one obtains the fundamental cutset matrix.

$$\mathbf{C}_f = \begin{array}{c} \\ c_2 \\ c_3 \\ c_5 \end{array} \begin{array}{c} \begin{array}{ccccc} R_1 & R_2 & V_1 & C_1 & D_1 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & -1 & -1 \\ 0 & 1 & 0 & -1 & -1 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix} \end{array}, \quad \mathbf{C}_T = \begin{array}{c} \\ c_2 \\ c_3 \\ c_5 \end{array} \begin{array}{c} \begin{array}{ccc} R_1 & R_2 & V_1 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \end{array}, \quad \mathbf{C}_L = \begin{array}{c} \\ c_2 \\ c_3 \\ c_5 \end{array} \begin{array}{c} \begin{array}{cc} C_1 & D_1 \end{array} \\ \begin{bmatrix} -1 & -1 \\ -1 & -1 \\ 1 & 1 \end{bmatrix} \end{array}.$$

Relation between A, B, C From theorem 2.3, partitioning incidence and loop matrices $\mathbf{A}, \mathbf{B}$ according to a tree T and dual links $L = \overline{T}$ as $\mathbf{A}_f = [\mathbf{A}_T \ \mathbf{A}_L]$, $\mathbf{B}_f = [\mathbf{B}_T \ \mathbf{I}_L]$ and using corrolary 2.1, one can show that the tree loop matrix $\mathbf{B}_T$ is related to the tree and link incidence matrices $\mathbf{A}_T$, $\mathbf{A}_L$ as follows $\mathbf{B}_T = -\mathbf{A}_T^{-1}\mathbf{A}_L$ [3].

Proof.

$$\mathbf{A}_f\mathbf{B}_f^{\mathsf{T}} = 0 \iff \begin{bmatrix} \mathbf{A}_T & \mathbf{A}_L \end{bmatrix} \begin{bmatrix} \mathbf{B}_T \\ \mathbf{I}_L \end{bmatrix} = 0 \iff \mathbf{A}_T\mathbf{B}_T + \mathbf{A}_L = 0 \iff \mathbf{B}_T = -\mathbf{A}_T^{-1}\mathbf{A}_L.$$

□

3. Note that tree loop matrix $\mathbf{B}_T$ and the link cutset matrices $\mathbf{C}_L$ are important objects that emerge when a Kirchhoff Dirac structure (see subsection 2.3.1) is reduced to an Hybrid Dirac structure.

2.3 Port-Hamiltonian representations of electronic circuits

We present here PH circuit representations and transformations that will be used in this thesis. The Kirchhoff–Dirac structure is presented in subsection 2.3.1, then its reduction as a Hybrid Dirac structure is shown in subsection 2.3.2. Transformation to semi-explicit pH-DAE using well chosen spanning trees is detailed in subsection 2.3.3. Finally reduction of pH-DAE to pH-ODE in detailed in subsection 2.3.4. An alternative refinement is presented in subsection 2.3.5 using thermodynamic embedding of pH-DAEs as conservative but irreversibly modulated pH-ODEs.

Voltage, current and bond spaces for circuits

Following [VdSM13, VdSM11] (see also [Sma00]), for a circuit graph $\mathcal{G}$ with n nodes and b branches, over each node (using the label $k = 0$) and branch (using the label $k = 1$) [4], *using the receiver convention for both*, we denote

- $\mathcal{V}_0 \sim \mathbb{R}^n$ the *node voltage space* and $\mathcal{I}_0 = \mathcal{V}_0^* \ (\sim \mathbb{R}^n)$ its dual the *node current space*,
- $\mathcal{V}_1 \sim \mathbb{R}^b$ the *branch voltage space* and $\mathcal{I}_1 = \mathcal{V}_1^* \ (\sim \mathbb{R}^b)$ its dual the *branch current space*,

with the duality pairings

$$\langle \mathbf{i}_k \,|\, \mathbf{v}_k \rangle_{\mathcal{B}_k} := \mathbf{i}_k \cdot \mathbf{v}_k, \quad \forall (\mathbf{i}_k, \mathbf{v}_k) \in \mathcal{I}_k \times \mathcal{V}_k, \quad k \in \{0, 1\}. \tag{2.5}$$

Together they generate respectively the *node bond space* $\mathcal{B}_0 = \mathcal{I}_0 \times \mathcal{V}_0$, the *branch bond space* $\mathcal{B}_1 = \mathcal{I}_1 \times \mathcal{V}_1$, and the *bond space* $\mathcal{B} = \mathcal{B}_0 \times \mathcal{B}_1$. respectively equipped with the quadratic forms (see (1.22) p.19)

$$Q_{\mathcal{B}_k}\left((\mathbf{i}, \mathbf{v})\right)_{\mathcal{B}_k} = 2 \langle \mathbf{i} \,|\, \mathbf{v} \rangle, \qquad \forall (\mathbf{i}, \mathbf{v}) \in \mathcal{B}_k, \quad k \in \{0, 1\}. \tag{2.6}$$

and their associated canonically defined indefinite bilinear form (see definition 1.12)

$$\left\langle (\mathbf{i}_1, \mathbf{v}_1), (\mathbf{i}_2, \mathbf{v}_2) \right\rangle_{\mathcal{B}_k} := \langle \mathbf{i}_1 \,|\, \mathbf{v}_2 \rangle_{\mathcal{B}_k} + \langle \mathbf{i}_2 \,|\, \mathbf{v}_1 \rangle_{\mathcal{B}_k}, \quad \forall (\mathbf{i}_1, \mathbf{v}_1), (\mathbf{i}_2, \mathbf{v}_2)) \in \mathcal{B}_k, \ k \in \{0, 1\}. \tag{2.7}$$

2.3.1 Kirchhoff-Dirac structure

Definition 2.19. Let $\mathbf{D} = \mathbf{A}^\mathsf{T}(\mathcal{G})$ be the reduced co-incidence matrix of a circuit graph $\mathcal{G}$. Kirchhoff Current and Voltage laws [a] can be expressed dually by

$$\mathbf{v}_1 = \mathbf{D}\mathbf{v}_0, \qquad \mathbf{i}_0 = -\mathbf{D}^\mathsf{T}\mathbf{i}_1 = 0. \tag{2.8}$$

This defines the following *Kirchhoff-Dirac structure*

$$\mathcal{D} = \left\{ (\mathbf{i}_0, \mathbf{v}_0, \mathbf{i}_1, \mathbf{v}_1) \in \mathcal{B}_0 \times \mathcal{B}_1 \;\middle|\; \begin{bmatrix} \mathbf{i}_0 \\ \mathbf{v}_1 \end{bmatrix} = \begin{bmatrix} \mathbf{0} & -\mathbf{D}^\mathsf{T} \\ \mathbf{D} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{v}_0 \\ \mathbf{i}_1 \end{bmatrix}, \quad \mathbf{i}_0 = 0. \right\} \tag{2.9}$$

a. The minus sign in front of $\mathbf{i}_0$ comes from the consistent use of the receiver convention for both nodes and branches: the sum of edge currents *entering* each node is zero.

4. This notation ($k = 0$, $k = 1$) is convenient and consistent with the k-junctions used in Bondgraph [Pay61]: **0**-junctions for nodes (shared voltage, parallel connection) and **1**-junctions for branches (shared current, serial connection). It is also a mnemonic to remember that lumped circuit equations arise from the spatial discretization of electro-magnetic 0-forms for nodes and 1-forms for branches.

Remark 2.2 (Interpretation)**.** Kirchhoff Current Laws can be interpreted as *zero boundary conditions* on the node currents [a]. The reduced co-incidence matrix $\mathbf{D}$ takes the status of a (lumped) differential operator $\mathbf{D} : \mathcal{V}_0 \to \mathcal{V}_1$, with adjoint the reduced incidence matrix $\mathbf{D}^\mathsf{T} : \mathcal{I}_1 \to \mathcal{I}_0$, i.e. we have the following diagram

$$\begin{tikzcd}
\mathbf{v}_0 \in \mathcal{V}_0 \arrow[r, "\mathbf{D}"] \arrow[d, leftrightarrow, "\langle\cdot|\cdot\rangle_{\mathcal{B}_0}"] & \mathbf{v}_1 \in \mathcal{V}_1 \arrow[d, leftrightarrow, "\langle\cdot|\cdot\rangle_{\mathcal{B}_1}"'] \\
\mathbf{i}_0 \in \mathcal{I}_0 & \mathbf{i}_1 \in \mathcal{I}_1 \arrow[l, "-\mathbf{D}^\mathsf{T}"]
\end{tikzcd} \tag{2.10}$$

a. If the charge is chosen as state variable for node and branches, this would correspoond to Neumann boundary conditions $\mathbf{i}_0 = \dot{\mathbf{q}}_0 = 0$.

Power Balance Left multiplying (2.9) by $[\mathbf{v}_0 \;\; \mathbf{i}_1]$, skew-symmetry of the Kirchhoff-Dirac structure leads to the power balance

$$\mathbf{v}_0 \cdot \mathbf{i}_0 + \mathbf{v}_1 \cdot \mathbf{i}_1 = \begin{bmatrix} \mathbf{v}_0 & \mathbf{i}_1 \end{bmatrix} \begin{bmatrix} \mathbf{0} & -\mathbf{D}^\mathsf{T} \\ \mathbf{D} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{v}_0 \\ \mathbf{i}_1 \end{bmatrix} = 0. \tag{2.11}$$

Tellegen theorem Furthermore since we have the KCL subconstraint $\mathbf{i}_0 = 0$ over the nodes, this yields Tellegen theorem (2.1) (the sum of a circuit branch power is zero) over the edges [5]

$$\mathbf{v}_1 \cdot \mathbf{i}_1 = 0. \tag{2.12}$$

Circuits and homology groups Using homology groups, one can interpret the Kirchhoff–Dirac structure as a realisation of a Stokes–Dirac structure [KML18] over 1-chains (edges) and 0-chains (nodes). See [VdSM11, VdSM13]. Kirchhoff laws can be rewritten canonically as $\delta\mathbf{i}_1 = 0$, and $\mathbf{v}_1 = \mathrm{d}\mathbf{v}_0$ where $\mathrm{d} \equiv \mathbf{D}$ denotes the *exterior derivative* and $\delta \equiv \mathbf{D}^\mathsf{T}$ denotes its dual the *co-differential*. See the thesis [Aba14, chap.3] for more details about algebraic topology and discrete Stokes relations (p.34) for electric circuits.

2.3.2 Reduced Hybrid Dirac structure

The dimensionality of the Kirchhoff-Dirac structure (2.9) can be reduced by eliminating node variables [6] which again yields a hybrid Dirac structure. Let T be a spanning tree (def. 2.9) of a circuit graph $\mathcal{G}$. Partitioning Kirchhoff laws (2.8) into tree (T) and link ($L = \overline{T}$) variables yields

$$\begin{bmatrix} \mathbf{v}_T \\ \mathbf{v}_L \end{bmatrix} = \begin{bmatrix} \mathbf{D}_T \\ \mathbf{D}_L \end{bmatrix} \mathbf{v}_0, \qquad \begin{bmatrix} \mathbf{D}_T^\mathsf{T} & \mathbf{D}_L^\mathsf{T} \end{bmatrix} \begin{bmatrix} \mathbf{i}_T \\ \mathbf{i}_L \end{bmatrix} = 0. \tag{2.13}$$

From theorem 2.2 and its corollary 2.1, having a spanning tree ensures that matrix $\mathbf{D}_T$ is invertible so that one can eliminate node voltages $\mathbf{v}_0$ using the relation

$$\mathbf{v}_0 = \mathbf{D}_T^{-1}\mathbf{v}_T. \tag{2.14}$$

5. Indeed [CDK87, p.30], any two of KCL, KVL and Tellegen theorem implies the third one.
6. This is the opposite of (Modified) Nodal Analysis [HRB75] which uses node voltages as main unknowns.

Fundamental loop and cutset form of Kirchhoff laws Substituting (2.14) in (2.13) and left multiplying the second equation of (2.13) by $\mathbf{D}_T^{-\mathsf{T}}$ yields the expression of Kirchhoff Voltage and Current Laws using *fundamental loop* and *fundamental cutset* matrices

$$\underbrace{\begin{bmatrix} -\mathbf{D}_L\mathbf{D}_T^{-1} & \mathbf{I}_L \end{bmatrix}}_{\text{fundamental loop matrix } \mathbf{B}_f} \begin{bmatrix} \mathbf{v}_T \\ \mathbf{v}_L \end{bmatrix} = 0, \qquad \underbrace{\begin{bmatrix} \mathbf{I}_T & \mathbf{D}_T^{-\mathsf{T}}\mathbf{D}_L^{\mathsf{T}} \end{bmatrix}}_{\text{fundamental cutset matrix } \mathbf{C}_f} \begin{bmatrix} \mathbf{i}_T \\ \mathbf{i}_L \end{bmatrix} = 0. \tag{2.15}$$

where the *tree loop matrix* $\mathbf{B}_T = -\mathbf{D}_L\mathbf{D}_T^{-1}$ and the *link cutset matrix* $\mathbf{C}_L = \mathbf{D}_T^{-\mathsf{T}}\mathbf{D}_L^{\mathsf{T}}$, are related by $\mathbf{C}_L = -\mathbf{B}_T^{\mathsf{T}}$. This is summarized by the following definition.

Definition 2.20 (Loop and cutset form of Kirchhoff Laws)**.** Let $\mathbf{B}_f$ and $\mathbf{C}_f$ be the fundamental loop and cutset matrices associated to a graph $\mathcal{G}$ with spanning tree T. then Kirchhoff laws can be written as

$$\mathbf{B}_f\mathbf{v} = 0, \quad \text{(KVL)} \qquad\qquad \mathbf{C}_f\mathbf{i} = 0. \quad \text{(KCL)} \tag{2.16}$$

where $\mathbf{B}_f = [\mathbf{B}_T\ \mathbf{I}_L]$ and $\mathbf{C}_f = [\mathbf{I}_T\ \mathbf{C}_L]$ and $\mathbf{C}_L = -\mathbf{B}_T^{\mathsf{T}}$.

Hybrid Dirac structure

Splitting voltages and currents according to tree and links in (2.15), one can express link voltages $\mathbf{v}_L$ in terms of tree voltages $\mathbf{v}_T$ and tree currents $\mathbf{i}_T$ in terms of link currents $\mathbf{i}_L$ as

$$\mathbf{v}_L = \mathbf{C}_L^{\mathsf{T}}\mathbf{v}_T, \qquad\qquad \mathbf{i}_T = -\mathbf{C}_L\mathbf{i}_L, \tag{2.17}$$

and gathering these informations yields the following definition.

Definition 2.21 (Hybrid Dirac structure)**.** Let $\mathbf{C}_L$ be the fundamental link cutset matrix associated to a graph $\mathcal{G}$ for a choice of spanning tree T, then the associated *Hybrid Dirac structure* is

$$\mathcal{D} = \left\{ (\mathbf{i}_T, \mathbf{v}_T, \mathbf{i}_L, \mathbf{v}_L) \in \mathcal{B}_T \times \mathcal{B}_L \;\middle|\; \begin{bmatrix} \mathbf{i}_T \\ \mathbf{v}_L \end{bmatrix} = \begin{bmatrix} \mathbf{0} & -\mathbf{C}_L \\ \mathbf{C}_L^{\mathsf{T}} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{v}_T \\ \mathbf{i}_L \end{bmatrix} \right\}. \tag{2.18}$$

i.e. we have the following diagram

$$\begin{tikzcd} \mathbf{v}_T \in \mathcal{V}_T \arrow[r, "\mathbf{C}_L^{\mathsf{T}}"] \arrow[d, leftrightarrow, "\langle\cdot\,|\,\cdot\rangle_{\mathcal{B}_T}"] & \mathbf{v}_L \in \mathcal{V}_L \arrow[d, leftrightarrow, "\langle\cdot\,|\,\cdot\rangle_{\mathcal{B}_L}"] \\ \mathbf{i}_T \in \mathcal{I}_T & \mathbf{i}_L \in \mathcal{I}_L \arrow[l, "-\mathbf{C}_L"] \end{tikzcd} \tag{2.19}$$

Example 2.8. In example 2.3, the reduced incidence matrix is

$$\mathbf{A}_f = \begin{array}{c} \\ \eta_1 \\ \eta_2 \end{array} \begin{array}{c} \begin{array}{cccc} R_1 & D_1 & C_1 & V_1 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 1 \\ -1 & 1 & 1 & 0 \end{bmatrix} \end{array} = \mathbf{D}^\mathsf{T}.$$

By consequence, according to (2.9), the corresponding Kirchhoff-Dirac structure is

$$\begin{bmatrix} i_1 \\ i_2 \\ v_{R_1} \\ v_{D_1} \\ v_{C_1} \\ v_{V_1} \end{bmatrix} = \left[\begin{array}{cc|cccc} . & . & -1 & 0 & 0 & -1 \\ . & . & 1 & -1 & -1 & 0 \\ \hline 1 & -1 & . & . & . & . \\ 0 & 1 & . & . & . & . \\ 0 & 1 & . & . & . & . \\ 1 & 0 & . & . & . & . \end{array}\right] \begin{bmatrix} v_1 \\ v_2 \\ i_{R_1} \\ i_{D_1} \\ i_{C_1} \\ i_{V_1} \end{bmatrix}, \qquad \begin{bmatrix} i_1 \\ i_2 \end{bmatrix} = \mathbf{0}.$$

Choosing a tree $T = \{V_1, C_1\}$ with cotree/links $L = \overline{T} = \{R_1, D_1\}$ yields the fundamental tree and link incidence matrices

$$\mathbf{A}_T = \begin{array}{c} \\ \eta_1 \\ \eta_2 \end{array} \begin{array}{c} \begin{array}{cc} V_1 & C_1 \end{array} \\ \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \end{array}, \qquad \mathbf{A}_L = \begin{array}{c} \\ \eta_1 \\ \eta_2 \end{array} \begin{array}{c} \begin{array}{cc} R_1 & D_1 \end{array} \\ \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix} \end{array}.$$

According to (2.15), we obtain the fundamental loop cutset matrix

$$\mathbf{C}_L = \mathbf{A}_T^{-1}\mathbf{A}_L = \begin{array}{c} \\ c_1 \\ c_2 \end{array} \begin{array}{c} \begin{array}{cc} R_1 & D_1 \end{array} \\ \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix} \end{array},$$

such that, according to (2.18), the Hybrid Dirac structure reduces to

$$\left[\begin{array}{c} i_{V_1} \\ i_{C_1} \\ \hline v_{R_1} \\ v_{D_1} \end{array}\right] = \left[\begin{array}{cc|cc} . & . & -1 & 0 \\ . & . & 1 & -1 \\ \hline 1 & -1 & . & . \\ 0 & 1 & . & . \end{array}\right] \left[\begin{array}{c} v_{V_1} \\ v_{C_1} \\ \hline i_{R_1} \\ i_{D_1} \end{array}\right].$$

Kernel form of Reduced Hybrid Dirac structure Using the fundamental loop and cutset matrices $\mathbf{B}_f$ and $\mathbf{C}_f$ from (2.16), one can obtain the kernel form of the reduced Dirac structure as follows. Define the matrices

$$\mathbf{E} = \begin{bmatrix} \mathbf{B}_T & \mathbf{I}_L \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \qquad \mathbf{F} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{I}_T & \mathbf{C}_L \end{bmatrix}, \tag{2.20}$$

then the kernel form of the reduced Dirac structure

$$\mathcal{D} = \left\{ (\mathbf{i}, \mathbf{v}) \in \mathcal{B}_1 \mid \mathbf{E}\mathbf{v} + \mathbf{F}\mathbf{i} = \mathbf{0} \right\}. \tag{2.21}$$

where one can verify that since $\mathbf{C}_L = -\mathbf{B}_T^\mathsf{T}$ it satisfies condition

$$\mathbf{E}\mathbf{F}^\mathsf{T} + \mathbf{F}\mathbf{E}^\mathsf{T} = \begin{bmatrix} \mathbf{0} & \mathbf{B}_T + \mathbf{C}_L^\mathsf{T} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{B}^\mathsf{T} + \mathbf{C}_L & \mathbf{0} \end{bmatrix} = \mathbf{0}. \tag{2.22}$$

Image form of Reduced Hybrid Dirac structure Finally, by transposition of the kernel Dirac structure (2.21) one obtains its dual image representation (which subsumes equation (2.17))

$$\mathcal{D} = \left\{ (\mathbf{i}, \mathbf{v}) \in \mathcal{B}_1 \;\middle|\; \mathbf{i} = \begin{bmatrix} \mathbf{B}_T^\mathsf{T} & \mathbf{0} \\ \mathbf{I}_L & \mathbf{0} \end{bmatrix} \boldsymbol{\lambda}, \quad \mathbf{v} = \begin{bmatrix} \mathbf{0} & \mathbf{I}_T \\ \mathbf{0} & \mathbf{C}_L^\mathsf{T} \end{bmatrix} \boldsymbol{\lambda}, \quad \forall \boldsymbol{\lambda} \in \mathbb{R}^b \right\}. \tag{2.23}$$

Note that, by inspection, the physical interpretation of the parameter $\boldsymbol{\lambda}$ corresponds to link currents $\mathbf{i}_L$ and tree voltages $\mathbf{v}_T$ as $\boldsymbol{\lambda} = \left[\begin{smallmatrix} \mathbf{i}_L \\ \mathbf{v}_T \end{smallmatrix}\right]$.

2.3.3 From hybrid Dirac structures to semi-explicit pH-DAE

The semi-explicit PHS representation from definition 1.24, is important for computer simulation. In particular, it fixes the choice of variables, it allows the formulation of a fixed-point equation, and it allows a structured interpretation of the power-balance.

$$\begin{bmatrix} \dot{\mathbf{x}} \\ \mathbf{w} \\ \mathbf{y} \end{bmatrix} = \mathbf{J} \begin{bmatrix} \nabla H(\mathbf{x}) \\ \boldsymbol{z}(\mathbf{w}) \\ \mathbf{u} \end{bmatrix}$$

In the context of a circuit, it is obtained by the following procedure: Let $\mathcal{G}$ be a circuit graph.

1. Denote $\mathbf{x}$ the state of *differential components* (capacitors and inductors) characterised by an energy potential $H(\mathbf{x})$, $\mathbf{w}$ the control variables of *passive algebraic components* characterized by a law $\boldsymbol{z}(\mathbf{w})$, and $\mathbf{u}$ the vector of *external* inputs with conjugated output variables $\mathbf{y}$.
2. Choose a spanning tree T of $\mathcal{G}$ such that current-controlled branches (voltage sources, capacitors, resistors, etc.) belong to the tree and all voltage-controlled branches (current sources, inductors, conductors, etc.) belong to the cotree $\overline{T}$.
3. Obtain the hybrid Dirac structure $\mathcal{D}$ of equation (2.18) and reorder rows and columns according to variables variables $(\mathbf{x}, \mathbf{w}, \mathbf{y})$ to obtain the skew-symmetric matrix $\mathbf{J}$.

Example 2.9. Reconsidering the diode clipper example, where $\mathbf{x} = q$, $\mathbf{w} = (v_R, v_D), \mathbf{y} = i_V, \mathbf{u} = v_V$, reordering the matrix and substituting component laws yields the pH-DAE

$$\begin{bmatrix} i_C = \dot{q} \\ v_R \\ v_D \\ i_V \end{bmatrix} = \left[\begin{array}{c|cc|c} . & 1 & -1 & . \\ \hline -1 & . & . & 1 \\ 1 & . & . & 0 \\ \hline . & -1 & 0 & . \end{array}\right] \begin{bmatrix} v_C = q/C \\ i_R = v_R/R \\ i_D = \mathrm{pn}(v_D) \\ v_V \end{bmatrix}$$

In step 2, it is not always possible to find a tree that satisfies these contraints. To address this problem, we propose the following approach

Finding a minimum spanning tree

We seek a spanning tree T that satisfies the following requirements

1. All Voltage sources and current controlled branches belong to the tree,
2. No current source and no voltage-controlled branch belong to the tree,
3. A maximum number of capacitors belong to the tree
4. A minimum number of inductors belong to the tree
5. Linear resistors and bijective algebraic components can belong to either tree or cotree.

This problem is similar to the Sequential Causality Assignment Procedure (SCAP) [VD95] in the bondgraph litterature (and its many variations [MFS00, WBK02]). This problem has also been addressed by Falaize with an ad-hoc algorithm in [Fal16, FH16a].

Zero-One-Linear integer programming problem in standard form For the b branches, let $\mathbf{x} \in \{0,1\}^b$ be the boolean vector representation of a subgraph T of a graph G (where $\mathbf{x}_\eta = 1$ if $\eta \in T$ and 0 otherwise). Its complement $\overline{T}$ is represented by the boolean vector $\bar{\mathbf{x}} = \mathbf{1} - \mathbf{x}$. A subgraph T is a tree of G (def. 2.8 p.48) if every node is reachable exactly once from the tree. This can be formalized using the graph Laplacian $\mathbf{L}(G) := \mathbf{A}^\mathsf{T}(G)\mathbf{A}(G)$ by the constraint

$$\mathbf{L}\mathbf{x} = \mathbf{1}. \tag{2.24}$$

where is the incidence matrix (see def. 2.12) of G. We formalize *preferred computational causalities*[7] constraints by the objective function

$$\Phi(\mathbf{x}) = \mathbf{w}_T \cdot \mathbf{x} + \mathbf{w}_{\overline{T}} \cdot \bar{\mathbf{x}} \tag{2.25}$$

with weights

$$\mathbf{w}_T(e) = \begin{cases} 1 & \text{if branch } e \text{ is current-controlled} \\ 0 & \text{otherwise} \end{cases}, \tag{2.26a}$$

$$\mathbf{w}_{\overline{T}}(e) = \begin{cases} 1 & \text{if branch } e \text{ is voltage-controlled} \\ 0 & \text{otherwise} \end{cases} \tag{2.26b}$$

Note that the objective function can be expressed with the number of branches b and a unique weighting function $\mathbf{w}$ as

$$\Phi(\mathbf{x}) = \mathbf{w}_T \cdot \mathbf{x} + \mathbf{w}_{\overline{T}} \cdot \bar{\mathbf{x}} = b + (\mathbf{w}_T - \mathbf{w}_{\overline{T}}) \cdot \mathbf{x}.$$

This leads to the Zero-One-Linear integer programming maximization problem in standard form

$$\begin{aligned} \text{maximize} \quad & b + \mathbf{w} \cdot \mathbf{x}, \\ \text{subject to} \quad & \mathbf{L}\mathbf{x} = \mathbf{1} \text{ and } \mathbf{x} \in \{0,1\}^b, \\ \text{with} \quad & \mathbf{w}(e) = \begin{cases} -1 & \text{if } e \text{ has voltage-controlled causality (I,L)} \\ 1 & \text{if } e \text{ has current-controlled causality (V,C,D,Q)} \\ 0 & \text{otherwise (R)} \end{cases}. \end{aligned} \tag{2.27}$$

7. arising either from numerical integration rules or from the availability of bijective algebraic maps.

Minimum spanning tree algorithm Since the cost function is restricted to spanning trees, and determined exclusively trough the tree vector $\mathbf{x}$, a significant simplification of the maximization problem (2.27) is to find a minimum spanning tree which solves the minimization problem

$$\begin{aligned} &\text{minimize} && -\mathbf{w}\cdot\mathbf{x}. \\ &\mathbf{x}\in\text{spanningtrees}(G) \end{aligned} \tag{2.28}$$

Note that this problem has an algorithmic complexity of $\mathcal{O}(b\log(n))$ when implemented using either the Prim–Dijkstra [Pri57] or the Kruskal [Kru56] algorithm.

A circuit with its minimum spanning tree is shown on figure 2.7. If a conforming spanning tree is found, then the number $\mathbf{w}\cdot\mathbf{x}$ should corresponds to the maximum number n_T of current-controlled edges in the circuit (here $n_T = 2$).

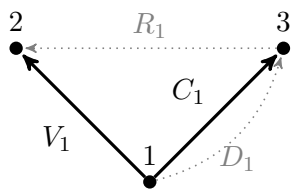


Figure 2.7 – Example of a minimum spanning tree that includes current-controlled branches.

Failure to satisfy the condition $\mathbf{w}\cdot\mathbf{x} = n_T$ can be used to detect the presence of topological problems such as hidden algebraic constraints (see figure 2.8).

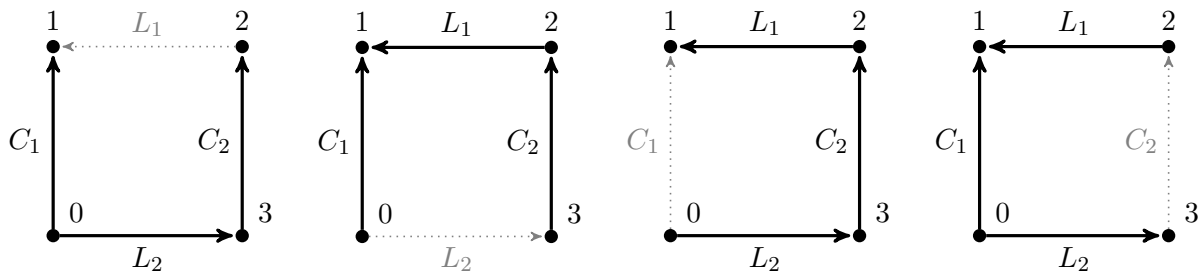


Figure 2.8 – Example of an LCLC circuit where there doesn't exists a spanning tree that includes all current-controlled branches and no voltage-controlled branches.

Note that, when a suitable minimum spanning tree cannot be found, so that the PH-DAE is semi-explicit, we proposed a fully-implicit numerical discretisation strategy in [MH20] which does not require causality assignment and can directly deal with such kind of implicit DAE constraints.

2.3.4 Reduction to Input-State-Output pH-ODE

In many cases, to study existence and uniqueness of solutions or to employ standard integration methods, it is desirable to reduce differential-algebraic equations to state-space ordinary differential equations. We show here how to transform a semi-explicit pH-DAE (definition 1.24) to an input-state-output pH-ODE (definition 1.23).

Consider a semi-explicit pH-DAE with Dirac structure $\mathcal{D}$ for a circuit graph $\mathcal{G}$ defined by the skew-symmetric matrix $\mathbf{S}$ partionned as follows

$$\begin{bmatrix} \dot{\mathbf{x}} \\ \mathbf{w} \\ \mathbf{y} \end{bmatrix} = \underbrace{\begin{bmatrix} \mathbf{J_x} & * & * \\ \mathbf{J_{xw}} & \mathbf{J_w} & * \\ \mathbf{J_{yx}} & \mathbf{J_{yw}} & \mathbf{J_y} \end{bmatrix}}_{\mathbf{S}} \begin{bmatrix} \nabla H(\mathbf{x}) \\ \boldsymbol{z}(\mathbf{w}) \\ \mathbf{u} \end{bmatrix}. \tag{2.29}$$

Case $\mathbf{J_w} = \mathbf{0}$ If $\mathbf{J_w} = \mathbf{0}$, which is a frequent case (no direct coupling between algebraic components), and there exists a symmetric positive definite matrix-valued function[8] $\boldsymbol{Z}(\mathbf{w})$ such that $\mathbf{z}(\mathbf{w}) = \boldsymbol{Z}(\mathbf{w})\mathbf{w}$, then one can reduce the dependance on $\mathbf{w}$ by reinjecting

$$\mathbf{w} = \mathbf{J_{xw}}\nabla H(\mathbf{x}) - \mathbf{J_{yw}^\top}\mathbf{u} \tag{2.30}$$

into (2.29) to obtain the nonlinear state-space system

$$\begin{bmatrix} \dot{\mathbf{x}} \\ \mathbf{y} \end{bmatrix} = \big(\mathbf{J} - \boldsymbol{R}(\mathbf{x},\mathbf{u})\big) \begin{bmatrix} \nabla H(\mathbf{x}) \\ \mathbf{u} \end{bmatrix} \tag{2.31}$$

where the skew-symmetrix matrix $\mathbf{J} = -\mathbf{J}^\top$ and the modulated symmetric positive definite matrix $\mathbf{R} = \mathbf{R}^\top \succeq 0$ are defined by

$$\mathbf{J} = \begin{bmatrix} \mathbf{J_x} & * \\ \mathbf{J_{yx}} & \mathbf{J_y} \end{bmatrix}, \qquad \boldsymbol{R}(\mathbf{x},\mathbf{u}) = \begin{bmatrix} \mathbf{J_{xw}^\top}\boldsymbol{Z}(\mathbf{x},\mathbf{u})\mathbf{J_{xw}} & -\mathbf{J_{xw}^\top}\boldsymbol{Z}(\mathbf{x},\mathbf{u})\mathbf{J_{yw}^\top} \\ -\mathbf{J_{yw}}\boldsymbol{Z}(\mathbf{x},\mathbf{u})\mathbf{J_{xw}} & \mathbf{J_{yw}}\boldsymbol{Z}(\mathbf{x},\mathbf{u})\mathbf{J_{yw}^\top} \end{bmatrix}. \tag{2.32}$$

and where by abuse of notation

$$\boldsymbol{Z}(\mathbf{x},\mathbf{u}) := \boldsymbol{Z}(\mathbf{w})\Big|_{\mathbf{w}=\mathbf{J_{xw}}\nabla H(\mathbf{x})-\mathbf{J_{yw}^\top}\mathbf{u}}. \tag{2.33}$$

Case $\mathbf{J_w} \neq \mathbf{0}$ When $\mathbf{J_w} \neq \mathbf{0}$, one needs to solve the implicit equation on $\mathbf{w}$

$$\mathbf{w} - \mathbf{J_w}\boldsymbol{z}(\mathbf{w}) = \mathbf{J_{xw}}\nabla H(\mathbf{x}) - \mathbf{J_{yw}^\top}\mathbf{u}. \tag{2.34}$$

Suppose the DAE is of index 1 such that the function $\boldsymbol{g}(\mathbf{w}) = \mathbf{w} - \mathbf{J_w}\boldsymbol{z}(\mathbf{w})$ can be inverted (algebraically or numerically) such that

$$\mathbf{w} = \boldsymbol{g}^{-1}(\mathbf{J_{xw}}\nabla H(\mathbf{x}) - \mathbf{J_{yw}^\top}\mathbf{u}).$$

then in general $\boldsymbol{z}(\mathbf{w})$ is no longer a separable function of $\nabla H(\mathbf{x})$ and $\mathbf{u}$. However if there exists matrix-valued functions $\boldsymbol{A}$, $\boldsymbol{B}$ such that

$$\boldsymbol{z}(\mathbf{w}) = \boldsymbol{A}(\mathbf{x},\mathbf{u})\nabla H(\mathbf{x}) + \boldsymbol{B}(\mathbf{x},\mathbf{u})\mathbf{u} \tag{2.35}$$

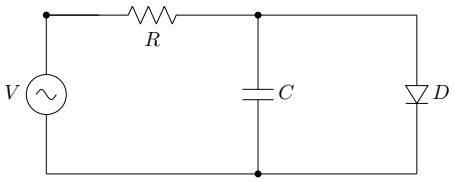
8. $\mathbf{Z}$ may not be positive definite if there exists conservative algebraic components, in which case $\mathbf{J}$ will be also modulated by $\mathbf{x}, \mathbf{u}$

then the algebraic feedback can be expressed by the matrix (not necessarily symmetric)

$$\boldsymbol{R}(\mathbf{x},\mathbf{u}) = \begin{bmatrix} \mathbf{J}_{\mathbf{xw}}^{\mathsf{T}}\boldsymbol{A}(\mathbf{x},\mathbf{u}) & -\mathbf{J}_{\mathbf{xw}}^{\mathsf{T}}\boldsymbol{B}(\mathbf{x},\mathbf{u}) \\ -\mathbf{J}_{\mathbf{yw}}\boldsymbol{A}(\mathbf{x},\mathbf{u}) & \mathbf{J}_{\mathbf{yw}}\boldsymbol{B}(\mathbf{x},\mathbf{u}) \end{bmatrix}. \tag{2.36}$$

Example 2.10. We reconsider the diode clipper example with semi-explicit representation ($\mathbf{x} = q$, $\mathbf{w} = (v_R, v_D), \mathbf{y} = i_V, \mathbf{u} = v_V$)

$$\begin{bmatrix} i_C = \dot{q} \\ v_R \\ v_D \\ i_V \end{bmatrix} = \left[\begin{array}{c|cc|c} . & 1 & -1 & . \\ \hline -1 & . & . & 1 \\ 1 & . & . & 0 \\ \hline . & -1 & 0 & . \end{array}\right] \begin{bmatrix} v_C = q/C \\ i_R = v_R/R \\ i_D = \mathrm{pn}(v_D) \\ v_V \end{bmatrix},$$



We remark that since current and voltage cannot be mixed in the Dirac structure, necessarily $\mathbf{J}_{\mathbf{x}} = 0$, $\mathbf{J}_{\mathbf{w}} = \mathbf{0}$, $\mathbf{J}_{\mathbf{y}} = 0$ and $\mathbf{J}_{\mathbf{yx}} = 0$. We have $v_R = v_S - q/C$ and $v_D = q/C$. It is then possible to solve explicitely for $\mathbf{w} = (v_R, v_D)$ to obtain the reduced state-space PHS with feed through (cf Equation (2.31) and Definition 1.23)

$$\begin{bmatrix} \dot{q} \\ y = i_V \end{bmatrix} = -\boldsymbol{R}(q) \begin{bmatrix} q/C \\ u = v_V \end{bmatrix}, \quad \boldsymbol{R}(q) = \frac{1}{R} \begin{bmatrix} 1 + R\,\mathrm{pn}(v_D)/v_D & -1 \\ -1 & 1 \end{bmatrix}_{v_D = q/C} \succeq 0.$$

where we already used in example 1.10 the fact that $\mathrm{pn}(v)/v \succeq 0$.

2.3.5 Dissipative pH-DAE to conservative pH-ODE embedding

We show how to embed a dissipative pH-DAE, as a conservative but irreversibly modulated input-state-output pH-ODE with extended thermodynamic state space. Our motivation is two-fold, first the DAE simplifies to a simpler state-space system, second the energy invariant of the system being explicit, we can use an energy-preserving ODE solver, rather than a DAE solver which only preserves a passive inequality. The transformation is similar to the approach presented in [EMVDS06] (see also the RS element [Bor09, p.52]). However, since we are not interested here in the thermodynamical details, it is simpler for our purpose to use the heat variable Q instead of the temperature T and entropy S.

We start from a semi-explicit skew-symmetric pH-DAE (2.29)

$$\begin{bmatrix} \dot{\mathbf{x}} \\ \mathbf{w} \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathbf{J}_{\mathbf{x}} & * & * \\ \mathbf{J}_{\mathbf{wx}} & \mathbf{J}_{\mathbf{w}} & * \\ \mathbf{J}_{\mathbf{yx}} & \mathbf{J}_{\mathbf{yw}} & \mathbf{J}_{\mathbf{y}} \end{bmatrix} \begin{bmatrix} \nabla H(\mathbf{x}) \\ \boldsymbol{z}(\mathbf{w}) \\ \mathbf{u} \end{bmatrix} \tag{2.37}$$

and *we consider dissipators as energy transducers converting electrical energy into heat.* To all resistive ports with power conjugate variables $(\mathbf{w}, \boldsymbol{z}(\mathbf{w}))$ and power $P(\mathbf{w}) = \boldsymbol{z}(\mathbf{w}) \cdot \mathbf{w}$, we associate a heat bath with thermodynamic potential

$$U(Q) = Q > 0 \tag{2.38}$$

with heat Q in Joule. It has energy variable Q with trivial co-energy variable $\nabla U = 1$.

Thermodynamic power balance Requiring that the dissipated power is absorbed by the thermodynamical potential U yields the thermodynamical power balance

$$\frac{\mathrm{d}}{\mathrm{d}t} U(Q) = \dot{Q} = \boldsymbol{z}(\mathbf{w}) \cdot \mathbf{w}. \tag{2.39}$$

Left multiplying the second row of (2.37) by $\boldsymbol{z}(\mathbf{w})^\mathsf{T}$ and factoring $\boldsymbol{z}(\mathbf{w})$ into the second column, yields the inhomogeneous ODE

$$\begin{bmatrix} \dot{\mathbf{x}} \\ \dot{Q} \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathbf{J_x} & * & * \\ \hat{\boldsymbol{J}}_{\mathbf{wx}}(\mathbf{w}) & \hat{\boldsymbol{J}}_{\mathbf{w}}(\mathbf{w}) & * \\ \mathbf{J_{yx}} & \hat{\boldsymbol{J}}_{\mathbf{yw}}(\mathbf{w}) & \mathbf{J_y} \end{bmatrix} \begin{bmatrix} \nabla H(\mathbf{x}) \\ \nabla U = 1 \\ \mathbf{u} \end{bmatrix} \tag{2.40}$$

with the following matrix-valued functions of the algebraic variable $\mathbf{w}$ defined by

$$\hat{\boldsymbol{J}}_{\mathbf{wx}}(\mathbf{w}) = \boldsymbol{z}(\mathbf{w})^\mathsf{T} \mathbf{J_{wx}}, \qquad \hat{\boldsymbol{J}}_{\mathbf{w}}(\mathbf{w}) = \boldsymbol{z}(\mathbf{w})^\mathsf{T} \mathbf{J_w} \boldsymbol{z}(\mathbf{w}), \qquad \hat{\boldsymbol{J}}_{\mathbf{yw}}(\mathbf{w}) = \mathbf{J_{yw}} \boldsymbol{z}(\mathbf{w}) \tag{2.41}$$

and where $\mathbf{w}$ is the solution [9] of $\mathbf{w} = \mathbf{J_{wx}} \nabla H(\mathbf{x}) + \mathbf{J_w} \boldsymbol{z}(\mathbf{w}) - \mathbf{J}_{\mathbf{yw}}^\mathsf{T} \mathbf{u}$.

Solving for w We introduce the function $\boldsymbol{g}(\mathbf{w}) = \mathbf{w} - \mathbf{J}_w \boldsymbol{z}(\mathbf{w})$. Under the hypothesis of the implicit function theorem (invertibility of the Jacobian $\boldsymbol{g}'$), we define the inverse function $\boldsymbol{w}$ to express the algebraic variable $\mathbf{w}$ as a function of state and input variables $\mathbf{x}, \mathbf{u}$

$$\boldsymbol{w}(\mathbf{x}, \mathbf{u}) = \boldsymbol{g}^{-1}(\mathbf{J_{wx}} \nabla H(\mathbf{x}) - \mathbf{J}_{\mathbf{yw}}^\mathsf{T} \mathbf{u}). \tag{2.42}$$

By substitution of (2.42) in (2.40) we define the modulated skew-symmetric matrix-valued function

$$\hat{\boldsymbol{J}}(\mathbf{x}, \mathbf{u}) := \begin{bmatrix} \mathbf{J_x} & * & * \\ \hat{\boldsymbol{J}}_{\mathbf{wx}}(\mathbf{x}, \mathbf{u}) & \hat{\boldsymbol{J}}_{\mathbf{w}}(\mathbf{x}, \mathbf{u}) & * \\ \mathbf{J_{yx}} & \hat{\boldsymbol{J}}_{\mathbf{yw}}(\mathbf{x}, \mathbf{u}) & \mathbf{J_y}. \end{bmatrix}, \qquad \hat{\boldsymbol{J}}^\mathsf{T} = -\hat{\boldsymbol{J}}. \tag{2.43a}$$

were the resulting matrix-valued functions of $\mathbf{x}$ and $\mathbf{u}$ are given by

$$\hat{\boldsymbol{J}}_{\mathbf{wx}}(\mathbf{x}, \mathbf{u}) := \boldsymbol{z}(\boldsymbol{w}(\mathbf{x}, \mathbf{u}))^\mathsf{T} \mathbf{J_{wx}}, \tag{2.43b}$$

$$\hat{\boldsymbol{J}}_{\mathbf{w}}(\mathbf{x}, \mathbf{u}) := \boldsymbol{z}(\boldsymbol{w}(\mathbf{x}, \mathbf{u}))^\mathsf{T} \mathbf{J_w} \, \boldsymbol{z}(\boldsymbol{w}(\mathbf{x}, \mathbf{u})), \tag{2.43c}$$

$$\hat{\boldsymbol{J}}_{\mathbf{yw}}(\mathbf{x}, \mathbf{u}) := \mathbf{J_{yw}} \, \boldsymbol{z}(\boldsymbol{w}(\mathbf{x}, \mathbf{u})) \tag{2.43d}$$

Thermodynamic pH-ODE Finally, introducing the total energy potential (Hamiltonian + Thermodynamical energy)

$$E(\mathbf{x}, Q) := H(\mathbf{x}) + U(Q) \tag{2.44}$$

and the extended state vector $\mathbf{X} = [\mathbf{x}, Q]^\mathsf{T}$, one obtains a conservative input-state-output pH-ODE (see def. 1.23) with *modulated matrix* $\hat{\boldsymbol{J}}$.

$$\begin{bmatrix} \dot{\mathbf{X}} \\ \mathbf{y} \end{bmatrix} = \hat{\boldsymbol{J}}(\mathbf{x}, \mathbf{u}) \begin{bmatrix} \nabla E(\mathbf{X}) \\ \mathbf{u} \end{bmatrix}. \tag{2.45}$$

9. Note that although the general case is implicit, it is frequent to have $\mathbf{J_w} = 0$

Example 2.11 (Conservative RLC)**.** Consider a Parallel RLC circuit with semi-explicit PHS representation

$$\begin{bmatrix} i_C = \dot{q} \\ v_L = \dot{\phi} \\ v_R \end{bmatrix} = \begin{bmatrix} 0 & -1 & -1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} v_C = q/C \\ i_L = \phi/L \\ i_R = v_R/R \end{bmatrix} \tag{2.46}$$

Using the thermodynamic embedding, we obtain the irreversibly modulated system with conserved total energy $E(q, \phi, Q) = q^2/2C + \phi^2/2L + Q$ (see figure 2.9)

$$\begin{bmatrix} \dot{q} \\ \dot{\phi} \\ \dot{Q} \end{bmatrix} = \begin{bmatrix} 0 & -1 & -q/(RC) \\ 1 & 0 & 0 \\ q/(RC) & 0 & 0 \end{bmatrix} \begin{bmatrix} q/C \\ \phi/L \\ 1 \end{bmatrix}. \tag{2.47}$$

Using the third row, and noticing that $v_R = q/C$, we recover the dissipative power transfer

$$\frac{\mathrm{d}}{\mathrm{d}t} U(Q) = 1 \cdot \dot{Q} = \frac{q}{C} \cdot \frac{q}{RC} = \frac{q}{C} \cdot \left(\frac{1}{R} \frac{q}{C} \right) = v_R \cdot i_R \geq 0. \tag{2.48}$$

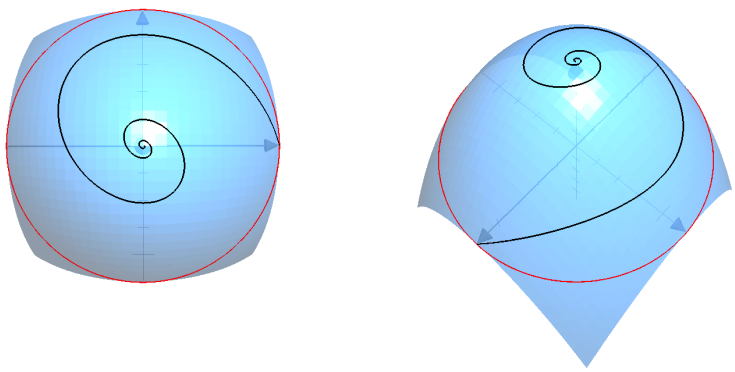


Figure 2.9 – Isothermal RLC. $x = q/\sqrt{C}$, $y = \phi/\sqrt{L}$, $z = (Q - Q_0)$. Iso-energy surface $\{(q, \phi, Q) \mid E(q, \phi, Q) = E(q_0, \phi_0, Q_0)\}$ (in blue). Reachable points are above the red circle.

Remark 2.3. It is possible to refine this representation in several ways.

a) use an isothermal heat bath $U(S) = TS$ with temperature T and entropy S,

b) keep track of the entropy variable for each component using the potential $U(S_1, \ldots, S_n) = T(S_1 + \ldots + S_n)$,

c) use distinct (and isolated) isothermal heat baths for each dissipative component $U(S_1, \ldots, S_n) = T_1 S_1 + \ldots + T_n S_n$,

d) replace the isothermal condition by heat diffusion.

2.4 Bond Graphs and Wave Digital Filters

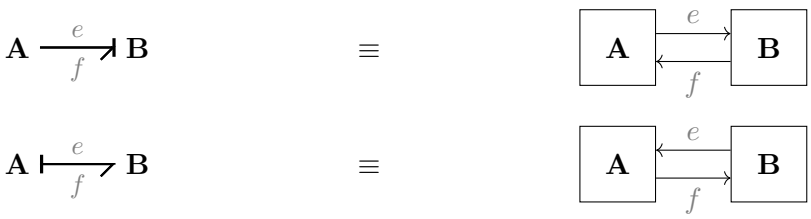
We leave equational algebraic representations to present two graphical network representations, namely bondgraphs and wave digital filters. Despite their notational differences, and the fact that bondgraphs use flow-effort variables while wave digital filters use wave variables, both notations are conceptually very similar and will be presented in parallel to highlight their similarities and differences. Both representations rely on breaking down a system into elementary n-port components, and connections between them.

We shortly present below the basics of both formalisms, for more details, please refer to the following references for bond graphs [Pay61, Bre86, Bro99b, GVdSBM03, Bor09] and [Fet86, Bil04, WNSA15, WBSS18, BS17] for WDF.

2.4.1 Bondgraphs

Bondgraphs are a multi-physics network modelling tool invented by Henri Paynter at the MIT in 1959. It models energy transfer as an oriented graph between subsystems A, B such that power $e \cdot f$ is positive in the direction of the half-arrow.

Note that the equivalent block diagram on the right is not oriented yet. To realize the block-diagram, it is necessary to assign a so-called *computational causality* which is indicated by a vertical bar toward the element that is *effort-driven* the other element being *flow-driven*.



Serial and parallel junctions As we have already seen, systems are connected together through power-preserving junctions structures. The basic building blocks to create more elaborated connections are the serial **1** and parallel junctions **0**

$f_1 = \ldots = f_n$ $\qquad$ $e_1 = \ldots = e_n$

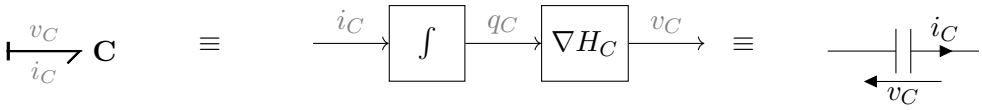
$e_1 + \ldots + e_n = 0$ $\qquad$ $f_1 + \ldots + f_n = 0$

We remark that, for parallel junctions, since all efforts are equal only one port can be effort-driven. Dually for serial junctions all flows being equal, only one port can be flow-driven.

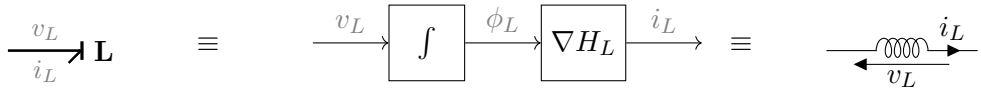
Transformer and Gyrators Two important Dirac structures, the Transformer and Gyrator are represented (with their admissible causalities) by

Common electronic components

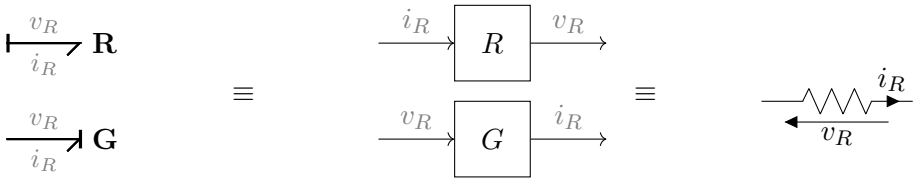
- Capacitor: the law of a (nonlinear) capacitor is $v_C(t) = \nabla H_C \left(q_C = \int_\infty^t i_C(\tau)\,\mathrm{d}\tau\right)$. This is formalized by the current-driven component.



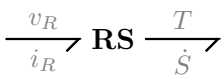
- Inductor: the law of a (nonlinear) inductor is $i_L(t) = \nabla H_L \left(\phi_L = \int_\infty^t v_L(\tau)\,\mathrm{d}\tau\right)$. This is formalized by the voltage-driven component.



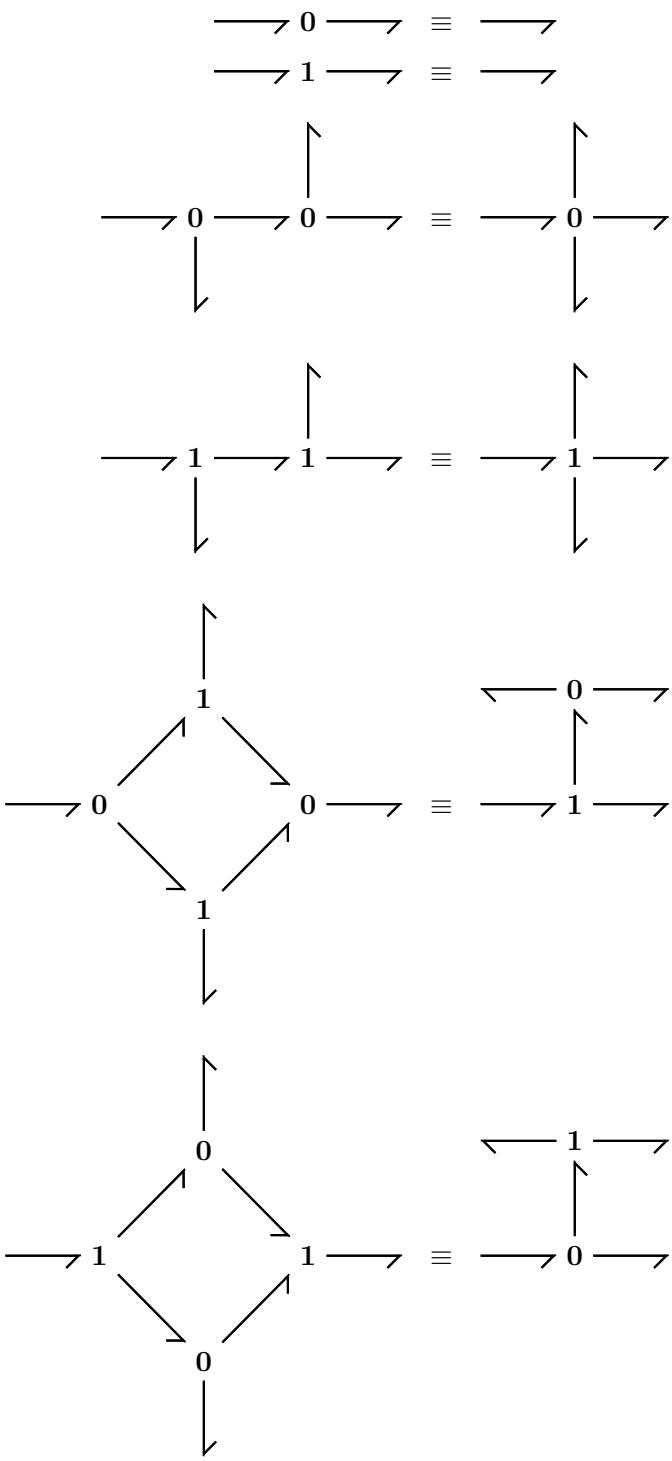
- Resistor / Conductor: (nonlinear) resistors (conductors) are charaterized by passive relations $R : i_R \mapsto v_R$, $(G : v_R \mapsto i_R)$



- RS element [Bor09, p.52]: In the bondgraph litterature, dissipators can also be considered as energy transducers converting non-thermal energy into heat satisfying the power balance $\dot{Q} = T\dot{S} = v_R \cdot i_R$.



Simplification rules We recall here some useful graphical bondgraph simplification rules (see [Bro99b, Bor09]). These can considerably reduce the number of elements and save tedious algebraic manipulations.

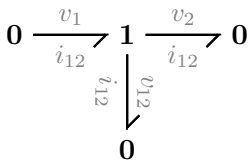


(2.49a)
(2.49b)
(2.49c)
(2.49d)
(2.49e)
(2.49f)

In particular, these rules are implemented in the 20-sim software [Bro99a]. We also note that since these identities only rely on (here Kirchhoff) conservation laws, they translate directly to Wave Digital Filters.

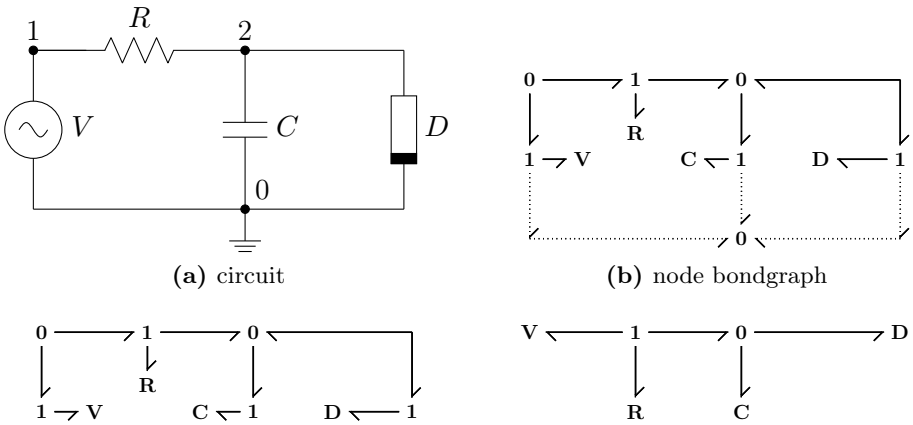
Automated conversion of circuits to Bondgraphs In reference [Bre86], Breedveld proposed an procedure to automatically convert a circuit to bond graphs. This systematic procedure is of great value in particular when working with pen and paper to avoid errors. It is summarized (here for electronic circuits) by the following steps

1. For each node η_i of the circuit create a parallel $\mathbf{0}_i$ junction (the node voltage v_i is shared at the $\mathbf{0}$ junction),
2. For each branch between two nodes η_i, η_j form the voltage difference $v_{ij} = v_i - v_j$ represented by a zero junction $\mathbf{0}_{ij}$ connected to a serial $\mathbf{1}_{ij}$ junction as follows [10]



3. Connect all ports of all components to the corresponding branch voltages,
4. Suppress the ground node and all its bonds,
5. (optional) use bond graph simplification rules

A step by step application of the method to the diode clipper test circuit is shown below on Figure 2.10.



(a) circuit **(b)** node bondgraph **(c)** node bondgraph mass removed **(d)** reduced bondgraph using simplification rules

Figure 2.10 – Automated Bondgraph modelling of the diode clipper circuit.

We note that we can layout the graph in a canonical way, in order to exhibit the fact that the junction structure of the unreduced bond graph is bipartite (i.e. a **1**-junction is necessarily connected to a **0**-junction) see Figure 2.11.

Causality assignment procedures As we have seen, to make a bondgraph computable, it is necessary to orient its equivalent block-diagram such that each port is either flow or effort driven. However in practice, some components such as voltage and current sources or non bijective dissipators have an imposed causality, dynamic components such as capacitors and inductors have a preferred integral causality while bijective algebraic components have no preferred causality. In

10. Mnemonic: n**0**de, v**0**ltage → **0**-junction, ser**1**al, **1**ntensity → **1**-junction.

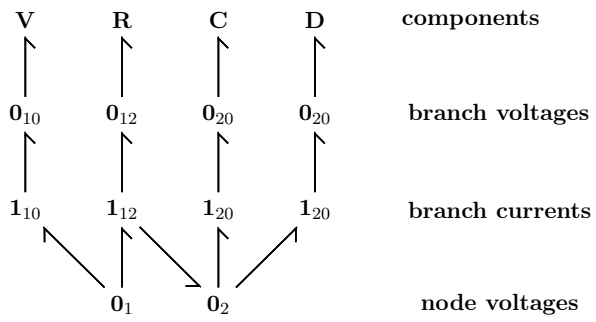


Figure 2.11 – Bi-partite bondgraph of the diode clipper circuit.

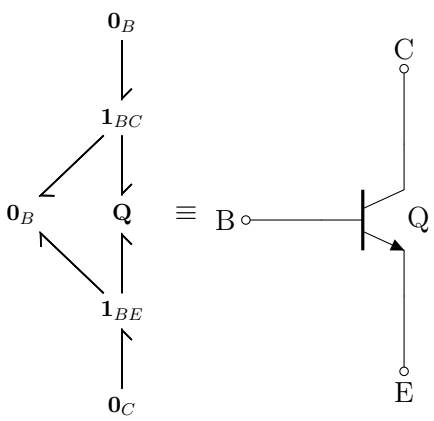
the Bondgraph litterature this problem is called the Sequential Causality Assignment Procedure (SCAP) [KR68] for which many variants have been proposed (see reference [MFS02] for a review). It can be summarized by the following steps

1. Assign causalities for all components that have fixed causalities
2. Propagate causalities through $\mathbf{0}, \mathbf{1}$ junctions, ideal transformers and gyrators
3. Repeat steps 1 and 2 with components having preferred causalities
4. While there remains unoriented bonds choose an orientation for one and propagate causalities
5. (optional) If causality conflicts are detected, backtrack choices made in step 3 and 4 and resume the procedure.

This problem is closely related to the problem presented in subsection 2.3.3 where we show how to formulate and efficiently solve causality assignment as a minimum spanning tree problem. In practice however, the procedure described above remains important to perform causality assignment graphically using only pen and paper and no computer.

Occurence of step 4 is an indicator of the presence of algebraic loops in the bond graph.

n-port and m-terminal elements Finally, to illustrate how to deal with elements that are represented either as n-ports or m-terminals, we show the bondgraph of a 2-port, 3-terminal: the Bipolar Junction Transistor.

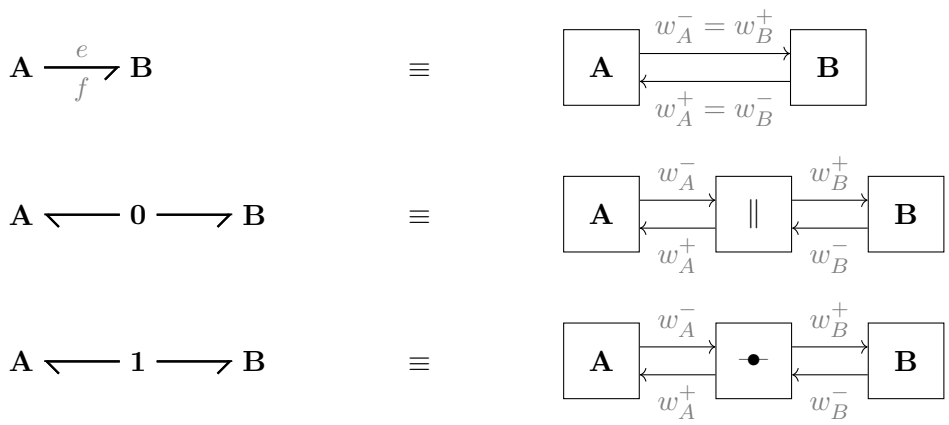


2.4.2 Wave Digital Filters (WDF)

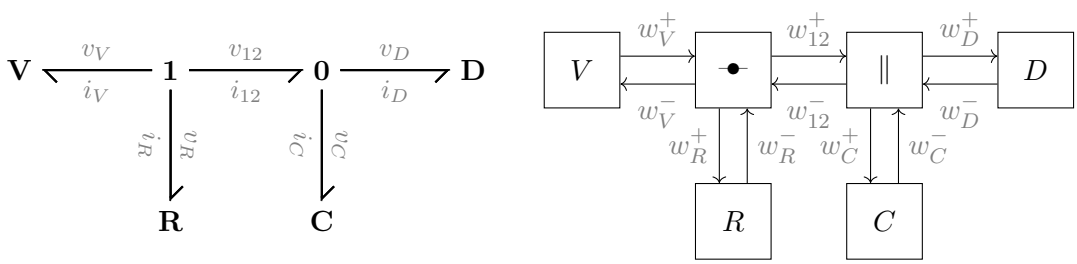
Now that the bondgraph formalism has been presented, we show similarities and differences between bond graphs and WDF. We rely on references [Fet86, Bil04], see also [FOO05, Wer16] for more recent developments (in particular SPQR trees). Compared to bondgraphs, the WDF formalism has some important differences:

1. wave variables (w^+, w^-) are used instead of flow-effort variables (f, e),
2. there is no need to assign computational causalities: block diagram inputs are incident wave w^+ and outputs are reflected waves w^-.
3. the variable change is done *after discretization*,
4. WDFs rely on adapting the port-impedance parameter R of the wave variable change to achieve reflection-free ports or break delay-free loops (i.e. obtain causal delayed reflected waves [11]).

The last property is perhaps the strongest computational advantage of WDF compared to standard methods. In term of graphical representations, we have the following equivalences



Continuing with the diode clipper example from figure 2.10, we obtain the equivalence between bondgraph and WDF shown on figure 2.12.

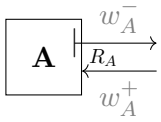


(a) reduced bondgraph **(b)** WDF

Figure 2.12 – Equivalence between circuit Bondgraph and WDF represention.

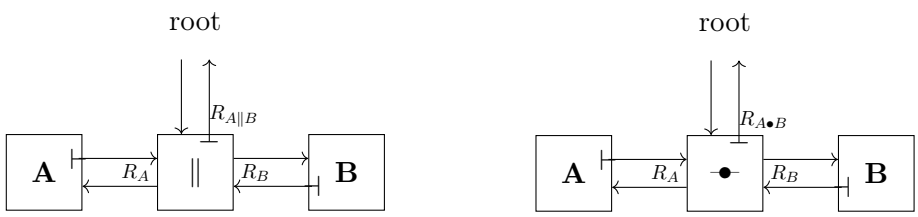
11. At time t_n, the reflected wave w_n^- does not depend on the incident wave w_n^+.

Port-Adaptation, Binary and SPQR connection trees In WDF, the port impedance can be chosen such that the reflected wave does not depend instantaneously on the incidence wave. This property (no instantaneous algebraic loop) is shown graphically by a vertical bar where the port is adapted.



$\equiv$ $w_A^-[n]$ does not depend on $w_A^+[n]$

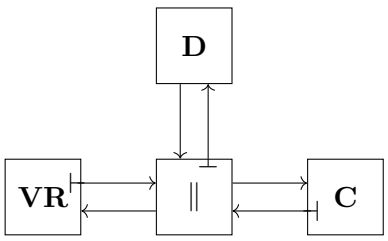
Similarily to the fact that for parallel (resp. serial) junctions, only one port can be effort-driven (resp. flow-driven), in the WDFs, only one port (called the root) can be adapted while the remaining ports (called the leaves) inherit their port-impedance from the connected components.



Serial/parallel Binary Connection Trees (BCT) Using this property, for many circuits, (by decomposing serial and parallel junction into 3 port adapters) it is possible to arrange elements into a serial-parallel binary connection tree.

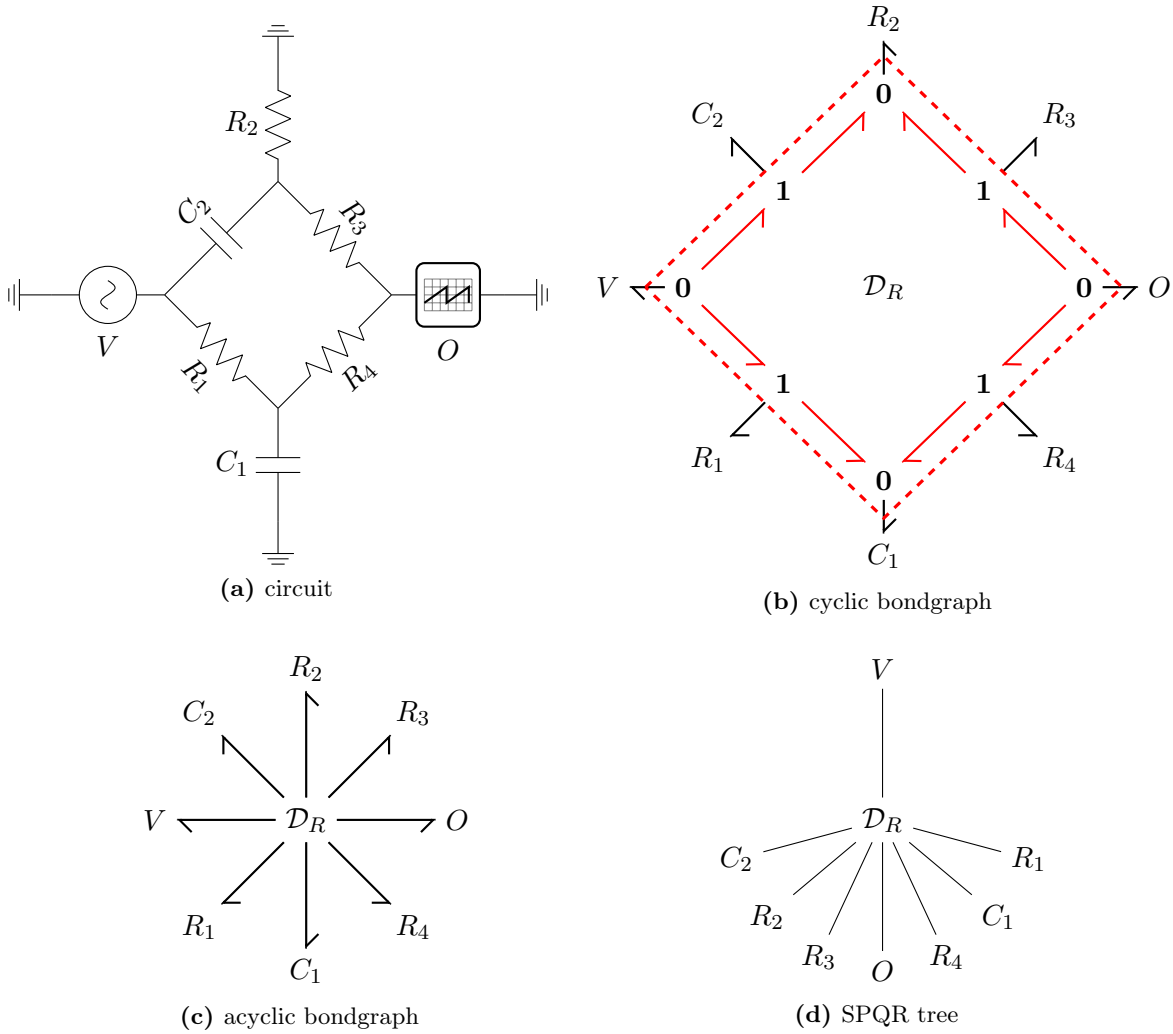
To numerically process the WDF tree at each time step, first reflected waves (which do not depend instantaneously on incident waves) are propagated from the leaves to the root. Then incident waves are propagated from the root to the leaves to update the state of stateful elements.

Using this approach it is possible to have a single nonlinear element at the root and use Newton iteration to solve the instantaneous algebraic loop. This is illustrated below: the diode clipper from Figure 2.12 has been redrawn with the nonlinear element **D** at the root of the tree, and the voltage source and resistor have been lumped into a resistive voltage source **VR** with port impedance R).



SPQR trees However the above approach fails for multiple nonlinearities or complex network topologies which stimulated research for alternative strategies [FOO05, WNSA15, Wer16, WBSS18]. An approach is to collect all nonlinear elements into a single multi-port situated at the root of the tree and to decompose the remaining elements into an SPQR tree [12] [DBT96]. The example in figure 2.13 illustrates that rigid nodes arise as soon as the bondgraph contains algebraic loops. To address this difficulty, these loops (red lines in figure 2.13b) are aggregated into irreducible Dirac structure multiports to obtain an acyclic bondgraph (in figure 2.13c). Then choosing a root (V in fig.2.13d), the graph can be transformed into an SPQR tree.

12. S for *serial* nodes, P for *parallel* nodes, R for *rigid* (strongly connected) nodes and Q for trivial nodes.



(a) circuit **(b)** cyclic bondgraph

(c) acyclic bondgraph **(d)** SPQR tree

Figure 2.13 – Example of a circuit containing a rigid node $\mathcal{D}_R$ transformed to a single-root SPQR tree (taken from the tone stack stage of the Big Muff π distortion pedal). I would like to thank Kurt Werner and Ólafur Bogason for the fruitful discussion on this topology at DAFx18 in Aveiro.

2.5 Port-variable changes

This section introduces the class of port variables changes that are linear, power-conserving, and that operate *across* ports. This class, different than that of wave variable changes (performing port by port linear combination of flow-effort pairs) is of interest to exploit circuit symmetries.

2.5.1 Conversion to common and differential ports

A common source of symmetry in physics happens when a system only depends on the difference between port variables. In electronics, differential amplifiers [13] (as the name suggests) are exactly designed for that purpose. However in practice, components are not perfect and are often characterised by their common mode rejection ratio, so that both common and differential ports are necessary. Furthermore it is often the case that topological symmetries in the circuit are broken by computational causality assignment. By consequence, in practice, the following theorem is useful for devices whose description is simpler in terms of common and differential ports. This is used in section 7.2.3 p.194 (see also the symmetries on circuits, fig. 7.24 p.195).

Theorem 2.5 (Common-differential 2-port)**.** *Consider a 2-port with conjugated port variables* $(f_1, e_1) \in \mathcal{F}_1 \times \mathcal{E}_1$, $(f_2, e_2) \in \mathcal{F}_2 \times \mathcal{E}_2$, *and the variable change* $(f_1, f_2, e_1, e_2) \leftrightarrow (f_\Delta, f_\Sigma, e_\Delta, e_\Sigma)$

$$f_\Delta = \alpha(f_1 - f_2), \qquad e_\Delta = \beta(e_1 - e_2), \tag{2.50a}$$
$$f_\Sigma = \alpha(f_1 + f_2), \qquad e_\Sigma = \beta(e_1 + e_2). \tag{2.50b}$$

where $\alpha\beta = 1/2$. *Then,* (2.50a)-(2.50b) *defines an equivalent* common-differential 2-port *parametrisation with the same power*

$$\langle f_\Sigma \,|\, e_\Sigma \rangle + \langle f_\Delta \,|\, e_\Delta \rangle = \langle f_1 \,|\, e_1 \rangle + \langle f_2 \,|\, e_2 \rangle \,. \tag{2.51}$$

Proof. Substituting (2.50a) (2.50b) into (2.51) and eliminating cross terms yields $\langle f_\Sigma \,|\, e_\Sigma \rangle + \langle f_\Delta \,|\, e_\Delta \rangle = \frac{1}{2}\left[\langle f_1 + f_2 \,|\, e_1 + e_2 \rangle + \langle f_1 - f_2 \,|\, e_1 - e_2 \rangle\right] = \langle f_1 \,|\, e_1 \rangle + \langle f_2 \,|\, e_2 \rangle \,.$ □

Example 2.12 (Amplifiers)**.** Consider a 4-port amplifier (here with lumped energy source) having input-output ports $\{I+, I-, O+, O-\}$, differential gain $K_\Delta \gg 1$ and common mode gain K_Σ. Its representation is the Σ-Δ domain by the diagonal matrix

$$\begin{bmatrix} e_O^\Delta \\ e_O^\Sigma \end{bmatrix} = \begin{bmatrix} K_\Delta & 0 \\ 0 & K_\Sigma \end{bmatrix} \begin{bmatrix} e_I^\Delta \\ e_I^\Sigma \end{bmatrix}, \qquad \begin{bmatrix} f_I^\Delta \\ f_I^\Sigma \end{bmatrix} = \mathbf{0}. \tag{2.52}$$

is more natural than in the original domain by

$$\begin{bmatrix} e_{O+} \\ e_{O-} \end{bmatrix} = \begin{bmatrix} K+\epsilon & -K+\epsilon \\ -K+\epsilon & K+\epsilon \end{bmatrix} \begin{bmatrix} e_{I+} \\ e_{I-} \end{bmatrix}, \qquad \begin{bmatrix} f_{I+} \\ f_{I-} \end{bmatrix} = \mathbf{0}. \tag{2.53}$$

A passive model of the operational amplifier is detailed in chapter 7.

13. Differential amplifiers are commonly used in guitar and microphone preamps, operational amplifiers or in the Moog synthesizer filter

2.5.2 Generalized linear port variables adapters

We generalize the previous variable change by interpreting it as a power-conserving Dirac structure adapter between multiports. This is illustrated by the block-diagram of figure 2.15.

Theorem 2.6. *Let $\mathcal{D}$ be linear multi-port adapter mapping vector port variables $(\mathbf{f}_a, \mathbf{e}_a) \in \mathcal{F}_a \times \mathcal{E}_a$ to vector port variables $(\mathbf{f}_b, \mathbf{e}_b) \in \mathcal{F}_b \times \mathcal{E}_b$ where $\mathcal{F}_a \sim \mathbb{R}^n$, $\mathcal{F}_b \sim \mathbb{R}^n$ according to*

$$\mathbf{f}_b = \mathbf{F}\mathbf{f}_a, \qquad \mathbf{e}_b = \mathbf{E}\mathbf{e}_a, \qquad \mathbf{F}^\mathsf{T}\mathbf{E} = -\mathbf{I}_n, \tag{2.54}$$

with full rank matrices $\mathbf{F}, \mathbf{E} \in \mathbb{R}^{n\times n}$. Then $\mathcal{D}$ defines a Dirac structure.

Proof. According to proposition 1.1, $\mathcal{D}$ is a Dirac structure if and only if $\langle \mathbf{f} \,|\, \mathbf{e} \rangle = 0$ and $\dim \mathcal{D} = \dim \mathcal{F}_a \times \mathcal{F}_b$. Indeed substituting (2.54) into the power-balance yields

$$\langle \mathbf{f} \,|\, \mathbf{e} \rangle = \langle \mathbf{f}_a \,|\, \mathbf{e}_a \rangle + \langle \mathbf{f}_b \,|\, \mathbf{e}_b \rangle = \mathbf{f}_a^\mathsf{T}\mathbf{e}_a + \mathbf{f}_a^\mathsf{T}\mathbf{F}^\mathsf{T}\mathbf{E}\mathbf{e}_a = \mathbf{f}_a^\mathsf{T}\mathbf{e}_a - \mathbf{f}_a^\mathsf{T}\mathbf{e}_a = 0.$$

And we have $\dim \mathcal{D} = \operatorname{rank}(\mathbf{F}) + \operatorname{rank}(\mathbf{E}) = 2n = \dim \mathcal{F}_a \times \mathcal{F}_b$. □

Lemma 2.1. *Let $\mathbf{F}$ be any unitary orthogonal transform and $\mathbf{E} = -\mathbf{F}$. Then this is a sufficient condition to have $\mathbf{F}^\mathsf{T}\mathbf{E} = -\mathbf{I}$, satisfying equation (2.54).*

Example 2.13 (Common-differential adapter)**.** The common-differential variable change from theorem 2.5 can be formalized as a common-differential adapater defined by

$$\begin{bmatrix} f_\Delta \\ f_\Sigma \end{bmatrix} = -\alpha \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \end{bmatrix}, \qquad \begin{bmatrix} e_\Delta \\ e_\Sigma \end{bmatrix} = \beta \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \end{bmatrix}. \tag{2.55}$$

Note the change of sign compared to theorem 2.5, so that the adapter uses the receiver convention. It is illustrated on figure 2.14.

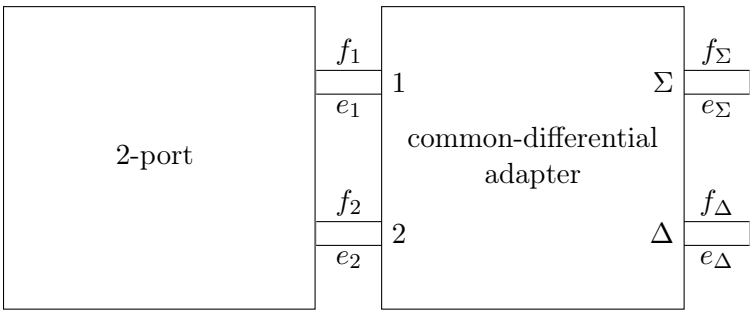


Figure 2.14 – Illustration of common-differential adaptation of a 2-port.

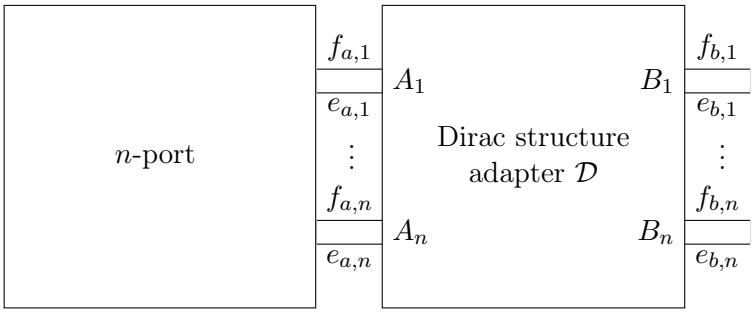


Figure 2.15 – Generalized n-port adapter.

Example 2.14 (Orthogonal adapters)**.** According to lemma 2.1, the common-differential adapter (2.55) is an instance of the more general class of unitary two-port adapters (for $\theta = \pi/4$, $\alpha = \beta = 1/\sqrt{2}$)

$$\begin{bmatrix} f_\Delta \\ f_\Sigma \end{bmatrix} = - \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \end{bmatrix}, \qquad \begin{bmatrix} e_\Delta \\ e_\Sigma \end{bmatrix} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \end{bmatrix}. \tag{2.56}$$

More generally, orthogonal n-port adapters (2.54) can diagonalise a coupled multi-dimensional relation (e.g. $\mathbf{e} = \mathbf{R}\mathbf{f}$ where $\mathbf{R} = \mathbf{R}^\mathsf{T} \succeq 0$ has an SVD decomposition $\mathbf{R} = \mathbf{U}\mathbf{S}\mathbf{U}^\mathsf{T}$).

Example 2.15 (common differential representation of a 2-port parallel jonction)**.** consider a parallel jonction defined by

$$e_1 = e_2, \qquad f_1 + f_2 = 0.$$

Then its common-differential representation becomes the trivial constraints

$$e_\Delta = 0, \qquad f_\Sigma = 0. \tag{2.57}$$

such that we have $P = e_\Sigma \cdot f_\Sigma + e_\Delta \cdot f_\Delta = 0$.

Example 2.16 (common differential representation of a 3-port parallel jonction)**.** consider a classical parallel junction defined by

$$e_1 = e_2 = e_3, \qquad f_1 + f_2 + f_3 = 0.$$

If we choose to transform ports $\{1, 2\}$ to common-differential $\{\Sigma, \Delta\}$ using (2.50b) with $\alpha = 1/2$, $\beta = 1$, we obtain the following singular skew-symmetric Dirac structure

$$\begin{bmatrix} e_\Delta \\ e_\Sigma \\ f_3 \end{bmatrix} = \begin{bmatrix} . & . & 0 \\ . & . & 1 \\ 0 & -1 & . \end{bmatrix} \begin{bmatrix} f_\Delta \\ f_\Sigma \\ e_3 \end{bmatrix}. \tag{2.58}$$

We can see that the differential port Δ has no influence on the behaviour of the circuit.

Conclusion

In this chapter, we have recalled the main results from network and circuit theory, we have seen how to obtain a PHS from a circuit graph and how to transform it to semi-explicit PH-DAE and PH-ODE. We have reviewed the topic of "computational causality assignment". Causality assignment is important for numerical reasons: in practice, it is usually preferable to obtain equations that are numerically integrated (integral causality) rather than numerically differentiated (differential causality). A strength of the PH framework is that under a week hypothesis (invertibility of the Jacobian of algebraic nonlinearities, see 1.2.2 p.14) many circuits are representable as (semi-explicit) index-1 DAE (and thus convertible to ODE). This property is important to study existence and uniqueness of solutions. To highlight their similarities and differences, we have presented two graphical network formalisms side by side: Wave Digital Filters and bond-graphs. Finally we have presented "accross ports" power-conserving variable changes. They can explicitly exploit network or component symmetries. In particular Σ-Δ variable changes can be used to avoid breaking symmetries of differential or push-pull circuits during computational causality assignment. The modelling framework being setup, we are ready to address the broad subject of power-balanced numerical simulation methods. This is the object of Part II which constitutes the core of this thesis.

Part II

Time-continuous power-balanced numerical methods

Approach and problem statement

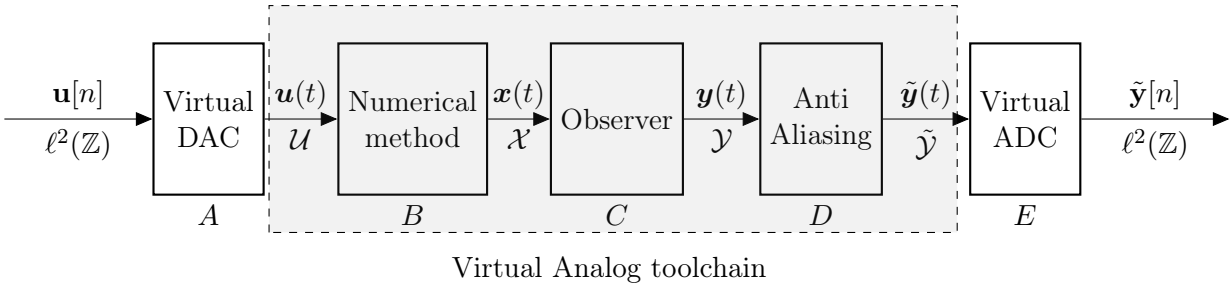


Figure 2.16 – (continuous-time virtual analog signal processing) block-diagram of the approach

Signal processing framework

In this part, we propose power-balanced numerical methods (block B) within a complete signal processing chain (A-E) described in figure 2.16. It is based on the following.

- *Reconstruction* (block A): A (bandlimited) sampled input $\mathbf{u}[n]$ is reconstructed by a virtual Digital to Analog Converter (Virtual DAC) to obtain a continuous-time signal $\boldsymbol{u}(t)$ represented over sequences of time frames.

- *Numerical Solver* (blocks B-C): for each time frame, given an input signal $\boldsymbol{u}(t)$ represented by parameters $\widehat{\mathbf{u}}$, a power-balanced numerical method produces an output signal $\boldsymbol{y}(t)$ with parameters $\widehat{\mathbf{y}}$,

- *Sampler* (Blocks D-E): the signal $y(t)$ is meant to be *listened* through a soundcard. For that purpose, a virtual antialiasing filter and sampler (Virtual ADC) are used to obtain the discrete-time signal $\tilde{\mathbf{y}}[n]$ based on bandlimited Shannon-Nyquist sampling (see thm. 3.1).

To precise our approach, continuous-time signal representations, generalized sampling theory and the implementation of virtual DAC, anti-aliasing and virtual ADC are discussed first in chapter 3. Subsequent chapters are dedicated to power-balanced numerical methods.

Power-balanced Numerical methods

This thesis is dedicated to build numerical methods to solve PH-ODE and PH-DAE whose numerical solutions are required to satisfy the following properties

P0. **Class of solutions** Numerical solutions are approximated in the time-continuous domain and represented with a finite number of parameters per time-frame.

P1. **Regularity** Numerical solutions inherit the global regularity of true solutions up to a controllable regularity order denoted k. Indeed, for a function $f(t)$ of class $\mathcal{C}^k$, its Fourier-spectrum $\left|\widehat{F}(\omega)\right|^2$ decreases asymptotically as $1/\omega^{2(k+1)}$. This property is important to reduce the requirements on the antialiasing module.

P2. **Accuracy** For each time frame, the approximation error between numerical solutions $\boldsymbol{X}(\tau)$ and true solutions $\boldsymbol{x}(\tau)$ is controllable, bounded and converges to zero for small time steps h, with a controllable accuracy order p (defined thereafter).

P3. **Power-balance** Numerical approximations satisfy the PH power-balance over each time-frame. In particular, for conservative PHS the Hamiltonian $H(\mathbf{x})$ must be exactly preserved from frame to frame, and for dissipative PHS, the Hamiltonian must decrease monotonically over time (in the absence of external input).

While the interplay between continuous and discrete time is a common theme in (digital) signal processing and control theory, in numerical analysis, many numerical methods (e.g. Finite Differences, Runge–Kutta, multistep) are discrete by design [14]: the underlying continuous-time signal model is not made explicit. We note some important exceptions which are relevant to us: Runge–Kutta methods with dense output [HNW93, II.6], Continuous Runge–Kutta Methods [OZ92], Time finite elements (TFEM) [Hul92, BB93, Bot97, BS00], time-continuous Galerkin (CG), time-discontinuous Galerkin (DG) [TS12, TSC17] and continuous-stage Runge–Kutta (CSRK) methods [Hai10, MB16, Tan18]. Continuous Galerkin and CSRK formulations will be considered in chapter 5 p.117.

Outline

Chapter 3 details the general continuous-time signal processing framework used to implement blocks A-E. We first review important results and notations about functional analysis, non-bandlimited signals and (generalized) sampling theory that are required thereafter. Then we review several realisation strategies and tradeoffs for the Virtual DAC (block A) and Virtual ADC modules (blocks D-E in fig. 2.16). Subsequent chapters 4-6 propose different methods for the realisation of blocks B-C.

Chapter 4 is of an introductory nature. Satisfaction of properties $\mathbf{P1} - \mathbf{P3}$ is considered using *adaptive collocation* for PH-ODEs. (Symmetric) Power balanced Adaptive collocation methods ((S)PAC) are introduced. Their analysis reveals that, using this approach, the existence domain of power-balanced solutions is bounded.

Chapter 5 proposed a more general framework. It relies on an alternative viewpoint: using the idea of *continuous-time functional projection*. We introduce the notion of a functional Dirac structure [15] over a time frame, for which a sufficient condition to preserve the power balance is established. Then, Regular Power balanced projection Methods (RPM) are introduced, with controllable projection and regularity orders. They are analysed and illustrated for both Port-Hamiltonian ODEs and DAEs.

Chapter 6 extends the ideas of chapter 5 and combines them with *exponential integrators* (which exactly solve the linear dynamic). First the exponential Average Vector Field (EAVF) method is introduced and shown to be energy-preserving (resp. dissipating) for autonomous systems. Then, input–output ports are considered. Finally, an extension strategy towards higher orders is proposed.

14. However, backward error analysis [HLW06] allows to interpret these schemes as sampled solutions of modified continuous-time approximation of the original system.
15. see definition 1.14 p.20.

Chapter 3

Non-bandlimited signal representations, reconstruction and antialiasing

Think analog, act digital

Michael Unser [Uns05]

Contents

3.1 Generalized-sampling theory and Finite Rate of Innovation 83
3.1.1 Short reminder on functional analysis 83
3.1.2 Class of signals and notations 85
3.1.3 Sampling signals with a Finite Rate of Innovation (FRI) 86
3.1.4 Piecewise polynomial frames 88
3.2 Input reconstruction (Virtual DAC) 89
3.2.1 B-spline spaces 89
3.2.2 Shifted linear interpolation 91
3.3 Output antialiasing and sampling (Virtual ADC) 93
3.3.1 Exact continuous-time filtering for LTI state-space systems 93
3.3.2 Approximation of (broken) piecewise polynomials on B-spline spaces 98
3.4 Application: "virtual analog" resampler 104

Before we address numerical methods (blocks B-C in fig.3.2), we detail the *virtual analog* (VA) continuous-time signal processing framework that will be used in the following of this manuscript and propose realisation strategies for blocks A,D,E.

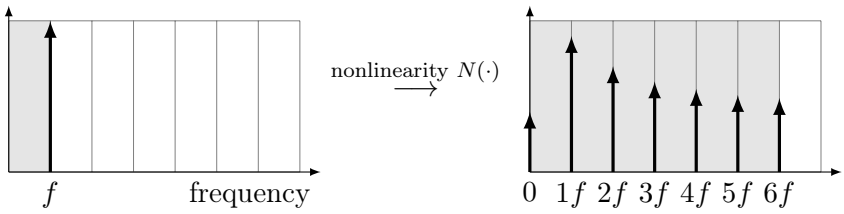
Non band-limited signals with a finite rate of innovation We observe the following facts:

a) Signals arising from nonlinear physical systems are usually *not bandlimited*. Furthermore the outputs of nonlinear systems usually have a *richer spectral content* than their inputs because of the bandwidth expansion of nonlinearities (see figure 3.1a),

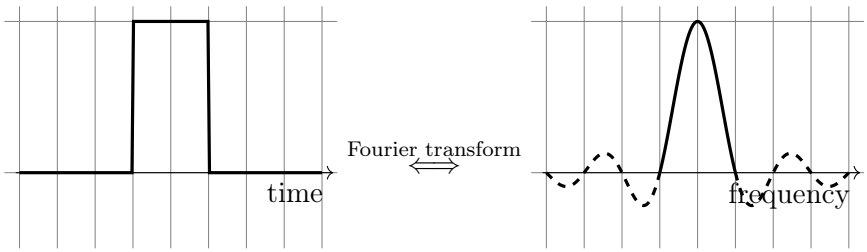
b) Real-time numerical time integration methods rely on *causal time-stepping*, meaning that any decomposition on basis functions must have finite and non-overlapping temporal support from frame to frame (see figure 3.1b)

c) Because of memory requirements, computer representations of continuous-time signals signals are necessarily *finite-dimensional*.

Because of (a), in this thesis, we rely on *generalized sampling theory* [Uns00, NH14]. The Shannon-Nyquist bandlimited hypothesis is replaced by a more flexible notion of limited bandwidth called the *finite rate of innovation* [VMB02]. This relaxed hypothesis is also of importance to address (b). Indeed, this allows exact representation of *piecewise defined signals* using basis functions that have *finite temporal support*. In particular, we will work with piecewise polynomials spaces [UAE93a, UAE93b], and piecewise exponential spaces [UB05, Uns05].



(a) Bandwidth expansion by a nonlinearity function $N(\cdot)$. The spectrum of a distorted sinusoid usually contains an infinite number of harmonics.



(b) Time-limited signals are not bandlimited

Figure 3.1 – Common sources of non-bandlimitednes: nonlinearities and finite temporal support.

Outline In section 3.1, we recall results and notations from generalized sampling theory and functional analysis. In section 3.2, we consider continuous-time input reconstruction, in piecewise-defined signal spaces, i.e. the realisation of the "Virtual DAC" module in figure 3.2 (block A). In section 3.3, we consider the realisation of the dual output anti-aliasing, and sampling modules, i.e. implementations strategies and choices to implement an anti-aliased "Virtual ADC" (blocks D-E). In particular we consider two problems: exact continuous-time solutions of LTI ARMA filters with piecewise polynomial inputs and projection of piecewise discontinuous polynomials on smooth B-spline spaces [UAE93a, UAE93b]. Finally, in section 3.4, as a validation test, we illustrate this "virtual analog" toolchain with an original implementation of a common audio effect: a "virtual analog" sampling rate reduction effect (emulating artefacts of old ADC-DAC).

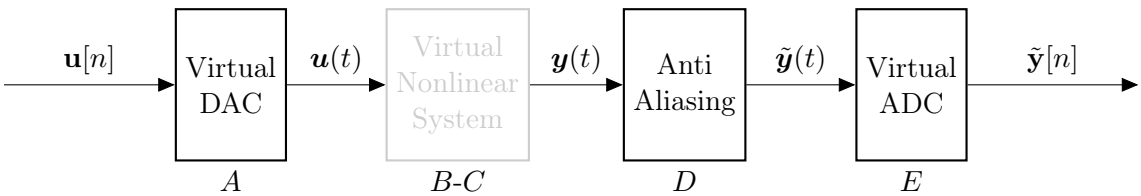


Figure 3.2 – (continuous-time virtual analog signal processing) block-diagram of the approach. In this chapter, input reconstruction (Virtual DAC) and output antialiasing/ sampling (virtual ADC) are considered.

3.1 Generalized-sampling theory and Finite Rate of Innovation

3.1.1 Short reminder on functional analysis

Here we provide a short reminder on functional analysis and fix some notations. For more details refer to the definitions in appendix C p.281 on Banach spaces, Hilbert spaces, Sobolev spaces, etc). Let $\Omega = (0, 1)$ be the unit interval and $\mathbb{I} \subseteq \mathbb{Z}$ a countable set.
The inner product of the Hilbert space of square integrable functions $L^2(\Omega, \mathbb{R}^n)$ is

$$\langle \boldsymbol{u}, \boldsymbol{v} \rangle_{L^2(\Omega,\mathbb{R}^n)} := \int_\Omega \boldsymbol{u}(\tau) \cdot \boldsymbol{v}(\tau)\, \mathrm{d}\tau, \qquad \forall \boldsymbol{u}, \boldsymbol{v} \in L^2(\Omega). \tag{3.1}$$

The inner product of the Hilbert space of square summable sequences $\ell^2(\mathbb{I}, \mathbb{R}^n)$ is

$$\langle \mathbf{u}, \mathbf{v} \rangle_{\ell^2(\mathbb{I},\mathbb{R}^n)} := \sum_{k\in\mathbb{I}} \mathbf{u}[k] \cdot \mathbf{v}[k], \qquad \forall \mathbf{u}, \mathbf{v} \in \ell^2(\mathbb{I}). \tag{3.2}$$

In this manuscript, we identify the space L^2 with its dual $(L^2)^* \simeq L^2$ (used as a pivot space). This means that for a space V and its (algebraic) dual V^* (def. C.19), we have the inclusions

$$V \subseteq L^2 \subseteq V^*,$$

where the (functional) duality product between V^* and V is (note that V and V^* can be swapped)

$$\langle \boldsymbol{u} \,|\, \boldsymbol{v} \rangle := \int_\Omega \boldsymbol{u}(\tau) \cdot \boldsymbol{v}(\tau)\, \mathrm{d}\tau, \quad \forall \boldsymbol{u}, \boldsymbol{v} \in V^* \times V. \tag{3.3}$$

Remark 3.1 (Dirac bra-ket notations)**.** To simplify proofs and enhance readability (without any reference to quantum mechanics) we use Dirac *bra-ket* notations (i.e. the functional analogs of a transposed vector and a vector).

- A *ket* $|\,\psi\rangle$ denotes a synthesis operator from coefficients to functions.
- A *bra* $\langle\phi\,|$ is an analysis functional that returns a number and receives a function.
- A *bra-ket* $\langle\phi\,|\,\psi\rangle$ denotes a contraction (or inner product). It returns a number.
- A *ket-bra* $|\,\phi\rangle\,\langle\psi\,|$ denotes an analysis-synthesis operator.
- $\langle \boldsymbol{u} \,|\, \mathcal{A} \,|\, \boldsymbol{v} \rangle$ is used as a shorthand for $\langle \boldsymbol{u}, \mathcal{A}\boldsymbol{v} \rangle_{L^2} = \langle \mathcal{A}^*\boldsymbol{u}, \boldsymbol{v} \rangle_{L^2}$ where $\mathcal{A}^*$ denotes the adjoint operator (see def. C.16 p.283). This is the functional equivalent of the matrix notation $\mathbf{u}^\mathsf{T}\mathbf{A}\mathbf{v} = \mathbf{u}^\mathsf{T}(\mathbf{A}\mathbf{v}) = (\mathbf{A}^\mathsf{T}\mathbf{u})^\mathsf{T}\mathbf{v}$.

Definition 3.1 (Frame [Chr16])**.** Let V be an inner product space and $F = \{\phi_k\}$ a set of vectors in V, then these vectors satisfy the *frame condition* if there are positive real numbers A and B such that $0 < A < B < \infty$ and for each $v \in V$

$$A\|v\|_V^2 \leq \sum_{k\in\mathbb{I}} \big|\langle \phi_k \,|\, v \rangle\big|^2 \leq B\|v\|_V^2\,. \tag{3.4}$$

Furthermore a frame F is said to be *tight* if $A = B$, a *Parseval frame* if $A = B = 1$. It is a *Riesz basis* if F is a basis, otherwise F is said to be an *overcomplete frame*. For example, an orthonormal basis, is at the same time, a tight frame, a Parseval frame and a Riesz basis.

The frame condition guarantees the well-posedness of analysis and resynthesis operators but not the uniqueness of their representation. By contrast, if F is a Riesz basis, then there exists a *unique* dual basis $\tilde{F}$ (defined below) such that $\langle \tilde{\phi}_i | \phi_j \rangle = \delta_{ij}$.

Frame synthesis operator For a basis or frame $\{\phi_k\}$ such that $V = \text{span}\{\phi_k\} \subset L^2(\Omega)$, we introduce the *frame synthesis operator* $\mathbf{\Phi} : \ell^2(\mathbb{I}) \to V$ defined by

$$\mathbf{\Phi} := \begin{bmatrix} \dots & |\,\phi_k\rangle & \dots \end{bmatrix}_{k\in\mathbb{I}}, \tag{3.5}$$

so that using the vector of coefficients $\vec{\mathbf{x}}$, we can compactly write a function as $\boldsymbol{x}(t) = (\mathbf{\Phi}\vec{\mathbf{x}})(t)$.

Frame analysis operator Dually, we define the *frame analysis operator* $\mathbf{\Phi}^* : L^2(\Omega) \to \ell^2(\mathbb{I})$.

$$\mathbf{\Phi}^* := \begin{bmatrix} \vdots \\ \langle\phi_k\,| \\ \vdots \end{bmatrix}_{k\in\mathbb{I}} \tag{3.6}$$

so that the coefficients $\vec{\mathbf{x}}^*$ of a function $\boldsymbol{x}(t)$ are given by $\vec{\mathbf{x}}^* = \mathbf{\Phi}^*\boldsymbol{x}$.

Gram Matrix The *Gram Matrix* (or gramian) of the frame $\mathbf{\Phi}$ is defined by

$$\mathbf{G}_{\mathbf{\Phi}} := \mathbf{\Phi}^*\mathbf{\Phi} = \left[\langle\phi_m\,|\,\phi_n\rangle\right]_{m,n\in\mathbb{I}}. \tag{3.7}$$

Dual Frame If $\mathbf{\Phi}$ is a frame, then a *dual frame* $\tilde{\mathbf{\Phi}}$, is a frame such that $\tilde{\mathbf{\Phi}}^*\mathbf{\Phi} = \mathbf{I}_{\mathbb{I}}$.

Dual Basis If $\mathbf{\Phi}$ is a basis, then its *dual basis* (or biorthogonal basis) $\tilde{\mathbf{\Phi}}$ is the linear combination of basis functions obtained using the inverse of the Gram Matrix.

$$\tilde{\mathbf{\Phi}} = \mathbf{\Phi}\mathbf{G}_{\mathbf{\Phi}}^{-1}. \tag{3.8}$$

Proof. Using (3.7)-(3.8), we have $\tilde{\mathbf{\Phi}}^*\mathbf{\Phi} = (\mathbf{\Phi}\mathbf{G}_{\mathbf{\Phi}}^{-1})^*\mathbf{\Phi} = \mathbf{G}_{\mathbf{\Phi}}^{-\mathsf{T}}\mathbf{\Phi}^*\mathbf{\Phi} = \mathbf{G}_{\mathbf{\Phi}}^{-1}\mathbf{\Phi}^*\mathbf{\Phi} = \mathbf{G}_{\mathbf{\Phi}}^{-1}\mathbf{G}_{\mathbf{\Phi}} = \mathbf{I}_{\mathbb{I}}$. □

Reproducing Kernel If $\{\phi_k(\tau)\}_{k\in\mathbb{I}}$ is an orthonormal basis of a space $V \subseteq L^2(\Omega)$, then according to Mercer's theorem, the *reproducing kernel* of V is

$$K_V(\tau,\sigma) := \sum_{k\in\mathbb{I}} \phi_k(\tau)\phi_k(\sigma), \tag{3.9}$$

so that we can express the projector $\mathcal{P}_V$ using the reproducing kernel K_V as

$$(\mathcal{P}_V u)(\tau) := \int_\Omega K_V(\tau,\sigma)u(\sigma)\,\mathrm{d}\sigma = \sum_{k\in\mathbb{I}} |\,\phi_k\rangle\,\langle\phi_k\,|\,u\rangle\,. \tag{3.10}$$

Resolutions of the Identity If $\mathbf{\Phi}$ represents an orthonormal basis, then by definition

$$\mathbf{\Phi}^*\mathbf{\Phi} = \left[\langle\phi_m\,|\,\phi_n\rangle\right]_{m,n\in\mathbb{I}} = \mathbf{I}_{\mathbb{I}}. \tag{3.11}$$

Conversely, the projector $\mathcal{P}_V : L^2(\Omega) \to V$ is given by

$$\mathbf{\Phi}\mathbf{\Phi}^* = \sum_{k\in\mathbb{I}} |\,\phi_k\rangle\,\langle\phi_k\,| = \mathcal{P}_V. \tag{3.12}$$

When $\mathcal{P}_V$ is restricted to functions in V, then $\mathcal{P}_V = \mathcal{I}_V$ where $\mathcal{I}_V$ denotes the identity operator.

Partition of unity A generator $\varphi(t)$ satisfies the *partition of unity* property if the sum of its integer translates sums to one.

$$\sum_{n\in\mathbb{Z}} \varphi(t-n) = 1, \quad \forall t \in \mathbb{R}. \tag{3.13}$$

3.1.2 Class of signals and notations

We introduce the class of signals and the notations that are used in this thesis. A vector-valued signal $\boldsymbol{x} : t \in \mathbb{R} \to \boldsymbol{x}(t) \in \mathbb{R}^m$ is represented as *a sequence of time frames* $\boldsymbol{x}_n(\tau)$

$$\boldsymbol{x}(t) = \sum_{n\in\mathbb{Z}} \underbrace{\left(\sum_{i=0}^{p-1} \phi_i \left(\frac{t-t_n}{h_n} \right) \mathbf{x}_{n,i} \right)}_{\boldsymbol{x}_n(\tau)}, \quad \text{where} \quad \tau = \frac{t-t_n}{h_n} \tag{3.14}$$

where

- $\mathbb{T} = \{t_n\}_{n\in\mathbb{Z}}$ is a monotonic partition of time ($t_n < t_{n+1}$),
- $h_n = t_{n+1} - t_n$ is the local *step size*,
- p is the *number of basis functions* and $\mathbb{I} = \{0, \dots p-1\}$,
- The generating functions $\{\phi_i(\tau)\}_{i\in\mathbb{I}}$ form the local *representation basis*,
- $\mathbf{x}_{n,i} \in \mathbb{R}^m$ are the vector-valued *coefficients* for each time-step n and basis index i,
- $\tau = \frac{t-t_n}{h_n}$ is the *normalized local time* for time-step n,
- $\boldsymbol{x}_n(\tau)$ is the local representation of $\boldsymbol{x}(t)$ at time-step n.

The *generating functions* $\phi_0, \dots, \phi_{p-1}$ and their translates span the approximation space

$$V = \operatorname{span} \left\{ \phi_i \left((t-t_n)/h_n \right), \quad \forall i \in \mathbb{I}, \quad n \in \mathbb{Z} \right\} \otimes \mathbb{R}^n.$$

This class of signals is related to (time) finite elements and multi-wavelets [1] (see [Uns00, section C]). For *causality of computations*, basis functions translates are *non overlapping*. When the context is not ambiguous, we drop the temporal subscript n. We talk about the *local trajectory*

$$\boldsymbol{x}(\tau) = \sum_{i=0}^{p-1} \phi_i(\tau) \mathbf{x}_i.$$

To simplify the presentation, we restrict to a constant step-size [2] h ($h_n = h$, $\forall n \in \mathbb{Z}$) so that the approximation space V is *integer shift-invariant*. Generating functions ϕ_i are defined over the open unit interval $\Omega = (0,1)$ with boundary $\partial\Omega = \{0,1\}$ and closure $\overline{\Omega} = \Omega \cup \partial\Omega = [0,1]$.

> **Remark 3.2.** The tensor of coefficients $[x_{n,i,j}]$ may be denoted by $\mathbf{x}_{n,i}$ or $\mathbf{x}_i[n]$ according to the way it is "sliced" in each context, i.e. when a clear distinction between the different roles of time index n, functional basis index i and "geometric" index j is required. We also use $\boldsymbol{x}[n](\tau)$ as a synonym for $\boldsymbol{x}_n(\tau)$ [a].
>
> *a.* For example, we use the notation $\mathbf{x}_i[n]$ (resp. $\boldsymbol{x}[n](\tau)$) to emphasize the sequence of coefficients (resp. functions) interpretation. This is particularly useful when working in the Z-domain.

1. In the multi-wavelets literature the generators $\phi_0, \dots, \phi_{p-1}$ are called multi-scaling functions.
2. The numerical methods in this thesis, in particular in chapter 5, do not require a constant step size, they can be adapted by taking the step size into account when computing derivatives of the solution.

3.1.3 Sampling signals with a Finite Rate of Innovation (FRI)

Classical bandlimited sampling

The vast majority of digital (audio) signal processing relies on the following theorem.

> **Theorem 3.1** (Shannon sampling theorem [Sha49]). *If a function $x(t)$ contains no frequencies higher than B cycles per second, it is completely determined by giving its ordinates at a series of points spaced $h = 1/2B$ seconds apart.*

The reconstruction formula that complements the sampling theorem is

$$x(t) = \sum_{n\in\mathbb{Z}} \operatorname{sinc}\left(\frac{t}{h} - n\right) x_n, \qquad \text{where} \qquad \operatorname{sinc}(x) := \frac{\sin(\pi x)}{\pi x} \qquad \text{and} \qquad x_n = f(hn). \tag{3.15}$$

Equation (3.15) is exact when x is *bandlimited* to $f_{max} < B$, called the *Nyquist* frequency. Coefficients $\{x_n\} \in \ell^2(\mathbb{Z})$ are called *samples* of x and $f_s = 2B$ is called the *sampling rate*.

Modern Sampling

Generalized sampling theory accounts for the fact that real world signal are not exactly bandlimited and ideal band-limiting filters do not exist. Nevertheless, perfect analysis and reconstruction of signals is still possible if we assume that they have a *finite rate of innovation*.

The paradigm shift in modern sampling is to realize that (3.15) is an *orthogonal decomposition* and that ideal bandlimiting and sampling is simply a way to compute the *projection coefficients*[3]

$$x(t) = \sum_{n\in\mathbb{Z}} \varphi_n(t) x_n, \qquad \text{where} \qquad \varphi_n(t) = \operatorname{sinc}(t/h - n), \qquad \text{and} \qquad x_n = \langle \varphi_n, x \rangle . \tag{3.16}$$

Shannon bandlimited sampling is an instance of the more general (and practical) situation. Let $\varphi(t)$ be a generating function such that $\left\{\varphi_n = \varphi(\cdot/h - n)\right\}_{n\in\mathbb{Z}}$ is a Riesz basis of the non-bandlimited integer shift invariant space $V_h(\varphi) = \operatorname{span}\left\{\varphi(\cdot/h - n)\right\}_{n\in\mathbb{Z}}$ in $L^2(\mathbb{R})$. One further requires that φ satisfies the partition of unity property[4] (3.13). Then there exists a dual basis $\{\tilde{\varphi}_n\}$ of V_h such that signals in V_h are perfectly reconstructed[5] according to

$$x(t) = \sum_{n\in\mathbb{Z}} \varphi_n(t) x_n \qquad \text{where} \qquad x_n = \langle \tilde{\varphi}_n, x \rangle . \tag{3.17}$$

Note that, by construction, signals spaces such as (3.14) fulfil the *finite rate of innovation* property. They can be *exactly represented* (over a multi-generator basis) using a finite number of degrees of freedom p per time-step h called the *generalized bandwidth* [VMB02]

$$B = \frac{p}{h}. \tag{3.18}$$

Also note that, in our case, it is enough to have the *constant reproduction property*[6] over each time step to fulfil the partition of unity (for all $t \in \mathbb{R}$). It turns out that constant reproduction is also a necessary condition to obtain *consistent numerical integration schemes* (eq. (5.21a) p.128).

3. It happens that the sinc system is both orthonormal in $L^2(\mathbb{R})$ and interpolating, i.e. the sinc function (and its integer translates) is the generator of the space of bandlimited signals.
4. This guarantees that the approximation is consistent, so that one can approximate any function of $L^2(\mathbb{R})$ over the space V_h as closely as desired (in norm) for a small enough sampling step h.
5. B-spline sampling is a typical example of perfect reconstruction in non-bandlimited spaces.
6. meaning that constant functions belongs to the approximation space.

Example 3.1. Piecewise polynomial signals are not *band-limited* in the sense of Shannon (see (3.14) where ϕ). For example, the discontinuities in a sequence of piecewise constant signals (at the output of a sample and hold circuit for example) have an *infinite spectrum* (see figure 3.1).

Approximation order, polynomial reproduction and Strang–Fix conditions

We recall result (3.22) from [Uns00, section IV] relating the approximation order of the sampling space, the spectral flatness of the approximation error in the Fourier domain and the capability of the approximation space to reproduce polynomials.

Let $\mathcal{Q}_h : L^2(\mathbb{R}) \to V_h(\varphi) \subset L^2(\mathbb{R})$ denote the linear approximation operator defined by

$$(\mathcal{Q}_h f)(t) = \sum_{n\in\mathbb{Z}} \varphi\left(\frac{t}{h} - n\right) \left\langle \varphi\left(\frac{\cdot}{h} - n\right), f \right\rangle \tag{3.19}$$

and the approximation error by $\epsilon_h(f) = \|f - \mathcal{Q}_h f\|_{L^2}$. Averaging ϵ_h over all time-shifts, it happens that one can characterise the average error in the frequency domain as

$$\bar{\epsilon}_h^2(f) := \frac{1}{h}\int_0^h \left\|f(\cdot - \tau) - \mathcal{Q}_h f(\cdot - \tau)\right\|^2 \mathrm{d}\tau = \int_{\mathbb{R}} E_\varphi(h\omega)\left|\widehat{f}(\omega)\right|^2 \frac{\mathrm{d}\omega}{2\pi}, \tag{3.20}$$

where $\widehat{f}$ denotes the Fourier transform of f and $E_\varphi(\omega)$ is the error kernel given by

$$E_\varphi(\omega) = \left|1 - \widehat{\tilde{\varphi}}^*(\omega)\widehat{\tilde{\varphi}}(\omega)\right| + \left|\widehat{\varphi}(\omega)\right|^2 \sum_{k\neq 0}\left|\widehat{\varphi}(\omega + 2k\pi)\right|^2. \tag{3.21}$$

One can predict the rate of decay of the approximation error from the degree of *flatness* of $E_\varphi(\omega)$ near the origin. If $E_\varphi(\omega) = C^2\omega^{2L} + \mathcal{O}(\omega^{2(L+1)})$ as $\omega \to 0$, then [Uns00, eq.45]

$$\|f - \mathcal{Q}_h f\|_{L^2} = Ch^L \left\|f^{(L)}\right\|_{L^2} \qquad \text{as} \qquad h \to 0. \tag{3.22}$$

for $f \in H^L(\mathbb{R})$. This implies that the error decays globally like $\mathcal{O}(h^L)$ and is called the *order of approximation*. It happens that through the *Strang–Fix conditions* [FS69, JL93, Cha99] (see also appendix C.3 p.285) property (3.22) is equivalent to the reproduction of polynomials of degree $L - 1$.

Remark 3.3 (Peano kernels)**.** In complement to the asymptotic error-bound estimate (3.22), the error shape can be analysed thanks to Peano kernels presented thereafter (see (5.30) p.131). In subsection 5.2.7, we have a closer look at error measures such as (3.22) by studying the Peano kernels of approximation operators used in power-balanced integration methods.

Remark 3.4 (Accuracy order and Strang–Fix conditions)**.** In the ODE literature, order conditions of one-step methods are usually investigated using the combinatorial theory of B-series [MMMKV17, HLW06]. As an interresting result, bridging sampling and numerics through Strang-Fix conditions, we show in subsection 5.2.6 p.128 that if (continuous-stage) Runge–Kutta methods are built on orthogonal projection (of the vector field) which reproduces polynomials of order p. Then, the local truncation error has accuracy order $2p$ (automatically fulfilling B-series order conditions). We note that this result reveals itself in the continuous-time setting whereas it remains hidden using standard (discrete) RK formulations.

3.1.4 Piecewise polynomial frames

Let $\mathbb{P}^n(\Omega, F)$ be the space of F-valued polynomials of maximal degree n over the domain Ω. We sometimes drop F when $F = \mathbb{R}$ and Ω when $\Omega = (0, 1)$. This section quickly mentions a few important polynomial bases and their main properties.

Monomial Basis The canonical basis of polynomials is given by the *monomial basis* $\{M_k(\tau)\}$ where

$$M_k(\tau) := \frac{\tau^k}{k!} \tag{3.23}$$

and satisfies the derivation property (i.e. they correspond to Green functions of $\frac{\mathrm{d}^i}{\mathrm{d}\tau^i}$)

$$\frac{\mathrm{d}^i}{\mathrm{d}\tau^i} M_k = \begin{cases} M_{k-i} & i \leq k \\ 0 & i > k \end{cases} \tag{3.24}$$

This basis is not orthogonal, which leads to bad conditioning for some numerical applications. However we will use it in subsection 3.3.1 to obtain closed-form filtering of sequences of polynomials.

Shifted Orthonormal Legendre polynomials By Gram-Schmidt orthogonalisation of the monomial basis in L^2, one obtains the *shifted orthonormal Legendre* polynomial basis. They have the explicit representation

$$L_k(\tau) = \frac{\sqrt{2k+1}}{k!} \frac{\mathrm{d}^k}{\mathrm{d}\tau^k} \tau^k (\tau - 1)^k. \tag{3.25}$$

Important properties of Legendre polynomials are detailed in appendix C.4. This is the main basis used in projection methods of chapter 5.

Bernstein polynomials Another useful basis of the polynomial space $\mathbb{P}^n$ is given by the *Bernstein basis* [Far12]

$$B_k^n(\tau) = \binom{n}{k} (1 - \tau)^{n-k} \tau^k. \tag{3.26}$$

This basis is not orthogonal, but it is useful to represent Bezier splines by their *control polygon* $\{\mathbf{x}_k\}$

$$\boldsymbol{x}(\tau) = \sum_{k=0}^{n} B_k^n(\tau) \mathbf{x}_k. \tag{3.27}$$

They satisfy a number of interesting properties. In particular the continuous derivative and integral operators translate to finite differences and finite sums of their discrete control points, and the curve is contained in the convex hull formed by the control polygon [Far12].

Hermite splines Hermite splines (defined in (C.22) p.287) are closely related to the Bernstein basis but the representation uses derivatives of functions on the left and right boundaries of the interval as coefficients. It is useful in *derivative sampling* and function interpolation. Hermite splines and their generalisation will appear in 5.2.7 p.129, to address $\mathcal{C}^k$-continuous trajectories.

B-splines (Cardinal) B-spline (see [UAE93a] [UAE93b]) are smooth finite-support continuous functions whose restriction to the unit interval are piecewise polynomials. B-splines are defined and used below in subsections 3.2.1 and 3.3.2 dedicated to input reconstruction and output projection.

3.2 Input reconstruction (Virtual DAC)

No matter how accurate simulations methods can be, the response of the overall system is limited by the quality of the input reconstruction. To reconstruct a continous-time input $u(t)$ from discrete samples u_n, it is not practical to use Shannon's bandlimited interpolation formula (3.15) because it is both *acausal* and the sinc kernel has *infinite temporal support*[7]. By consequence, the bandlimited input reconstruction is not computable. Instead, using generalised sampling theory (see the overview paper [Uns00]), we consider *computable* non bandlimited approximations of bandlimited spaces whose synthesis functions have finite temporal support.

3.2.1 B-spline spaces

Following the standard approach in [UAE93a] we consider reconstruction of the input in compactly supported B-spline spaces[8] (B-splines basis functions are shown on figure 3.4)

$$u(t) = \sum_{n=-\infty}^{\infty} \beta_m\left(\frac{t}{h} - n\right)\tilde{u}_n \quad \text{where} \quad \widehat{\beta_m}(\omega) := \left(\frac{e^{j\frac{\omega}{2}} - e^{-j\frac{\omega}{2}}}{j\omega}\right)^{m+1} = \operatorname{sinc}^{m+1}\left(\frac{\omega}{2}\right) \tag{3.28}$$

where function β_m denotes the centred B-spline of order m and $\widehat{\beta_m}$ its Fourier transform.

Prefiltering The coefficients $\tilde{u}_n$ are computed from the cardinal samples $u(t_n)$ using the *discrete B-spline IIR pre-filter* $S_m(z)$, whose Z-transform is the *inverse of the B-spline FIR filter* $B_m(z)$

$$S_m(z) = \frac{1}{B_m(z)} \qquad \text{with} \qquad B_m(z) = \sum_{k=-\lceil m/2\rceil}^{\lceil m/2\rceil} \beta_m(k) z^k. \tag{3.29}$$

The block diagram of the method is shown on figure 3.3 (where $\star$ is the convolution operator).

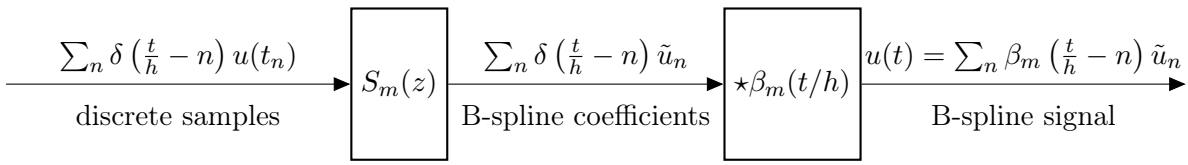


Figure 3.3 – Digital IIR prefiltering scheme to obtain B-spline coefficients $\{\tilde{u}_n\}$ such that the reconstructed function $u(t)$ interpolates the cardinal samples $\{u(t_n)\}$.

Piecewise polynomials Since B-splines are piecewise polynomials, for each time-frame $\Omega_n = (t_n, t_n + h)$, $t_n = hn$, the restriction of the signal $u(t_n + h\tau)$ to the interval Ω_n is exactly representable as a polynomial, it is thus suitable for use in our one-step simulation framework, which requires inputs to be specified as *sequences of time frames*. It is given by

$$u(t_n + h\tau)\Big|_{\Omega_n} = \sum_{k=n-(m+1)/2}^{n+(m+1)/2} \beta_m(\tau - k)\Big|_{[0,1]} \cdot \tilde{u}_k, \qquad u(t_n + h\tau)\Big|_{\Omega_n} \in \mathbb{P}^m([0,1]). \tag{3.30}$$

7. A finite approximation of the Shannon bandlimited interpolation formula and approximate integration of windowed sinc interpolation using quadratures has been proposed in [SH11]

8. This approach is more suitable for our time-stepping framework and it is known that the limit when $m \to \infty$ converges to bandlimited spaces.

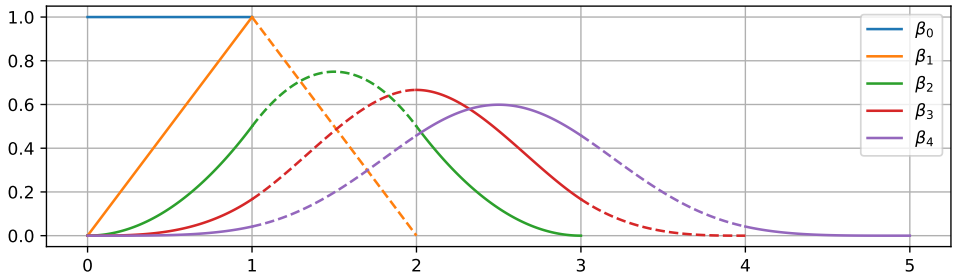


Figure 3.4 – B-splines (non centered). Piecewise polynomial segments are emphasised using alternating solid and dashed lines.

Cardinal interpolating splines It is possible to combine[9] B-splines with their prefilter. This gives the following interpolation formula expressed using the cardinal interpolating splines β^m_{int}

$$u(t) = \sum_{k=-\infty}^{\infty} \beta_m^{\mathsf{int}}\left(\frac{t}{h} - k\right) u(t_k) \qquad \text{where} \qquad \widehat{\beta}_m^{\mathsf{int}}(\omega) = \frac{\widehat{\beta^m}(\omega)}{B_m(e^{j\omega})}. \tag{3.31}$$

It is shown on figure 3.7 (d) that the prefilter has the role of a pre-emphasis filter that compensates the lowpass characteristic of B-splines so that the magnitude response of cardinal splines is maximally flat below the Nyquist frequency. Time and frequency responses of B-spline and corresponding cardinal interpolating splines are shown on figure 3.7 .

Causality The above approach is adapted in image processing where causality is not an issue, however for audio signal processing, acausality of the discrete prefilter $S_m(z)$ is an important issue that needs to be addressed. Several approaches can be considered:

- If phase linearity (i.e. constant delay) is considered more important than latency, it is possible to approximate the IIR filter $S_m(z)$ by an optimal FIR $S_m^{\mathsf{FIR}}(z)$. Furthermore since the impulse reponse $s_m[n]$ of the filter $S_m(z)$ decays quickly, an accurate approximation can be obtained with short FIR filters (see figure 3.8).

- If instead a minimal group delay is desired, it is possible to convert S_m to minimum phase so that both filters share the same magnitude response $\left|S_m(e^{j\omega})\right| = \left|S_m^{\mathsf{minphase}}(e^{j\omega})\right|$ while the minimum phase filter has a stable realization because it only has stable poles.

If we restrict to piecewise affine spaces, a cost-effective approach consists in using *shifted-linear interpolation* which is detailed thereafter (see figure 3.5).

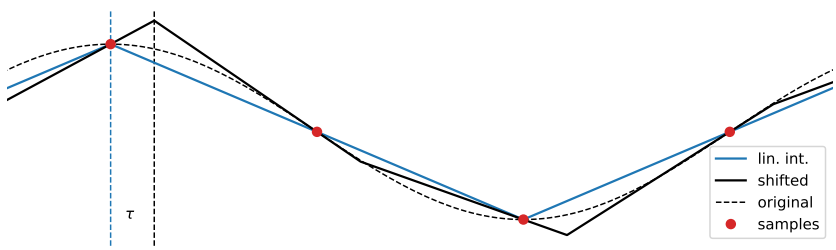


Figure 3.5 – Comparison of shifted and standard linear interpolation.

9. In practice, since interpolating splines are infinitely supported, it is computationally more interesting to work with finitely supported B-splines, and rely on IIR pre-filtering to obtain their coefficients.

3.2.2 Shifted linear interpolation

We restrict to B-spline spaces of degree 1, following the approach presented in [BTU04]. Instead of using standard linear interpolation whose frequency response is $\mathrm{sinc}^2(\omega/2)$, by relaxing phase linearity, it is possible to both obtain a causal IIR prefilter and to improve the frequency response of the interpolator. The mean to obtain this improvement is to use *shifted linear interpolation* (see figure 3.5). The main idea is the following: instead of using the following (trivial) B-spline prefilter to obtain a cardinal interpolating spline (i.e. here $\beta_1 = \beta_1^{\mathsf{int}}$)

$$S_1(z) = \frac{1}{\beta_1(-1)z + \beta_1(0) + \beta_1(1)z^{-1}} = 1 \qquad \text{where} \qquad \beta_1(t) := |1-t|_+ ,$$

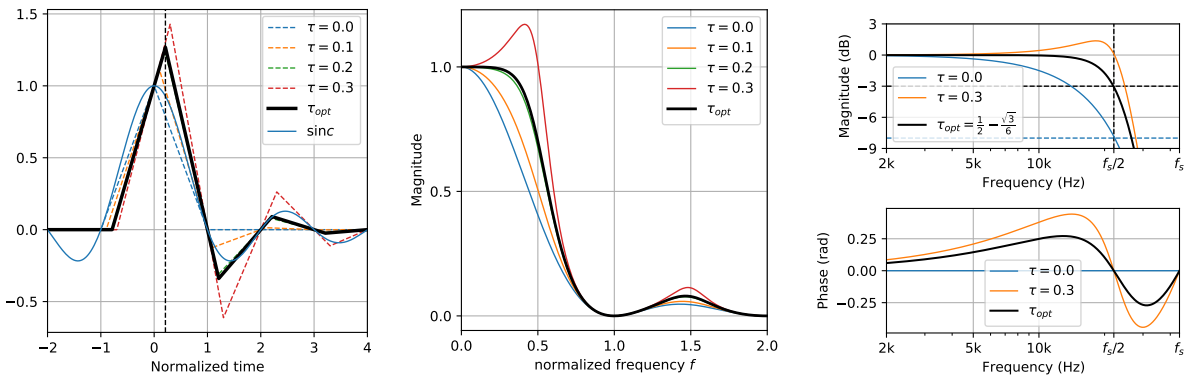
one can use the inverse of the *shifted B-spline FIR filter* $\beta_1(\cdot - \tau)$ to pre-filter the samples $\{u(t_n)\}$

$$S_{1,\tau}(z) = \frac{1}{(1-\tau) + \tau z^{-1}}. \tag{3.32}$$

It turns out [BTU04] that there exists an optimal shift [10] $\tau_{\mathsf{opt}} = \frac{1}{2} - \frac{\sqrt{3}}{6}$ for which the magnitude response of the cardinal interpolating spline is maximally flat. This gives the optimal IIR prefilter

$$S_1^{\mathsf{opt}}(z) = \frac{b_0}{1 + a_1 z^{-1}}, \qquad \text{where} \qquad b_0 = \frac{1}{1-\tau_{\mathsf{opt}}}, \qquad a_1 = \frac{\tau_{\mathsf{opt}}}{1-\tau_{\mathsf{opt}}}. \tag{3.33}$$

The frequency response is shown on figure 3.6. To conclude this paragraph on shifted linear interpolation, for only a small additional cost (a causal discrete first order IIR pre-filter followed by standard linear interpolation), the frequency response of linear interpolation is significantly improved and can compete with higher order cardinal interpolating splines from figure 3.7 .

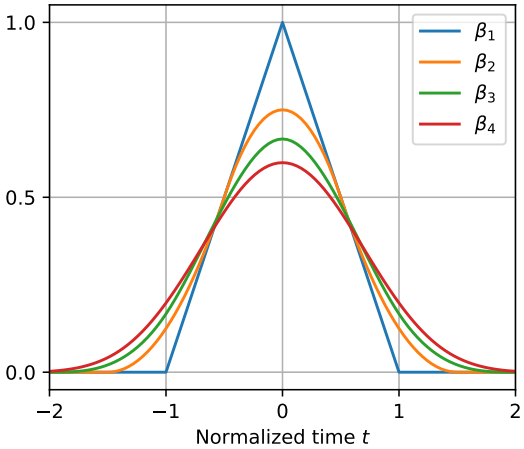


(a) Shifted interpolating splines **(b)** Fourier kernels $\widehat{\beta_{1,\tau}^{\mathsf{int}}}$ **(c)** Fourier kernels (zoomed)

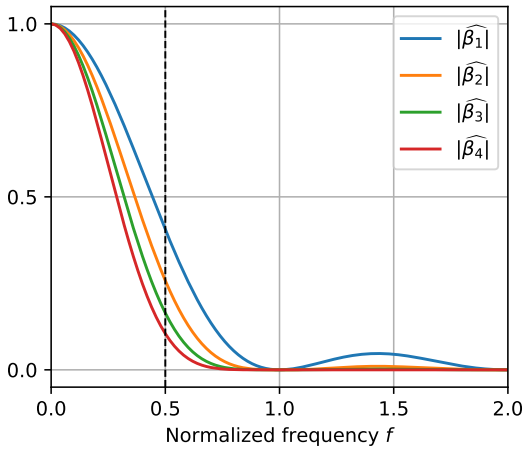
Figure 3.6 – Time and frequency response of shifted linear interpolation: $\widehat{\beta_{1,\tau}^{\mathsf{int}}}(\omega) = \widehat{\beta}_1(\omega) S_{1,\tau}^{\tau}(z = e^{j\omega})$. Note that cardinal splines are interpolating on the integer grid, but their maximum value is reached for the timeshift τ. For the optimal shift τ^{opt}, the magnitude response is improved by up to 5dB between 5kHz and $f_s/2$ compared to standard linear interpolation.

To sum up: for low order reconstruction, shifted linear interpolation is both causal and cost-effective; for higher order reconstruction, causal approximations of B-spline prefilters and higher latency are required (see figure 3.8).

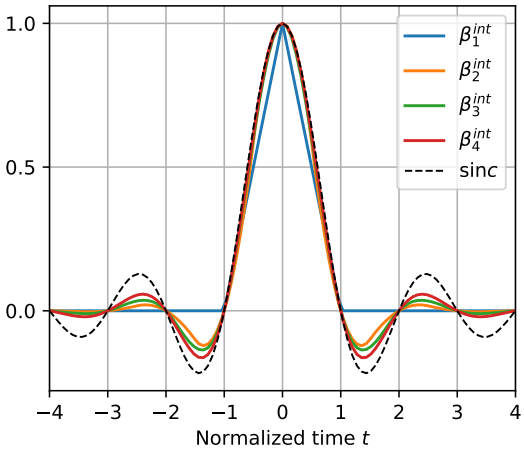
10. We note by anticipation, that the optimal shift corresponds to a Gauss quadrature node ($\frac{1}{2} - \frac{\sqrt{3}}{6}$ is the smallest root of the second shifted Legendre polynomials P_2 which is used in Gauss-Legendre numerical integration methods [HLW06]). This is the second time in this chapter (see remark 3.4 above) that we discover unexpected connections between numerical analysis and signal processing. A dedicated study would be required to reveal the fundamental causes behind these apparent co-incidences. Legendre polynomials are detailed in appendix C.4 p.286.



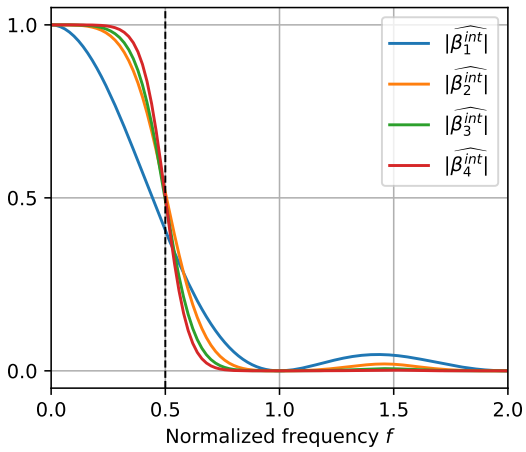
(a) (centred) B-splines



(b) Fourier transform of B-splines



(c) cardinal interpolating B-splines



(d) Fourier transform of cardinal interpolating B-splines

Figure 3.7 – Comparison between B-splines and cardinal interpolating B-splines (3.31). B-splines have finite support and a lowpass frequency response (both time and frequency representations converges to gaussians when order is increasd). By contrast, cardinal interpolating B-splines have infinite support in both time and frequency (but both decay quickly). The major difference, comes from the the fact that cardinal B-splines are interpolating (they vanish on the uniform grid except in 0) and their frequency response below the Nyquist frequency is much sharper: it converges to the ideal bandlimited rectangular kernel when order is increased.

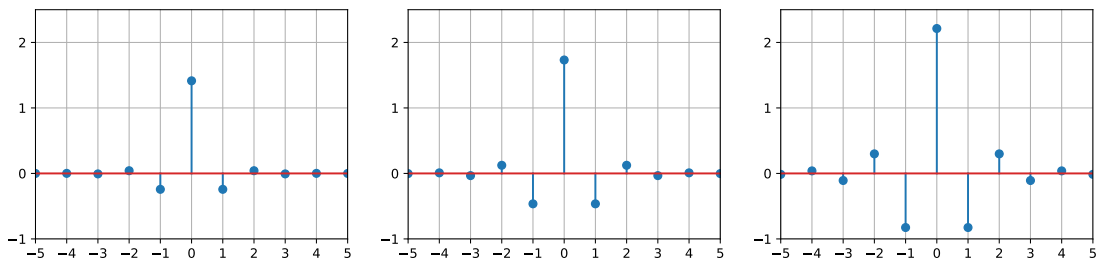


Figure 3.8 – Impulse responses of cardinal interpolating B-spline pre-filters $s_2[n]$, $s_3[n]$, $s_4[n]$ (see equation(3.29)).

3.3 Output antialiasing and sampling (Virtual ADC)

In this section, we consider Virtual Analog to Digital Converters (vADC), their implementation and different design tradeoffs. We propose two approaches. First, in subsection 3.3.1, we consider the exact implementation of continuous-time Linear Time-Invariant ARMA filters represented as state-space systems. This strategy allows the use of all analog filter design tools to implement anti-aliasing filters (Butterworth, Chebyshev, Elliptic, etc). Second, in subsection 3.3.2, to mirror input reconstruction in shift-invariant B-splines spaces, we propose an alternative strategy. Given a (potentially discontinuous) signal $y(t)$ defined as a (broken) piecewise polynomial, we look for the best approximant $\tilde{y}(t)$ in B-splines spaces (the dual problem of input reconstruction).

3.3.1 Exact continuous-time filtering for LTI state-space systems

Let $u(t)$ be a non band-limited signal with a finite rate of innovation B (see 3.18). For band-limiting purposes [11], we would like to apply an exact continuous-time antialiasing filter.

We consider the class of Linear Time-Invariant (LTI) state-space filters

$$\dot{\boldsymbol{x}}(t) = \mathbf{A}\boldsymbol{x}(t) + \mathbf{B}u(t), \tag{3.34a}$$

$$y(t) = \mathbf{C}\boldsymbol{x}(t) + \mathbf{D}u(t), \tag{3.34b}$$

and assume that the input signal $u(t)$, is locally defined for each time step by $u(t_n + h\tau) = u[n](\tau)$ for $\tau \in (0,1)$ over a basis $\Phi = \{\phi_1(\tau), \dots, \phi_p(\tau)\}$ as follows

$$u[n](\tau) = \sum_{i=1}^{p} \phi_k(\tau)\mathbf{u}_i[n]. \tag{3.35}$$

For simplicity of notation, in the following, we drop the indices n and assume a normalized time step $h = 1$ over the unit interval $\tau \in [0,1]$. It is well known that the Green kernel of the operator

$$\mathcal{L} = \frac{\mathrm{d}}{\mathrm{d}\tau} - \mathbf{A}, \tag{3.36}$$

is given by ($\Theta(\tau)$ denotes the Heaviside unit step function)

$$\boldsymbol{K}_{\mathbf{A}}(\tau,\sigma) = \Theta(\tau - \sigma)e^{\mathbf{A}(\tau-\sigma)} \tag{3.37}$$

For an initial condition $\mathbf{x}_0$, the state $\boldsymbol{x}$ is obtained by convolution with the kernel $\boldsymbol{x} = \mathcal{L}^{-1}(\delta_0\mathbf{x}_0 + \mathbf{B}u) = \int \boldsymbol{K}_{\mathbf{A}}(\tau,\sigma)\left(\delta_0(\sigma)\mathbf{x}_0 + \mathbf{B}u(\sigma)\right)\mathrm{d}\sigma$. It yields the basis representation

$$\boldsymbol{x}(\tau) = \sum_{i=0}^{p} \boldsymbol{\varphi}_i(\tau)\mathbf{x}_i \qquad \text{where} \qquad \mathbf{x}_i = \mathbf{B}\mathbf{u}_i \quad i > 0. \tag{3.38}$$

The basis functions are defined by $\boldsymbol{\varphi}_i := \Psi_i[\mathbf{A},\Phi]$, $i = 0,\dots,p$ where the generator of exponential basis functions Ψ, parametrized by the matrix $\mathbf{A}$ and basis Φ, is defined as follows

$$\Psi_i[\mathbf{A},\Phi](\tau) := \begin{cases} \displaystyle\int_0^1 \boldsymbol{K}_{\mathbf{A}}(\tau,\sigma)\delta_0(\sigma)\,\mathrm{d}\sigma = \exp(\mathbf{A}\tau) & i = 0, \\ \displaystyle\int_0^1 \boldsymbol{K}_{\mathbf{A}}(\tau,\sigma)\phi_i(\sigma)\,\mathrm{d}\sigma & i = 1,\dots,p. \end{cases} \tag{3.39}$$

11. i.e. if we need to resample a signal in a (quasi)-bandlimited sense: for audition via a soundcard or for communication with digital audio processing chains inside of a Digital Audio Workstation.

Looking at the output equations (3.34b), we find that the output signal $\boldsymbol{y}$ belongs to the space spanned by the union of input and exponential basis $\{\boldsymbol{\varphi}_k(\tau)\} \cup \{\mathbf{I} \otimes \phi_k(\tau)\}$

$$y(\tau) = \mathbf{C}\left(\sum_{i=0}^{p} \boldsymbol{\varphi}_i(\tau)\mathbf{x}_i\right) + \mathbf{D}\left(\sum_{i=1}^{p} \phi_i(\tau)\mathbf{u}_i\right). \tag{3.40}$$

By sampling the functions for $\tau = 1$, we obtain the discrete state-space filtering scheme

$$\mathbf{x}_0[n+1] = \sum_{i=0}^{p} \boldsymbol{\varphi}_i(1)\mathbf{x}_i[n], \quad \text{where} \quad \mathbf{x}_i[n] = \mathbf{B}\mathbf{u}_i[n] \quad \text{for } i > 1, \tag{3.41a}$$

$$y[n+1] = \mathbf{C}\mathbf{x}_0[n+1] + \mathbf{D}\left(\sum_{i=1}^{p} \phi_i(1)\mathbf{u}_i[n]\right). \tag{3.41b}$$

Exact representation of the state $\boldsymbol{x}(\tau)$ over the basis $\{\boldsymbol{\varphi}_0, \dots \boldsymbol{\varphi}_p\}$ relies on the ability to have computable formulae for functions $\boldsymbol{\varphi}$. In the following, we consider the case of a polynomial input space, for which we provide exact integration results.

Polynomial input spaces

We consider piecewise polynomial inputs, locally represented by polynomials $\boldsymbol{u}(\tau) \in \mathbb{P}^{p-1}(\Omega, \mathbb{C}^m)$ of maximal degree $p-1$ over the unit time interval $\Omega = (0, 1)$. In numerical applications, signals will often be represented using orthogonal polynomials. However in the following, the use of the monomial basis $\boldsymbol{M}$ leads to simpler formulae (see appendix D.1 for a detailed derivation)

$$\boldsymbol{M} = \{M_k\}_{k=1}^{p} \qquad \text{where} \qquad M_k(\tau) := \frac{\tau^{(k-1)}}{(k-1)!}. \tag{3.42}$$

In this section, the basis functions $\{\boldsymbol{\varphi}_k\}$ [12] (see figure 3.9) are generated from $\boldsymbol{M}$ using (3.39). They are defined by the convolution (see [MVL78, CI01, MVL03] to compute $\exp(\mathbf{A}\tau)$)

$$\boldsymbol{\varphi}_k(\mathbf{A};\tau) := \Psi_k[\mathbf{A}, \boldsymbol{M}](\tau) = \begin{cases} \exp(\mathbf{A}\tau), & k = 0, \\ \displaystyle\int_0^{\tau} \exp(\mathbf{A}(\tau-\sigma))\frac{\sigma^{(k-1)}}{(k-1)!}\,\mathrm{d}\sigma, & k > 0. \end{cases} \tag{3.43}$$

If $\mathbf{A} = \mathbf{0}$, the operator $\mathcal{L}$ reduces to an integrator, it is then immediate that

$$\boldsymbol{\varphi}_k(\mathbf{A};\tau) = \mathbf{I}\frac{\tau^k}{k!} = \mathbf{I}M_{k+1}(\tau). \tag{3.44}$$

If $\mathbf{A}$ is invertible, the following recurrence relation can be used for practical computations

$$\boldsymbol{\varphi}_{k+1}(\mathbf{A};\tau) = \mathbf{A}^{-1}\left(\boldsymbol{\varphi}_k(\mathbf{A};\tau) - \boldsymbol{\varphi}_k(\mathbf{0},\tau)\right). \tag{3.45}$$

By reccurence, we also have the explicit representation

$$\boldsymbol{\varphi}_k(\mathbf{A};\tau) = \mathbf{A}^{-k}\left(\exp(\mathbf{A}\tau) - \mathbf{I}\sum_{i=0}^{k-1}\frac{\tau^i}{i!}\right). \tag{3.46}$$

12. We have used the same notation for the so-called φ-functions that have an important role in the literature on exponential integrators [HO10]. Note however that here we are not only interested in discrete time-stepping, but also on all the continuous-time values between time-stepping instants. This will be important in the resampling application example.

Reorganising terms, we note the following interpretation for functions $\boldsymbol{\varphi}$: the term $\mathbf{A}^k\boldsymbol{\varphi}(\mathbf{A};\tau)$ is the remainder of the Taylor series expansion of $\exp(\mathbf{A}\tau)$ truncated after k terms

$$\exp(\mathbf{A}\tau) = \sum_{i=0}^{k-1} \frac{(\mathbf{A}\tau)^i}{i!} + \mathbf{A}^k\boldsymbol{\varphi}_k(\mathbf{A};\tau).$$

Remark 3.5 (lower incomplete gamma function)**.** φ-functions are closely related to the lower incomplete gamma function

$$\gamma(\kappa,\tau) = \int_0^\tau \sigma^{\kappa-1}e^{-\sigma}\,\mathrm{d}\sigma.$$

Indeed with $\mathbf{A} = -1$ and $\kappa = k$, we have $\varphi_k(-1;\tau) = \gamma(k,\tau)$

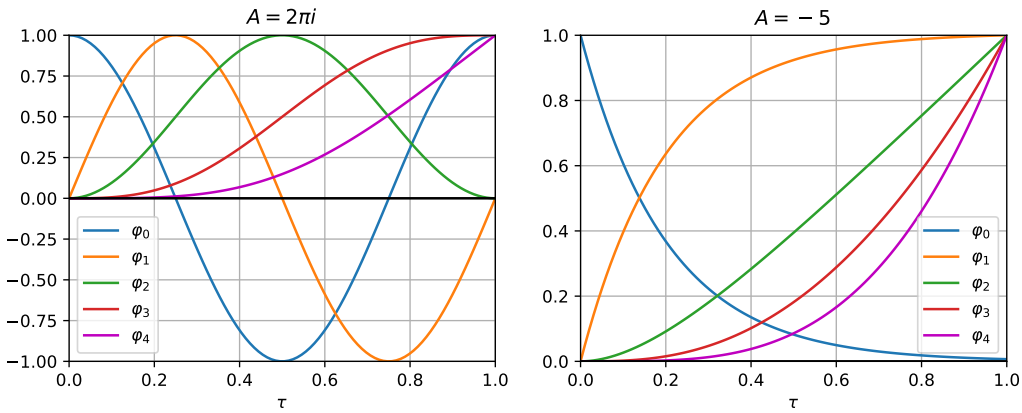


Figure 3.9 – Normalized filtered polynomial φ-functions for $k \in \{0\dots4\}$ for a complex pole $\mathbf{A} = 2\pi i$ (left plot) and a real pole $\mathbf{A} = -5$ (right plot) over the unit interval $\tau \in [0,1]$. The left plot only shows the real part of each function. blue: impulse response φ_0, orange: step response φ_1, green: ramp response φ_2, red: quadratic ramp response φ_3, magenta: cubic ramp response φ_4.

Implementation

Remark 3.6 (Diagonalised state-space and parallelisation)**.** To avoid using matrix-valued function and forming the matrix exponential, for diagonalizable matrices $\mathbf{A}$, it is advantageous to use the eigenvalue decomposition $\mathbf{A} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{-1}$ with eigenvalues $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1,\dots,\lambda_n)$. We define the complex variable change $\boldsymbol{z}(t) := \mathbf{U}^{-1}\boldsymbol{x}(t)$ to obtain the diagonalized state space system

$$\begin{aligned} \dot{\boldsymbol{z}}(t) &= \boldsymbol{\Lambda}\boldsymbol{z}(t) + \hat{\mathbf{B}}u(t), & \hat{\mathbf{B}} &= \mathbf{U}^{-1}\mathbf{B}, \\ y(t) &= \hat{\mathbf{C}}\boldsymbol{z}(t) + \mathbf{D}u(t), & \hat{\mathbf{C}} &= \mathbf{C}\mathbf{U}. \end{aligned}$$

The LTI state-space filter implementation can then be parallelised using scalar complex-valued φ-functions and the output space belongs to $\mathrm{span}\left\{\varphi_k(\lambda_i,\tau)\right\}_{k,i} \cup \left\{\phi_k(\tau)\right\}_k$.

Examples

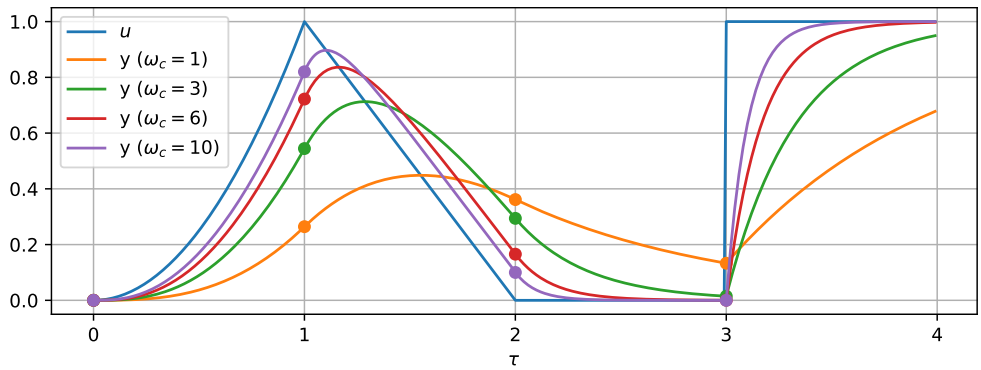


Figure 3.10 – Exact piecewise continuous-time output of a first order low-pass filter for a time sequence of local polynomials $\{\tau^2, 1-\tau, 0, 1\}$ and several values of $\omega_c \in \{1, 3, 6, 10\}$.

Example 3.2 (First order lowpass filter)**.** We consider a first order lowpass filter with the following state-space and Laplace transfer function representations for a cutoff pulsation $\omega_c \in \mathbb{R}^+$

$$\begin{cases} \dot{x}(t) = \omega_c(u(t) - x(t)) \\ y(t) = x(t) \end{cases} \quad \overset{\text{Laplace transform}}{\Longleftrightarrow} \quad Y(s) = H\left(\frac{s}{\omega_c}\right) U(s) \quad \text{where} \quad H(s) = \frac{1}{1+s}.$$

The filter is driven by a piecewise polynomial input signal $u(t)$. It is defined by the sequence of local polynomials (on the left) with corresponding monomial coefficients (on the right) by

$$\{u_n(\tau)\} = \{\tau^2, 1-\tau, 0, 1\}, \qquad \Longleftrightarrow \qquad \mathbf{u} = \left\{ \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \right\} \begin{matrix} \tau^0 \\ \tau^1. \\ \tau^2 \end{matrix}$$

The input and output signals are shown on Figure 3.10.

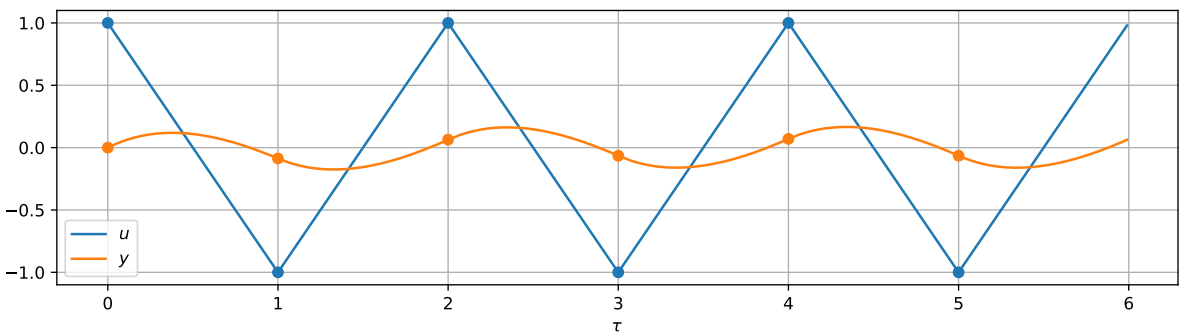


Figure 3.11 – Exact piecewise continuous-time response $y(t)$ of a third order Butterworth filter with cutoff pulsation $\omega_c = \pi$ to a triangle input signal $u(t)$ at the Nyquist frequency.

Example 3.3 (Triangle signal at the Nyquist frequency)**.** To illustrate the non-bandlimited representation capacity of piecewise polynomials, and the effectiveness of the continuous-time filtering scheme, consider a non-bandlimited triangular signal $u(t)$ oscillating at the Nyquist frequency, which is shown on figure 3.11. It is locally represented over each time step by

$$\left\{u_n(\tau)\right\} = \left\{(-1)^n(2\tau - 1)\right\}, \quad \forall n \in \mathbb{N} \qquad \Longleftrightarrow \qquad \mathbf{u} = \left\{\begin{bmatrix} -1 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 \\ -2 \end{bmatrix}, \dots\right\} \begin{matrix} \tau^0 \\ \tau^1 \end{matrix}.$$

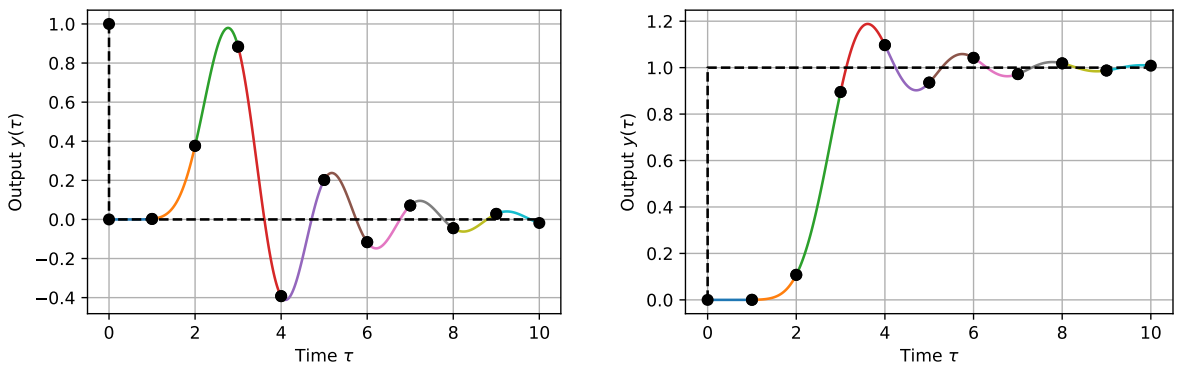
We filter this signal by a third order Butterworth [But30] filter $H\left(\frac{s}{\omega_c}\right)$ whose cutoff is set to the Nyquist pulsation $\omega_c = \pi$. The normalized Laplace transfer function prototype $H(s)$ is separated in partial fractions

$$H(s) = \frac{1}{(s^2 + s + 1)(s + 1)} = \frac{c_1}{s - \lambda_1} + \frac{c_2}{s - \lambda_2} + \frac{c_3}{s - \lambda_3}, \tag{3.47}$$

and realized in complex canonical diagonal form by the state-space system

$$\begin{cases} \dot{\boldsymbol{x}}(t) = \mathbf{A}\boldsymbol{x}(t) + \mathbf{B}u(t) \\ y(t) = \mathbf{C}\boldsymbol{x}(t) \end{cases} \qquad \mathbf{A} = \begin{bmatrix} \lambda_1 & & \\ & \lambda_2 & \\ & & \lambda_3 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} c_1 & c_2 & c_3 \end{bmatrix}, \tag{3.48}$$

with poles $\lambda_1 = \frac{-1-i\sqrt{3}}{2}$, $\lambda_2 = \frac{-1+i\sqrt{3}}{2}$, $\lambda_3 = -1$ and coefficients $c_1 = \frac{-3+i\sqrt{3}}{6}$, $c_2 = \frac{-3-i\sqrt{3}}{6}$, $c_3 = 1$. The continuous-time response of the filter is shown on figure 3.11. The values at the sampling instants are shown as black dots. To show that the method generalizes easily to any order using the same approach (and that we can easily use dirac deltas distributions as inputs), the exact impulse and step responses of an order 12 Butterworth filter are shown on figure 3.12 (Note the higher group delay which is due to the higher order of the causal minimum phase Butterworth filter).



(a) Impulse response **(b)** Step response

Figure 3.12 – Exact piecewise continuous-time impulse and step responses of an order 12 Butterworth filter. Inputs are plotted in dashed black, piecewise output segments with colours.

3.3.2 Approximation of (broken) piecewise polynomials on B-spline spaces

In the simulation methods developed in chapters 4 and 5, the time-continuous-output $y(t)$ will often be defined as a (possibly broken) piecewise polynomial. Furthermore, in section 3.2, we have considered input reconstruction in quasi-bandlimited B-spline spaces (with continuity order m). It is natural in this context to look for the dual process: finding a B-spline approximation $\tilde{y}(t)$ having the same continuity order m (or a higher continuity order if smoothing is seeked) and a rate of innovation equal to the output virtual ADC sampling rate (block E in figure 3.2). Furthermore, for implementation purposes, we want such an approximation be both local and causal.

Theory of operation The outline of the idea (shown on figure 3.13) is the following:

- Since B-splines of degree m, are piecewise polynomials with finite temporal support, there exists an invertible matrix operator $\mathbf{L}$ of dimenson $m+1$ converting from the restriction of any B-splines over the interval $[n, n+1]$ to its Legendre coefficients (see figure 3.14).

- Conversely, for each polynomial $y_n(t)$ on $\Omega_n = [n, n+1]$, the inverse operator $\mathbf{L}^{-1}$ yields a smooth extension operator: the resulting B-spline $\tilde{y}_n(t)$ (with extended temporal support) is such that its restriction to Ω_n yields the same polynomials, i.e. $\tilde{y}_n(t)|_{\Omega_n} = y_n(t)$.

- Note that each polynomial $y_n(t)$ yields a *different* local B-spline extension $\tilde{y}_n(t)$: we have an overcomplete representation with $m+1$ candidate coefficients $\tilde{\mathbf{y}}_n[k]$ for each B-spline basis function $\beta_m(t-k)$. To obtain a unique output $\tilde{y}(t)$, we need a strategy for the fusion of coefficients. It is then natural to think of weighted Overlap Add (which is a very common tool in signal processing based on the Short Time Fourier Transform).

- From frame theory [Chr16], we know that the combination of multiple bases using barycentric weights w_k (summing to one) constitute a frame. Furthermore, since the choice of (positive) weights is free, a natural idea is to use a weighting scheme proportional to the *area of influence* of each B-spline $\beta_m(t-k)$ (see (3.49)) on the interval $[0, 1]$ (see figure 3.15).

An example of B-spline projection from L^2 signals is shown in figures 3.17 and 3.16. A similar idea called Bezier projection for NURBS [13] and T-splines in the context of Isogeometric Analysis [HCB05] was proposed in reference [TSE+15].

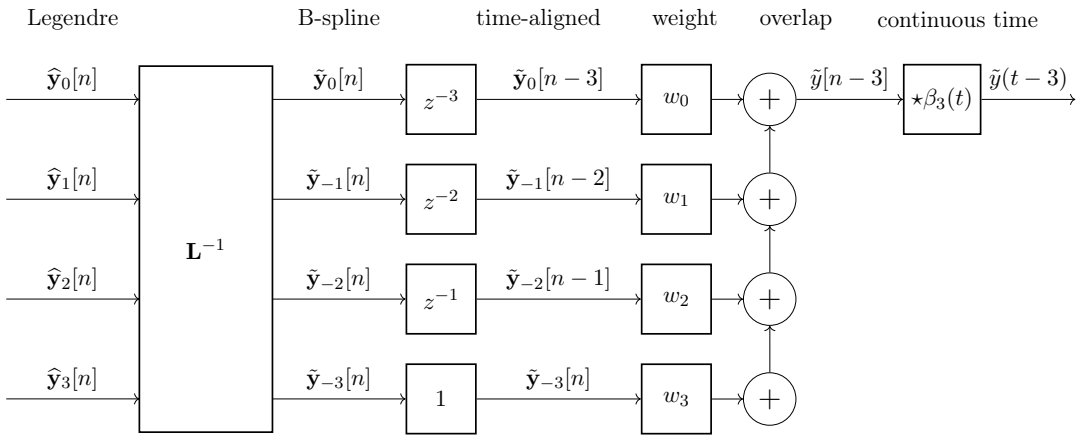


Figure 3.13 – (vADC) Block diagram of causal Legendre to cubic B-spline projection filterbank.

13. Non Uniform Rational B-splines

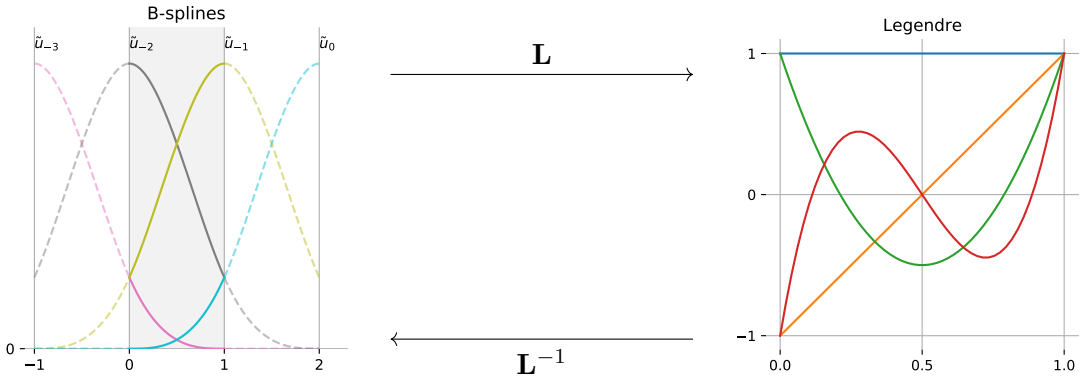


Figure 3.14 – Conversion of local cubic B-splines to Legendre polynomials using operator $\mathbf{L}$. Dually, for each function $u(t)$ defined over the Legendre polynomials on $[0,1]$, there is a smooth B-spline extension with coefficients $\{\tilde{u}[-1],\ldots,\tilde{u}[2]\}$ induced by $\mathbf{L}^{-1}$.

m	$\beta^m(t)$	$\mathbf{L}$	$\mathbf{L}^{-1}$
0	$\mathbf{1}_{[0,1]}(t)$	$[1]$	$[1]$
1	$(t)_+ - 2(t-1)_+ + (t-2)_+$	$\begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ -\frac{\sqrt{3}}{6} & \frac{\sqrt{3}}{6} \end{bmatrix}$	$\begin{bmatrix} 1 & -\sqrt{3} \\ 1 & \sqrt{3} \end{bmatrix}$
2	$\frac{(t)_+^2 - 3(t-1)_+^2 + 3(t-2)_+^2 + (t-3)_+^2}{2!}$	$\begin{bmatrix} \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \\ -\frac{\sqrt{3}}{12} & 0 & \frac{\sqrt{3}}{12} \\ \frac{\sqrt{5}}{60} & -\frac{\sqrt{5}}{30} & \frac{\sqrt{5}}{60} \end{bmatrix}$	$\begin{bmatrix} 1 & -2\sqrt{3} & 4\sqrt{5} \\ 1 & 0 & -2\sqrt{5} \\ 1 & 2\sqrt{3} & 4\sqrt{5} \end{bmatrix}$
3	$\sum_{i=0}^{4}(-1)^i\binom{4}{i}\frac{(t-i)_+^3}{3!}$,	$\begin{bmatrix} \frac{1}{24} & \frac{11}{24} & \frac{11}{24} & \frac{1}{24} \\ -\frac{\sqrt{3}}{40} & -\frac{11\sqrt{3}}{120} & \frac{11\sqrt{3}}{120} & \frac{\sqrt{3}}{40} \\ \frac{\sqrt{5}}{120} & -\frac{\sqrt{5}}{120} & -\frac{\sqrt{5}}{120} & \frac{\sqrt{5}}{120} \\ -\frac{\sqrt{7}}{840} & \frac{\sqrt{7}}{280} & -\frac{\sqrt{7}}{280} & \frac{\sqrt{7}}{840} \end{bmatrix}$	$\begin{bmatrix} 1 & -3\sqrt{3} & 11\sqrt{5} & -33\sqrt{7} \\ 1 & -\sqrt{3} & -\sqrt{5} & 9\sqrt{7} \\ 1 & \sqrt{3} & -\sqrt{5} & -9\sqrt{7} \\ 1 & 3\sqrt{3} & 11\sqrt{5} & 33\sqrt{7} \end{bmatrix}$

Table 3.1 – B-spline to Legendre conversion operators $\mathbf{L}$ and $\mathbf{L}^{-1}$. The weights $\{w_k\}$ correspond to the first row of operator $\mathbf{L}$ (i.e. projection of β_k^m on the first Legendre polynomial $P_0 = 1$).

Causal B-splines In order to align polynomials with the integer grid, here we use the causal definition of B-splines [14] as the m-fold convolution (see figure 3.4)

$$\beta^m(t) := \beta^0(t) \star \ldots \star \beta^0(t) = \sum_{i=0}^{m+1} (-1)^i \binom{m+1}{i} \frac{(t-i)_+^m}{m!}, \quad \text{where} \quad \beta^0(t) = \mathbf{1}_{[0,1]}(t). \tag{3.49}$$

We define the spline space $\mathbb{S}_m := \operatorname{span}\left\{\beta^m(t-k)\right\}_{k\in\mathbb{Z}} \subset L^2(\mathbb{R})$.

Local polynomial space Denote $\beta_k^m(t)$ the restriction to the unit interval $\Omega = [0,1]$ of the B-spline $\beta^m(t+k)$, i.e. $\beta_k^m(t) = \beta^m(t+k)\big|_\Omega$, so that the restriction to Ω of a function $u(t)$ from the spline space $\mathbb{S}_m$ is locally represented in the polynomial space $\mathbb{P}^m(\Omega)$ by

$$u(t)\Big|_\Omega = \sum_{k=0}^{m} \beta_k^m(t)\tilde{u}_{-k} = |\,\boldsymbol{\beta}\rangle\,\tilde{\mathbf{u}}. \tag{3.50}$$

where $|\,\boldsymbol{\beta}\rangle = \left[|\,\beta_k^m\rangle\right]_{k=0}^m$ denotes the B-spline synthesis operator and $\tilde{\mathbf{u}} = \left([\tilde{u}_{-k}]_{k=0}^m\right)^\mathsf{T}$ are the B-spline coefficients corresponding to times $\left\{\frac{m+1}{2} - k\right\}_{k=0}^m$ (see figure 3.14).

B-spline to Legendre representation We are interested in the Legendre representation.

$$u(t)\Big|_\Omega = \sum_{k=0}^{m} P_k(t)\widehat{u}_k = |\,\boldsymbol{P}\rangle\,\widehat{\mathbf{u}}. \tag{3.51}$$

where $|\,\boldsymbol{P}\rangle = \left[|\,P_0\rangle, \ldots, |\,P_m\rangle\right]$ denotes the Legendre synthesis operator (Legendre polynomials are defined in appendix C.4 p.286) for the Legendre coefficients $\widehat{\mathbf{u}} = [\widehat{u}_0, \ldots, \widehat{u}_m]^\mathsf{T}$. Since both representations correspond to the same function in the polynomial space $\mathbb{P}^m(\Omega)$, there exists an invertible operator $\mathbf{L}$ such that $\widehat{\mathbf{u}} = \mathbf{L}\mathbf{u}$ given by

$$\mathbf{L} = \langle \boldsymbol{P}\,|\,\boldsymbol{\beta}\rangle\,. \tag{3.52}$$

Proof. The result follows from the relations

$$u(t)|_\Omega \overset{a}{=} |\,\boldsymbol{\beta}\rangle\,\tilde{\mathbf{u}} = |\,\boldsymbol{P}\rangle\,\widehat{\mathbf{u}} \qquad \overset{b}{\Longleftrightarrow} \qquad \underbrace{\langle \boldsymbol{P}\,|\,\boldsymbol{\beta}\rangle}_{\mathbf{L}}\,\tilde{\mathbf{u}} = \langle \boldsymbol{P}\,|\,\boldsymbol{P}\rangle\,\widehat{\mathbf{u}} \overset{c}{=} \widehat{\mathbf{u}}$$

using (a) representation of $u(t)$ in both basis, (b) left multiplication by the dual Legendre analysis operator $\langle \boldsymbol{P}\,|$, (c) orthonormality of the Legendre polynomial basis $\langle \boldsymbol{P}\,|\,\boldsymbol{P}\rangle = \mathbf{I}_p$. □

Inverse Legendre to B-spline operator Conversely for a sequence of Legendre coefficients $\left\{\widehat{\mathbf{u}}[n]\right\}_{n\in\mathbb{Z}}$, the inverse operator yields $m+1$ sequences of B-spline coefficients

$$\left\{\tilde{\mathbf{u}}[n]\right\} = \left\{\mathbf{L}^{-1}\widehat{\mathbf{u}}[n]\right\} = \begin{bmatrix} \ldots & \tilde{u}_0[n] & \ldots \\ & \vdots & \\ \ldots & \tilde{u}_{-m}[n] & \ldots \end{bmatrix}. \tag{3.53}$$

We call $\mathbf{L}^{-1}$ the B-spline extraction operator. Examples are shown on table 3.1 and figure 3.14.

14. From the spectral definitions of causal B-splines with Laplace transform $\widehat{\beta}^m(s) = (1-e^{-s})^{m+1}/s^{m+1}$. The binomial coefficients and time-shifts comes from the expansion of the finite difference operator $(1-e^{-s})^{m+1}$ while $t_+^m/m!$ comes from the inverse Laplace transform of the repeated integration operator $1/s^{m+1}$.

Weighted barycentric overlap-add Finally, the B-spline coefficients $\{\tilde{u}[n]\}_{n\in\mathbb{Z}}$ of $\tilde{u}(t)$ are obtained by combining the $m+1$ B-spline estimates using the barycentric average

$$\tilde{u}\,[n-m] = \sum_{k=0}^{m} w_k \tilde{u}_{m-k}\,[n-k]\,, \qquad \text{where} \qquad w_k = \frac{\int_0^1 \beta_k^m(t)\,\mathrm{d}t}{\sum_{k=0}^{m}\int_0^1 \beta_k^m(t)\,\mathrm{d}t}. \tag{3.54}$$

The weights w_k are chosen proportional to the intersection of their area with the unit interval.

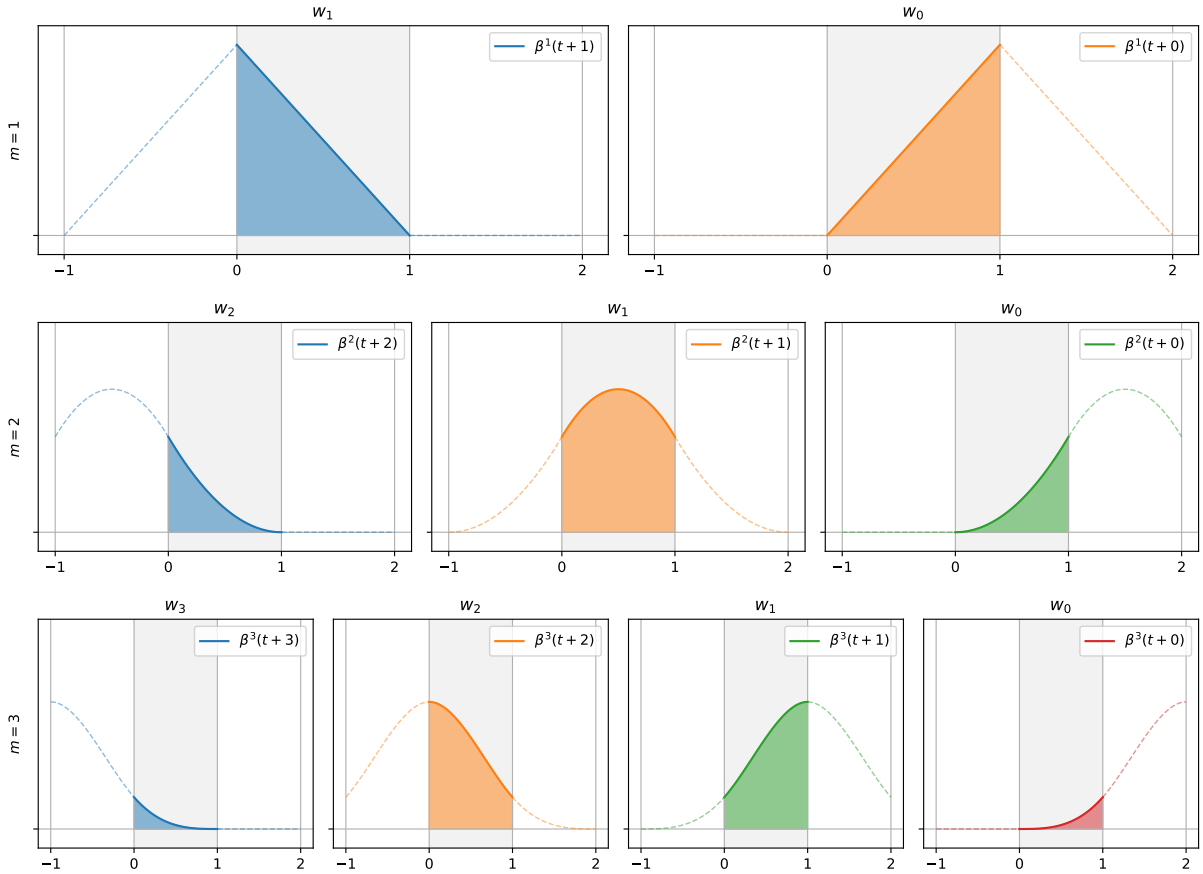


Figure 3.15 – Barycentric overlap-add weights for linear, parabolic and cubic splines.

Formalisation of the approximation operator Denote $\mathcal{Z}^\tau : u(t) \mapsto u(t+\tau), \tau \in \mathbb{R}$ the timeshift operator. Combining equations (3.49) to (3.54) according to the block diagram in figure 3.13, the analysis-synthesis process $\mathcal{Q}_m$ is defined by

$$(\mathcal{Q}_m u)(t) = \sum_{n\in\mathbb{Z}} \mathcal{Z}^{-n} \sum_{i=0}^{m} w_i \mathcal{Z}^{i-m} \sum_{j=0}^{m} \mathcal{Z}^i \beta^m(t) \mathbf{L}_{i,j}^{-1} \left\langle P_j \,\middle|\, \mathcal{Z}^n u \right\rangle. \tag{3.55}$$

Proposition 3.1. *Operator $\mathcal{Q}_m$ defined by (3.55) reproduces the spline space $\mathbb{S}_m$ up to a constant delay of size m, i.e.*

$$\mathcal{Q}_m \beta^m = \mathcal{Z}^{-m} \beta^m. \tag{3.56}$$

Proof. Substituting (a) $u = \beta^m$ in (3.55), then, using (b) the local B-splines $\beta_n^m = \beta^m(t+n)|_\Omega$ (see (3.50)), and the definition of operator $\mathbf{L}$ (3.52), (c) $\mathbf{L}^{-1}\mathbf{L} = \mathbf{I}_m$, (d) the barycentric weight property $\sum_{i=0}^m w_i = 1$ (see (3.54)), we obtain

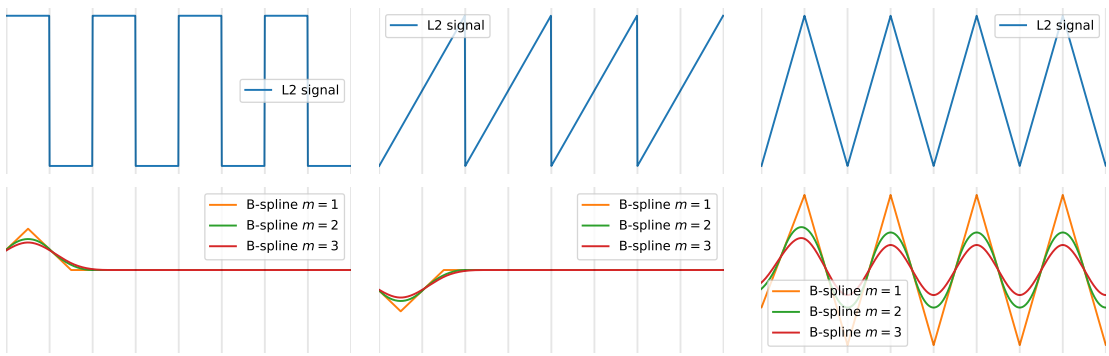
$$
\begin{aligned}
(\mathcal{Q}_m\beta^m)(t) &\overset{a}{=} \sum_{n\in\mathbb{Z}} \mathcal{Z}^{-n} \sum_{i=0}^m w_i \mathcal{Z}^{i-m} \sum_{j=0}^m \mathcal{Z}^i \beta^m(t) \mathbf{L}_{i,j}^{-1} \left\langle P_j \,\middle|\, \mathcal{Z}^n \beta^m \right\rangle \\
&\overset{b}{=} \sum_{n=0}^m \mathcal{Z}^{-n} \sum_{i=0}^m w_i \mathcal{Z}^{i-m} \mathcal{Z}^i \beta^m(t) \sum_{j=0}^m \mathbf{L}_{i,j}^{-1} \underbrace{\left\langle P_j \,\middle|\, \beta_n^m \right\rangle}_{\mathbf{L}_{jn}} \\
&\overset{c}{=} \sum_{n=0}^m \mathcal{Z}^{-n} \sum_{i=0}^m w_i \mathcal{Z}^{-m} \beta^m(t) \delta_{i,n} = \sum_{n=0}^m \mathcal{Z}^{-n} \mathcal{Z}^{-m} \beta^m(t) \left(\sum_{i=0}^m w_i \delta_{i,n} \right) \\
&\overset{d}{=} \sum_{n=0}^m \mathcal{Z}^{-n} \mathcal{Z}^{-m} \beta^m(t) \delta_{0,n} = \mathcal{Z}^{-m} \beta^m(t).
\end{aligned}
$$

By integer shift-invariance, by conclude that $\mathcal{Q}_m$ reproduces the spline space $\mathbb{S}_m$. □

Numerical experiments In order to assess the qualitative approximation properties of operator $\mathcal{Q}_m$, we perform two numerical experiments (a detailed quantitative study is left for future work).

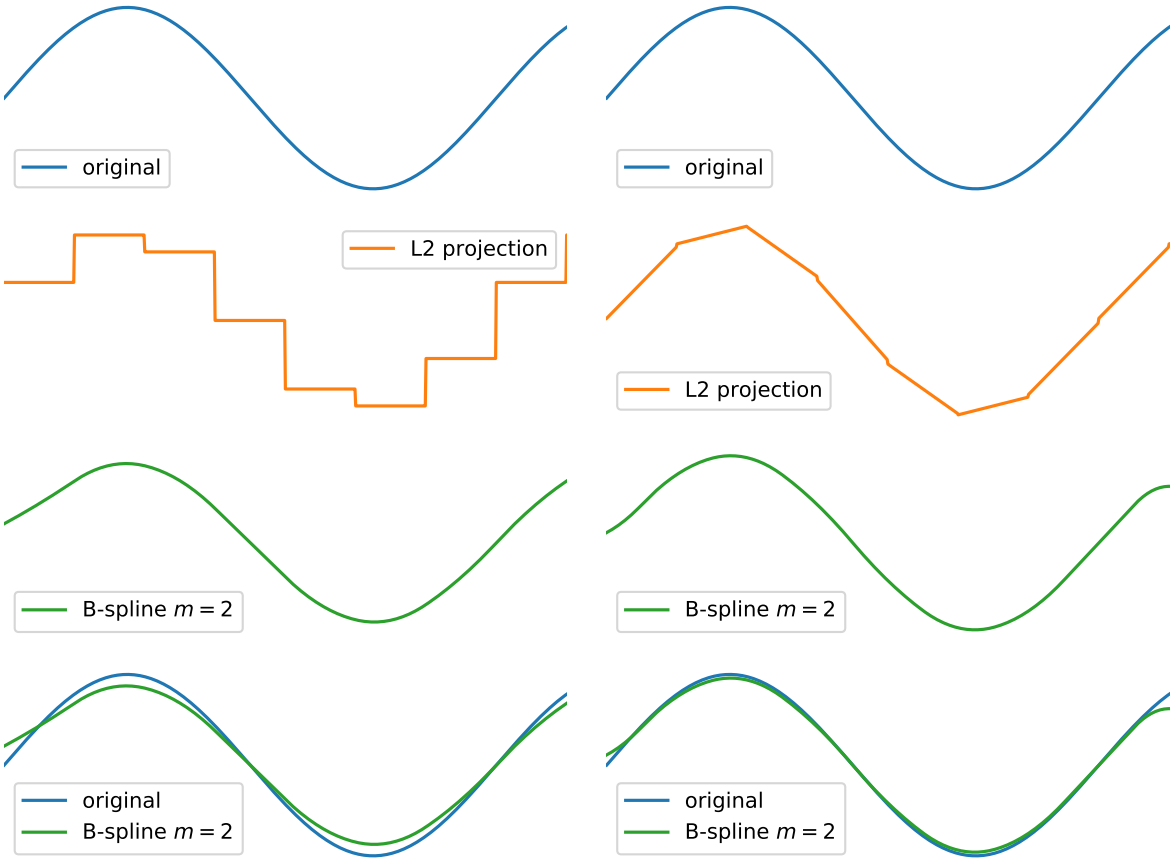
- First (figure 3.16), we approximate piecewise discontinuous square, sawtooth and triangle polynomial signals over B-spline spaces of increasing smoothness $\mathbb{S}_1, \mathbb{S}_2, \mathbb{S}_3$. We note that square and sawtooth belong to the kernel of the B-spline projector and are exactly filtered after an initial transient. The triangle is exactly reproduced by first order B-splines, but it is progressively filtered when increasing the B-spline smoothness.

- Second (figure 3.17), to anticipate signals from chapter 5, we project (first row) a (smooth, bandlimited) sinusoid over piecewise constant and piecewise affine subspaces of $L^2(\mathbb{R})$ (this yields non-bandlimited approximations, second row), then we reconstruct its $\mathcal{C}^1$ approximations over the B-spline space $\mathbb{S}_2$ using operator $\mathcal{Q}_2$ (third row). We note that even for low smoothness $m = 2$ and crude piecewise constant approximations, signals are qualitatively well recovered. Furthermore we notice the increased accuracy of the piecewise affine reconstruction (see section 3.1.3).



(a) square **(b)** saw **(c)** triangle

Figure 3.16 – B-spline aproximation of square, saw and triangle oscillations at Nyquist.



(a) Piecewise constant **(b)** Piecewise affine

Figure 3.17 – Reconstruction of a sinusoid in the B-spline space $\mathbb{S}_2$ after (discontinuous) piecewise constant (left column) and piecewise affine (right column) L^2 approximations . Signals have been time-aligned to compensate for the causal delay of size m (see (3.56)). Edge differences are due to to the fact that the smoothing operator operates on truncated signals with finite support (L^2 signals are implicitly extended to zero outside of the approximation window, while $\mathbb{S}_2$ signals in orange are smoothly extended according to the temporal support of the B-spline β^2).

3.4 Application: "virtual analog" resampler

As an illustration of the virtual analog toolchain, a real-time, variable rate, "virtual analog", resampler (fig. 3.18) has been implemented in UVI Falcon software [UVI21]. It is constituted of:

- First order B-spline DAC with sampling rate f_s (see section 3.2) to convert discrete-time signal to continous-time (and a second optional one with virtual sampling rate f'_s),
- a continuous-time anti-aliasing Butterworth lowpass filter (see figure 3.12) with cutoff frequency $f'_c < f'_s/2$ to approximately limit the bandwidth of the signal to f'_c,
- a variable rate sampler with virtual sampling rate $f'_s < f_s$, to downsample the signal at a lower sampling rate (with the effect of periodising the spectrum above $f'_s/2$),
- a second exact high-order continuous-time anti-image Butterworth lowpass filter with cutoff frequency $f_c < f_s/2$ to bandlimit the signal to $f_s/2$ (voluntarily [15] keeping spectral images between f'_c and f_c).
- a fixed sampler to resample the signal back to the original sampling rate f_s.

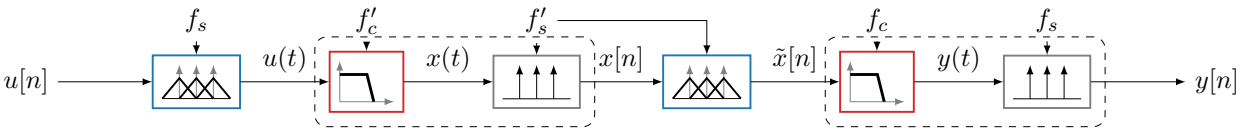


Figure 3.18 – (Virtual Analog resampler) block-diagram.

It can be interpreted as a cascade of two multi-rate polyphase resamplers [VL88], except that using virtual continuous-time signal processing, we have an infinite number of 'phases' between sampling instants. Blocks in dashed line on figure 3.18 corresponds to approximate projection (see (3.16)) on spaces of bandlimited signals with respective bandwidths $f'_s/2$ and $f_s/2$.

Spectral periodisation about the virtual sampling rate $f'_s = 4$ kHz (and its multiples) is illustrated on figure 3.19. The quasi-band-limiting effect of the two Butterworth filters is clearly visible: we still observe some aliasing in the crossover region about the virtual Nyquist frequency $f'_s/2 = 2$ kHz (an its images at 6 KHz, 10 kHz, etc) but it is maintained below -84 dB.

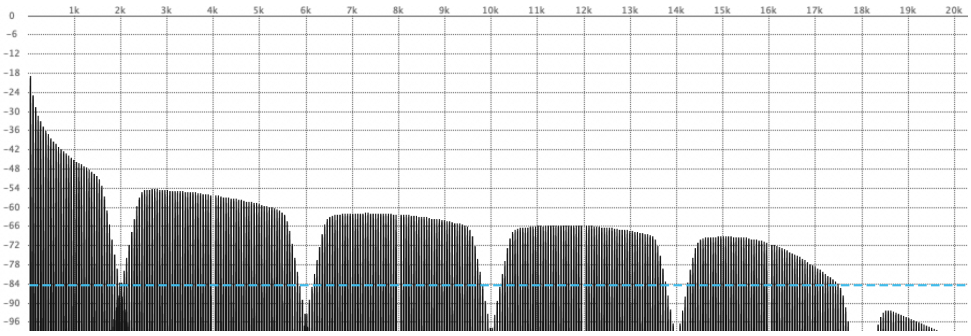


Figure 3.19 – (Virtual Analog resampler) spectrum periodisation.

15. The whole purpose of sampling rate reduction audio effects is to *keep the spectrum periodisation artefacts* of the *virtual sampling rate* f'_s (to emulate the sound of old analog-to-digital and digital-to-analog converters) and at the same time to *avoid spectral aliasing artefacts* that are linked to the current simulation sampling rate f_s.

Conclusion

In this chapter, we have reviewed the foundations of the non-bandlimited signal representation framework used in this thesis. Instead, we use sequences of time-frames having a finite number of parameters per time frame, i.e. signals with a *finite rate of innovation*. The tools of generalized sampling theory allows consistent analysis-resynthesis of such non-bandlimited signals. Extended bandwidth is useful to resolve the extended spectrum of nonlinear systems (for example a sawtooth signal is not bandlimited in the Shannon-Nyquist sense, but its rate of innovation is finite and proportional to its frequency), Having minimal disjoint temporal supports is also a critical ingredient to obtain causal numerical integration schemes.

We have revisited the topic of continuous-time input reconstruction in B-spline spaces from discrete signal samples. B-spline signal processing theory is now well established, yet discrete B-spline pre-filters are sometimes omitted so that B-splines can be wrongly described as being too smooth. In our context, causality is perhaps the most limiting factor, For that purpose, we have seen that shifted linear interpolation is a causal and cost-effective way to improve the frequency response of traditional linear interpolation at the expense of phase linearity.

We have also considered exact causal continuous-time ARMA filtering of piecewise defined signals. This strategy allows to use the vast literature on analog filter design tools (e.g. Butterworth, Chebyshev, Elliptic, etc) for the realisation of the continuous-time anti-aliasing stage. As an alternate approach: we consider the approximation of piecewise (discontinuous) polynomials on smooth B-splines spaces. Indeed, it is known that in the limit of infinite smooothness, the interpolating kernel in B-splines spaces converge to the sinc kernel of band-limited signal spaces. The ARMA approach has the advantage of being very general and causal with steep anti-aliasing filters for a relatively low filter order. The price to pay is the lack of phase linearity and lack of idempotence of the bandlimiting operator. Alternatively, B-spline output approximation works as a projector (with delay), so we have causality, phase linearity (idempotence with delay). We face the same kind of design tradeoffs as is usual in the choice between Finite Impulse Response (FIR) and Infinite Impulse Response (IIR) filters for a given application. Note that signal reconstruction in quasi-bandlimited spaces such as the ones generated by Hammerich pulses [Ham07, KZ17] looks promising for audio use but is left for future work.

Chapter 4

Power-balanced Adaptive collocation

If an idea works once it's a trick. If it works twice it's a technique. If it works three times it's a method.

Unknown source

Contents

4.1 Satisfying the power-balance using adaptive collocation 108
4.2 Method A: adaptive collocation 109
4.3 Method B: symmetric adaptive collocation 111
4.4 Increasing regularity: SPAC methods 113

In this chapter we restrict our investigation to input-state-output PHS systems, defined in definition 1.22, of the form

$$\dot{\boldsymbol{x}} = (\mathbf{J} - \mathbf{R})\nabla H(\boldsymbol{x}) + \mathbf{G}\boldsymbol{u}, \tag{4.1a}$$

$$\boldsymbol{y} = \mathbf{G}^{\mathsf{T}}\nabla H(\boldsymbol{x}). \tag{4.1b}$$

Although the approach is general, we focus the study on linear systems.

For a unit interval $\Omega = (0, 1)$, and time step h, properties $\mathbf{P}_0$–$\mathbf{P}_3$ (see p.79) are expressed by

$\mathbf{P}_0$ the trajectory is locally approximated on $[t_n, t_n + h]$ by a polynomial $\boldsymbol{X}_n$ on Ω such that

$$\boldsymbol{X}_n(\tau) \approx \boldsymbol{x}(t_n + h\tau), \quad \forall \tau \in \Omega,$$

$\mathbf{P}_1$ derivatives match on frame boundaries up to a controllable continuity order $k \geq 0$

$$\boldsymbol{X}_n^{(m)}(0) = \boldsymbol{X}_{n-1}^{(m)}(1), \quad \forall m \in \{0, \dots, k\},$$

$\mathbf{P}_2$ the local truncation error ϵ has a controllable accuracy order s:

$$\epsilon(h) = \boldsymbol{x}(t_0 + h) - \boldsymbol{X}(1) = \mathcal{O}(h^{s+1}),$$

$\mathbf{P}_3$ the trajectory $\boldsymbol{X}_n$ satisfies the power balance over each frame

$$H(\boldsymbol{X}_n(1)) - H(\boldsymbol{X}_n(0)) = -h\int_0^1 \nabla H(\boldsymbol{X}_n(\tau))^{\mathsf{T}}\mathbf{R}\nabla H(\boldsymbol{X}_n(\tau))\,\mathrm{d}\tau + h\int_0^1 \boldsymbol{y}(\tau)^{\mathsf{T}}\boldsymbol{u}(\tau)\,\mathrm{d}\tau.$$

Outline Our strategy, is detailed in section 4.1. It uses (adaptive) collocation (see [HLW06]) to satisfy all of the above properties: the vector field and its derivatives is exactly satisfied at fixed *collocation instants* to obtain both accuracy and smoothness. Additional collocation points are used and adaptively optimised for each time frame to satisfy the power balance.

In section 4.2 we propose a first instance of the approach. We obtain the adaptive Euler method whose solutions are $\mathcal{C}^0$-regular. We study its accuracy order $s \in \{1, 2\}$, its stability function, and the existence domain of power balanced solutions. This shows that with this formulation, there exists a maximal dissipation rate above which power-balanced solutions do not exist anymore. Numerical simulations show that despite the lower accuracy order, thanks to the power-balance, qualitative aspects such as orbit and dissipation rate are improved compared to the mid-point method.

In section 4.3, in order to improve the deficiencies (low accuracy and regularity orders) of the first method, we add symmetry and smoothness. This leads to a $\mathcal{C}^1$-regular method. We study its numerical properties showing that it is unconditionaly A-stable with an accuracy order $s \in \{4, 6\}$ (for linear systems). The existence domain of power-balanced solutions is also improved.

Finally in section 4.4, we generalise the approach to any number of derivatives and collocation points with the definition of (Symmetric [1]) Power-balanced Adaptive collocation methods (PAC and SPAC). We use symbolic computer algebra to automate the study of their stability function, accuracy order, leading error term and maximal dissipation rate. The existence domain of power-balanced solutions is also shown in the complex plane. The domains are different but closely reminiscent of the theory of order stars [WHN78].

4.1 Satisfying the power-balance using adaptive collocation

For a local trajectory $\boldsymbol{X}(\tau)$, $\tau \in [0, 1]$, we define the local vector field

$$\boldsymbol{f}_h(\boldsymbol{X}) := h\left((\mathbf{J} - \mathbf{R})\nabla H(\boldsymbol{X}) + \mathbf{G}\boldsymbol{u}\right), \tag{4.2}$$

and the vector field approximation error operator

$$\boldsymbol{E}(\boldsymbol{X}) := \dot{\boldsymbol{X}} - \boldsymbol{f}_h(\boldsymbol{X}). \tag{4.3}$$

Finally we introduce the power balance error, defined by the functional

$$\rho(\boldsymbol{X}) := \left\langle \nabla H(\boldsymbol{X}) \,\middle|\, \boldsymbol{E}(\boldsymbol{X}) \right\rangle = \int_0^1 \nabla H(\boldsymbol{X}(\tau))^\mathsf{T} \boldsymbol{E}(\boldsymbol{X}(\tau))\, \mathrm{d}\tau. \tag{4.4}$$

> **Remark 4.1** (Power balance orthogonality condition)**.** In the absence of external ports, the power balance $\rho(\boldsymbol{X}) = 0$ can be interpreted as an orthogonality condition between the vector field approximation error $\boldsymbol{E}(\boldsymbol{X})$ and the gradient of the Hamiltonian $\nabla H(\boldsymbol{X})$.

Our first strategy, inspired by Runge-Kutta collocation methods [HLW06] is to use a first set of fixed collocation points C, and a second set of variables ones $\tilde{C}$ such that

$$\boldsymbol{E}(\boldsymbol{X}(c_i)) = \mathbf{0}, \quad \forall c_i \in C \cup \tilde{C}.$$

The set C is used to achieve numerical accuracy (and continuity). The set $\tilde{C}$ is devoted to satisfy the power balance: the variable parameters $\tilde{c}_i \in [0, 1]$ are optimised so that

$$\rho(\boldsymbol{X}) = 0.$$

To obtain a practical numerical method, existence and uniqueness of power-balanced solutions must be investigated. To study this problem, we propose a family of (Symmetric) Power-balanced Adaptive collocation methods respectively called PAC and SPAC and study three instances of increasing complexity. We restrict the analysis to autonomous linear ODEs, for which we provide stability functions, accuracy analysis and analytical bounds on the existence of power-balanced solutions (based on the maximal dissipation rate).

1. i.e. such that the method is invariant under time reversal and has an even accuracy order.

4.2 Method A: adaptive collocation

We first consider the minimal requirements to satisfy properties **P0** – **P3**,

Method 4.1. The one-point Power-balanced Adaptive collocation method PAC(1) is defined implicitly by the following constraints:

P0. (Model) The trajectory $\boldsymbol{X}_\alpha(\tau) \in \mathbb{P}^1$ is an affine polynomial with 2 degrees of freedom

$$\boldsymbol{X}_\alpha(\tau) = \mathbf{X}_0 + \tau\delta\mathbf{X}_\alpha. \tag{4.5a}$$

P1. ($\mathcal{C}^0$-Continuity) The trajectory satisfies the initial condition

$$\boldsymbol{X}_\alpha(0) = \boldsymbol{X}_0 = \mathbf{x}_0 \in \mathbb{R}^n. \tag{4.5b}$$

P2. (Accuracy order $s \geq 1$) The vector field is satisfied for the collocation point $\alpha \in [0,1]$

$$\dot{\boldsymbol{X}}_\alpha(\alpha) = \delta\boldsymbol{X}_\alpha = \boldsymbol{f}_h(\boldsymbol{X}_\alpha(\alpha)) \in \mathbb{R}^n. \tag{4.5c}$$

P3. (Power balance) The PB is satisfied if there exists an optimal value $\alpha^\star$ satisfying

$$\alpha^\star \in \left\{\alpha \in [0,1] \mid \rho(\boldsymbol{X}_\alpha) = 0\right\} \neq \emptyset. \tag{4.5d}$$

The method is completed by the time-stepping map $\Phi_h : \mathbf{x}_0 \mapsto \mathbf{x}_1 := \boldsymbol{X}(1)$.

Method 4.1 defines a nonlinear problem with $n+1$ parameters to solve with respect to $(\delta\mathbf{X}_\alpha, \alpha)$. A difficulty is that the parameter α appears recursively in $\delta\mathbf{X}_\alpha$. To study this problem, we consider the autonomous linear case.

Autonomous Linear analysis

Let $H(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\mathsf{T}\mathbf{Q}\mathbf{x} = \frac{1}{2}\|\mathbf{x}\|^2_{\mathbf{Q}}$, with $\mathbf{Q} = \mathbf{Q}^\mathsf{T} \succeq 0$ be a quadratic Hamiltonian, $\mathbf{A} = h(\mathbf{J}-\mathbf{R})\mathbf{Q}$ and $\mathbf{G} = 0$. We rewrite (4.1a) as the autonomous ODE

$$\dot{\boldsymbol{X}} = \mathbf{A}\boldsymbol{X}, \quad \boldsymbol{X}(0) = \mathbf{x}_0. \tag{4.6}$$

Solving the collocation constraint (4.5c): $\delta\mathbf{X}_\alpha = \mathbf{A}(\mathbf{x}_0 + \alpha\delta\mathbf{X}_\alpha)$ leads to $\delta\mathbf{X}_\alpha = (\mathbf{I} - \alpha\mathbf{A})^{-1}\mathbf{A}\mathbf{x}_0$. Substitution in (4.5a) yields the following family of candidate solutions parametrised by α

$$\mathbf{X}_\alpha(\tau) = \left(\mathbf{I} + \tau(\mathbf{I} - \alpha\mathbf{A})^{-1}\mathbf{A}\right)\mathbf{x}_0 = (\mathbf{I} - \alpha\mathbf{A})^{-1}\left(\mathbf{I} + (\tau - \alpha)\mathbf{A}\right). \tag{4.7}$$

Evaluating $\mathbf{x}_1 = \mathbf{X}_\alpha(\tau = 1)$, yields the time stepping scheme $\mathbf{x}_1 = R_\alpha(\mathbf{A})\mathbf{x}_0$, where the time-stepping operator is

$$R_\alpha(\mathbf{A}) = (\mathbf{I} - \alpha\mathbf{A})^{-1}(\mathbf{I} + (1-\alpha)\mathbf{A}). \tag{4.8}$$

Substituting the matrix $\mathbf{A}$ by a complex pole $\lambda \in \mathbb{C}$, we obtain

Property 4.1 (stability fonction)**.** For the Dahlquist test equation, $\dot{x} = \lambda x$, $\lambda \in \mathbb{C}$, approximated using method 4.1, we obtain $x_1 = R_\alpha(\lambda)x_0$, the stability function (see def. B.4 p.276) is thus

$$R_\alpha(\lambda) = \frac{1 + (1-\alpha)\lambda}{1 - \alpha\lambda}. \tag{4.9}$$

Remark 4.2. This classical result corresponds to the stability function of extended Euler methods. Using Taylor series expansion, the time-stepping approximation error is given by

$$\epsilon(\lambda) = \exp(\lambda) - R_\alpha(\lambda) = \lambda^2 \left(\frac{1}{2} - \alpha\right) + \mathcal{O}(\lambda^3). \tag{4.10}$$

- The method has accuracy order $s \geq 1$, $\forall \alpha \in [0, 1]$. It reaches accuracy order $s = 2$ for $\alpha = \frac{1}{2}$ and $R_{1/2}(\lambda)$ is the Padé approximant of $\exp(\lambda)$ of order $(1, 1)$.
- If $\alpha \geq \frac{1}{2}$, then the method is A-stable: $\left|R_\alpha(\lambda)\right| \leq 1$ for $\Re(\lambda) \leq 0$, (see def. B.5 p.276). If $\alpha = \frac{1}{2}$, then the method is conservative: $\left|R_\alpha(\lambda)\right| = 1$ for all $\lambda \in i\mathbb{R}$.

The following result shows that, even in the linear dissipative case, there is a maximal dissipation rate above which it is not possible to satisfy the power balance (see figure 4.1).

Property 4.2 (bounded power balanced domain)**.** Let $\lambda = -\sigma$, $\sigma \in \mathbb{R}^+$. If $\sigma \in [0, \sqrt{3}]$, then the power balance (4.5d) is satisfied for the optimal collocation point

$$\alpha^\star = \frac{(\sigma - 1) + \sqrt{1 - \frac{\sigma^2}{3}}}{2\sigma} \in [0, 1]. \tag{4.11}$$

Proof. Substituting equation (4.7) in the power balance functional (4.4), and integrating symbolically (see appendix E.1 p.309) we btain

$$0 = \rho(X_\alpha) = \int_0^1 X_\alpha(\tau) \left(\dot{X}_\alpha(\tau) - f(X_\alpha(\tau))\right) \mathrm{d}\tau = \left(\alpha^2\sigma + \alpha\,(1 - \sigma) + \frac{2\sigma - 3}{6}\right) \frac{\sigma^2 x_0^2}{(1 + \sigma\alpha)^2}.$$

This quadratic equation has a unique real branch in $[0, 1]$ given by (4.11) for $|\sigma| \in [0, \sqrt{3}]$. □

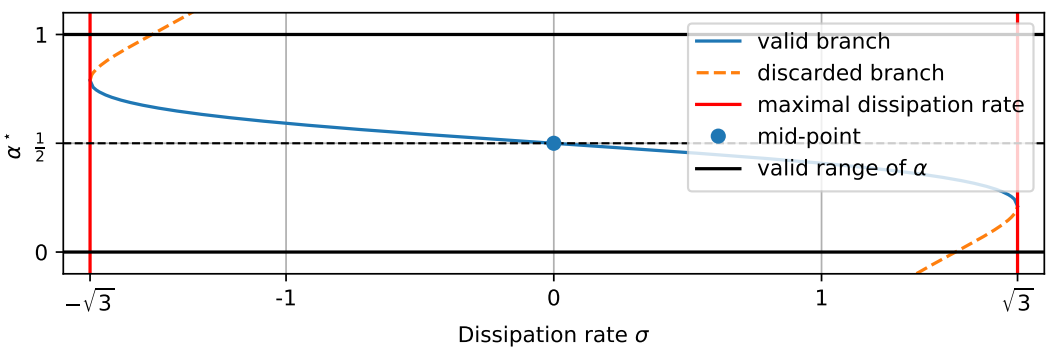


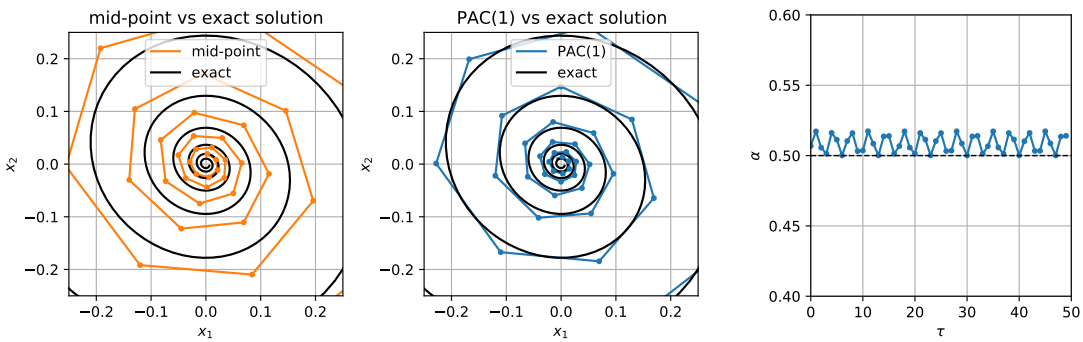
Figure 4.1 – (PAC(1)) Optimal parameter $\alpha^\star$ as a function of the dissipation rate σ. Note that in the absence of dissipation ($\sigma = 0$), the optimal parameter ($\alpha^\star = 1/2$) corresponds to the mid-point method. When the dissipation rate increases ($\sigma > 0$), the method goes towards the Forward Euler method ($\alpha^\star < 0.5$). Conversely if the pole is unstable ($\sigma < 0$), the method goes towards the Backward Euler scheme ($\alpha^\star > 0.5$). For $|\sigma| > \sqrt{3}$, it has no real solution.

It may seem that the method relies on the stability margin provided by the dissipation rate to solve the power-balance constraint. To demonstrate that solving the power balance does not require the use of artificial numerical damping (or emphasis), a symmetric power balanced adaptive collocation method that is always A-stable is presented in section 4.3.

Example 4.1 (Damped Oscillator)**.** Consider a damped oscillator with normalised ($h = 1$) pulsation ω and dissipation rate σ (for example a parallel RLC circuit) with vector field

$$\boldsymbol{f}(\boldsymbol{X}) = \begin{bmatrix} -\sigma & -\omega \\ \omega & 0 \end{bmatrix} \boldsymbol{X}.$$

A numerical simulation of this system is shown on Figure 4.2. The mid-point method ($\alpha = \frac{1}{2}$), which is second order accurate, is compared to the PAC(1) (adaptive Euler), which is only first-order accurate (in general). Despite the lower local numerical accuracy, we remark that two qualitative aspects of the exact solutions have been *improved thanks to the power balance*: the dissipation rate and the distance to the exact dissipative orbit.



(a) orbits **(b)** α values

Figure 4.2 – (PAC(1) - Damped RLC) Mid-point method vs PAC(1). Despite the lower accuracy order of PAC(1), we remark that the orbit and dissipation (in blue on the right) are improved compared to the mid-point method (in orange on the left).

4.3 Method B: symmetric adaptive collocation

To generalize to $\mathcal{C}^1$ solutions and to obtain a symmetric A-stable method, we introduce

Method 4.2 (SPAC(2))**. P0**. (Model) The trajectory is a polynomial $\boldsymbol{X}_\alpha \in \mathbb{P}^4(\Omega, \mathbb{R}^n)$,

P1-2. ($\mathcal{C}^1$-continuity) $\boldsymbol{X}_\alpha$ satisfies an initial condition and collocation of the vector field on the boundary of the interval $\partial\Omega = \{0, 1\}$

$$\boldsymbol{X}_\alpha(0) = \mathbf{X}_0, \qquad \dot{\boldsymbol{X}}_\alpha(0) = \boldsymbol{f}_h(\boldsymbol{X}_\alpha(0)), \qquad \dot{\boldsymbol{X}}_\alpha(1) = \boldsymbol{f}_h(\boldsymbol{X}_\alpha(1)), \tag{4.12a}$$

P2-3. (Power balance) the vector field is satisfied on symmetric adaptive collocation points

$$\dot{\boldsymbol{X}}_\alpha(\alpha) = \boldsymbol{f}_h(\boldsymbol{X}_\alpha(\alpha)), \qquad \dot{\boldsymbol{X}}_\alpha(1-\alpha) = \boldsymbol{f}_h(\boldsymbol{X}_\alpha(1-\alpha)), \tag{4.12b}$$

The PB is satisfied if there exists an $\alpha^\star$ such that

$$\alpha^\star \in \left\{\alpha \in [0,1] \mid \rho(\boldsymbol{X}_\alpha) = 0\right\}. \tag{4.12c}$$

We study the behaviour of method 4.2 and its validity domain. In the linear case, we have the following property

Property 4.3 (stability fonction)**.** For the Dahlquist test equation, $\dot{x} = \lambda x$, $\lambda \in \mathbb{C}$, approximated using method 4.2, the time stepping map is $x_1 = R_\beta(\lambda)x_0$ with the stability function

$$R_\beta(\lambda) = \frac{1 + \frac{\lambda}{2} + (1-\beta)\frac{\lambda^2}{12} + \beta\frac{\lambda^3}{24}}{1 - \frac{\lambda}{2} + (1-\beta)\frac{\lambda^2}{12} - \beta\frac{\lambda^3}{24}}, \quad \text{and} \quad \beta = \alpha(1-\alpha). \tag{4.13}$$

Proof. The proof is omitted. The result can be derived using CAS such as in E.1 p.309. □

Remark 4.3. The method is A-stable for all values of β. Using Taylor series expansion, the approximation error is

$$\epsilon(\lambda) = \exp(\lambda) - R_\beta(\lambda) = (5\beta - 1)\left(\frac{\lambda^5 + \lambda^6}{720}\right) + \mathcal{O}\left(z^7\right). \tag{4.14}$$

By consequence the method

- has (linear) accuracy order $s \geq 4$, $\forall \beta \in [0, \frac{1}{4}]$,
- reaches accuracy order $s = 6$ for $\beta = \frac{1}{5}$ (i.e. $\alpha = \frac{1}{2} \pm \frac{\sqrt{5}}{10}$). In this case, $R_\beta(\lambda)$ corresponds to the Padé approximation of $\exp(\lambda)$ of order $(3,3)$ (see also D.7 p.297).

For a purely dissipative test equation, we also have the following result

Property 4.4. Let $\lambda = -\sigma$, $\sigma > 0$. The power balance $\rho(X_\beta) = 0$ has a unique solution

$$\beta = \frac{2520 + \sigma^2(9\sigma^2 - 84) - \sqrt{3\left(-\sigma^8 + 112\sigma^6 + 2116800\right)}}{42\sigma^2\left(\sigma^2 - 10\right)} \in \left[0, \frac{1}{4}\right], \tag{4.15}$$

subject to $\sigma \in [0, \sigma_{max})$ where $\sigma_{max} \approx 10.651$ (see fig 4.3).

Proof. As in property 4.2, solving the power balance $\rho(X_\beta) = 0$ yields a quadratic equation

$$a\beta^2 + b\beta + c = 0, \tag{4.16}$$

with $a = 21\sigma^2(\sigma^2 - 10)$, $b = -9\sigma^4 + 84\sigma^2 - 2520$, $c = 504 - 12\sigma^2 + \sigma^4$. It admits a unique solution for $\beta \in \left[0, \frac{1}{4}\right]$ (i.e. $\alpha \in [0, 1/2]$) which is given by (4.15). □

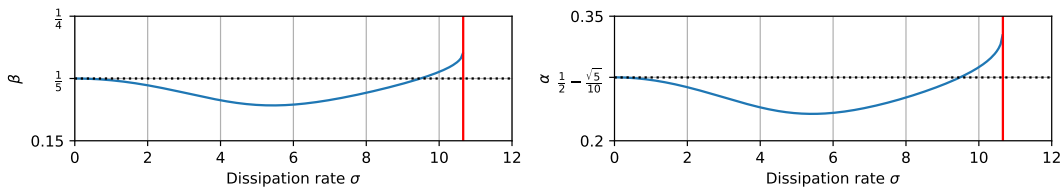


Figure 4.3 – Optimal value of β (and α) as a function of the dissipation rate σ.

4.4 Increasing regularity: SPAC methods

In order to increase the regularity and accuracy orders, we combine the previously presented approach with multi-derivative Hermite-Obreshkoff collocation methods [HNW93, Nør74, Obr40]. We summarize and extend the previous methods with the following definition.

Method 4.3 ((S)PAC)**.** Denote k the $\mathcal{C}^k$-regularity order and $d = 2k+1$ (resp. $d = 2k+2$) the polynomial degree. Denote $t = t_0 + h\tau$, $\tau \in \Omega = [0,1]$ the time and $\mathcal{D} = \frac{1}{h}\frac{\mathrm{d}}{\mathrm{d}\tau}$ the time derivative ($\equiv \frac{\mathrm{d}}{\mathrm{d}t}$). The (Symmetric) Power-balanced Adaptive collocation method of regularity k, in short *(S)PAC(k)*, is defined by

- **P0** (Model) $\boldsymbol{X}_\alpha \in \mathbb{P}^d(\Omega, \mathbb{R}^n)$ is a polynomial over the interval Ω,
- **P1,P2** ($\mathcal{C}^k$-continuity). $\boldsymbol{X}_\alpha$ satisfies an initial condition and multi-derivative collocation of the vector field on the boundaries of the interval $\partial\Omega = \{0,1\}$.

$$\boldsymbol{X}_\alpha(0) = \mathbf{x}_0, \tag{4.17a}$$

$$\left(\mathcal{D}^m \boldsymbol{X}_\alpha\right)(c) = \left(\mathcal{D}^{m-1}\boldsymbol{f}\left(\boldsymbol{X}_\alpha(\tau)\right)\right)(c), \qquad \forall c \in \partial\Omega, \quad \forall m \in \{1, \dots k\}. \tag{4.17b}$$

- **P3** (power balance) The vector field is satisfied over the set $\tilde{C} = \{\alpha\}$, $\alpha \in D = (0,1)$ for PAC (resp. $\tilde{C} = \{\alpha, 1-\alpha\}$, $\alpha \in D = (0,1/2)$ for SPAC) such that

$$\mathcal{D}\boldsymbol{X}_\alpha(c) = \boldsymbol{f}\left(\boldsymbol{X}_\alpha(c)\right), \qquad \forall c \in \tilde{C}. \tag{4.17c}$$

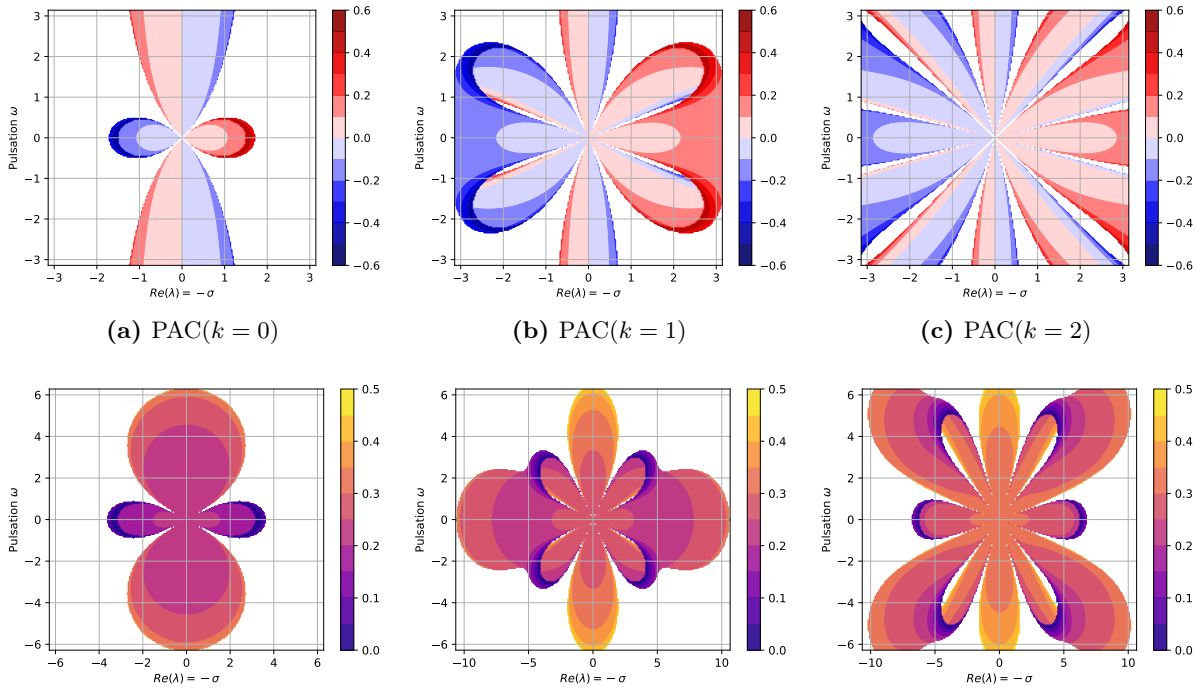
The power balance is satisfied if there exists an $\alpha^\star$ such that

$$\alpha^\star \in \left\{\alpha \in D \mid \rho(\boldsymbol{X}_\alpha) = 0\right\}. \tag{4.17d}$$

Automating proofs using CAS, as in E.1 p.309, we obtain the properties in table 4.1.

Method	Stability function $R(\lambda)$	Leading error	s	$\sigma_{\max}$
PAC(0)	$\dfrac{1+(1-\alpha)\lambda}{1-\alpha\lambda}$	$-(2\alpha-1)\dfrac{\lambda^2}{2}$	1 to 2	≈ 1.73
PAC(1)	$\dfrac{1+\frac{\lambda(4-2\alpha)}{6}+\frac{\lambda^2(1-\alpha)}{6}}{1-\frac{\lambda(2\alpha+2)}{6}+\frac{\alpha\lambda^2}{6}}$	$(2\alpha-1)\dfrac{\lambda^4}{72}$	3 to 4	≈ 6.66
PAC(2)	$\dfrac{1+\frac{\lambda(72-24\alpha)}{120}+\frac{\lambda^2(18-12\alpha)}{120}+\frac{\lambda^3(2-2\alpha)}{120}}{1-\frac{\lambda(24\alpha+48)}{120}+\frac{\lambda^2(12\alpha+6)}{120}-\frac{\alpha\lambda^3}{60}}$	$-(2\alpha-1)\dfrac{\lambda^6}{7200}$	5 to 6	≈ 4.96
SPAC(0)	$\dfrac{1+\frac{\lambda}{2}+\frac{\lambda^2\beta}{2}}{1-\frac{\lambda}{2}+\frac{\lambda^2\beta}{2}}$	$(6\beta-1)\left(\dfrac{\lambda^3+\lambda^4}{12}\right)$	2 to 4	≈ 3.66
SPAC(1)	$\dfrac{1+\frac{\lambda}{2}+(1-\beta)\frac{\lambda^2}{12}+\beta\frac{\lambda^3}{24}}{1-\frac{\lambda}{2}+(1-\beta)\frac{\lambda^2}{12}-\beta\frac{\lambda^3}{24}}$	$(5\beta-1)\left(\dfrac{\lambda^5+\lambda^6}{720}\right)$	4 or 6	≈ 10.65
SPAC(2)	$\dfrac{1+\frac{\lambda}{2}+\frac{\lambda^2(24\beta+72)}{720}+\frac{\lambda^3(12\beta+6)}{720}+\frac{\lambda^4(2\beta)}{720}}{1-\frac{\lambda}{2}+\frac{\lambda^2(24\beta+72)}{720}-\frac{\lambda^3(12\beta+6)}{720}+\frac{\lambda^4(2\beta)}{720}}$	$(14\beta-3)\left(\dfrac{\lambda^7+\lambda^8}{302400}\right)$	6 to 8	≈ 6.38

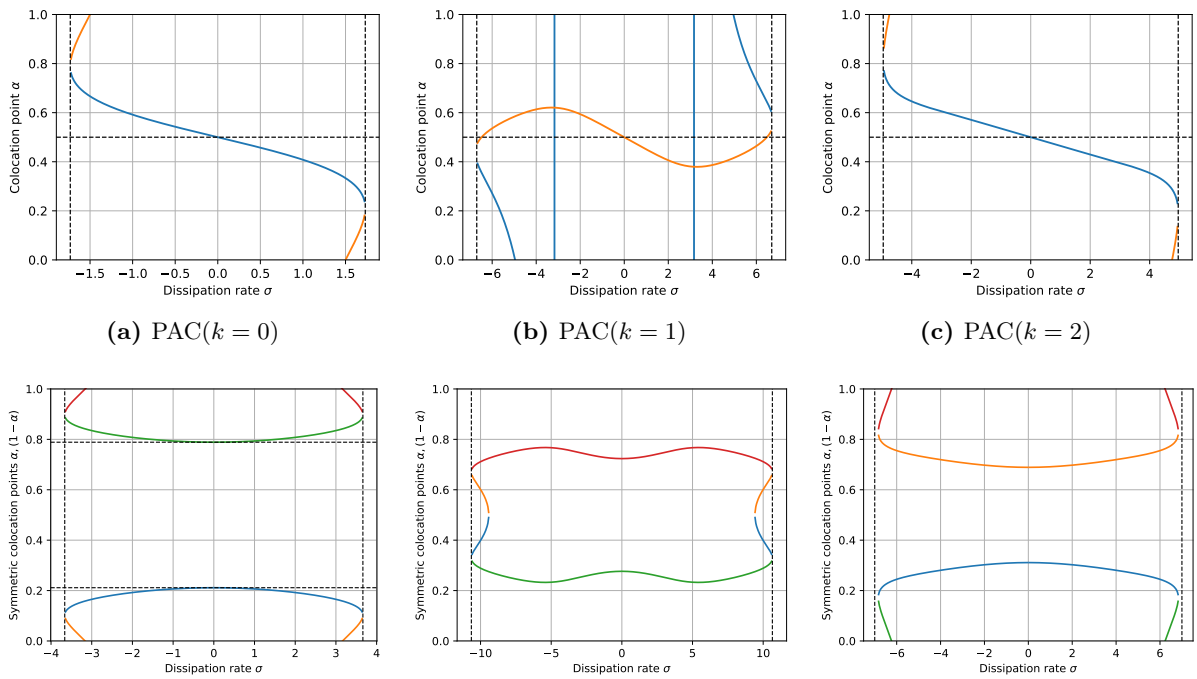
Table 4.1 – (SPAC methods) Linear properties. Remind that $\beta = \alpha(1-\alpha)$, s denotes accuracy order and the leading error is the first nonzero term in Taylor series expansion of the error.



(a) PAC($k = 0$) **(b)** PAC($k = 1$) **(c)** PAC($k = 2$)

(d) SPAC($k = 0$) **(e)** SPAC($k = 1$) **(f)** SPAC($k = 2$)

Figure 4.4 – ((S)PAC) Power Balanced regions satisfying $\rho(\boldsymbol{X}_\alpha) = 0$ and $\alpha \in [0, 1]$ for the Dahlquist test equation $\dot{x} = \lambda x$, $\lambda = -\sigma + j\omega$. Contour plots are shaded according to α for SPAC and $\alpha - \frac{1}{2}$ for PAC. Note that multiple solution branches are overlaid using transparency.



(a) PAC($k = 0$) **(b)** PAC($k = 1$) **(c)** PAC($k = 2$)

(d) SPAC($k = 0$) **(e)** SPAC($k = 1$) **(f)** SPAC($k = 2$)

Figure 4.5 – ((S)PAC) Optimal values of power-balanced collocation point(s) α as a function of the dissipation rate σ ($\lambda = -\sigma$). Note that multiple solutions are plotted with different colours .

Discussion about (S)PAC methods Optimal power balanced adaptive collocation points $\alpha^\star$ are shown in figure 4.5 according to dissipation rate for real poles. Power balanced regions for complex poles are shown on figure 4.4 for (S)PAC methods for regularity orders $k = 0, 1, 2$. Analysing table 4.1 and figures 4.4, we make the following observations:

- Power-balanced regions are closely related but different from the theory of order stars [2] which was introduced in [WHN78] to study the stability of numerical methods.

- We remark on figure 4.4 that for both PAC and SPAC, increasing the regularity k increases the surface of power balanced regions. However, we also notice in table 4.1 that the maximal dissipation rate shrinks for $k = 2$. A tradeoff seems to operate between the maximal dissipation rate and the total area of the power-balanced region.

- In Table 4.1, for PAC methods, the leading error term vanishes for the roots of the Legendre and Lobatto polynomials [3]. These polynomials play an important role in the construction of Gauss–Legendre and Gauss–Lobatto Runge–Kutta methods (see [HLW06]).

- In the absence of dissipation, for both PAC and SPAC methods, the power-balance yields balanced A-stable Padé approximations of the exponential with optimal accuracy order s.

- In the presence of dissipation, PAC methods may use locally expansive stability functions (blue zones in figures 4.4a-c). The method relaxes accuracy order to satisfy the power balance. Nevertheless, one can see on figure 4.2 that the orbit of the power-balanced approximation (with lower accuracy order $s = 1$) is closer to the orbit of the true solution when compared to the orbit of an A-stable approximation having higher accuracy order ($s = 2$) and the same number of collocation points.

- The previous observation indicates that the local truncation error, commonly used to measure accuracy order, is only one metric among others based on a discrete simulation grid: minimising specific *continuous-time error metrics* (such as the power-balance functional $\rho(\boldsymbol{X})$ (eq. (4.4)) or the vector-field approximation error $\boldsymbol{E}(\boldsymbol{X})$ (eq. (4.3))), can be beneficial to capture or improve important features of the dynamics (such as energy-conservation, orbit shapes or dissipation rate).

- SPAC methods are all symmetric, A-stable, time-reversible and of even accuracy order (independently of the dissipation rate) by symmetry of their collocation points.

A Python code example to produce results of table 4.1 and graphics of figures 4.4 and 4.5 is shown in listing E.1 p.309.

2. Order star theory uses the regions $A = \left\{\lambda \in \mathbb{C} \,|\, |S(\lambda) > 1|\right\}$ with $S(\lambda) = R(\lambda)/\exp(\lambda)$ to study stability, In (S)PAC we use power balanced regions of the complex plane for which $\rho(\boldsymbol{X}_\alpha) = 0$ can be satisfied.

3. For all PAC methods, the leading error term in table 4.1 vanishes for $\alpha = 1/2$, the root of the Legendre polynomial $P_1(\alpha) = 2\alpha - 1$. Expanding $\beta = \alpha(1 - \alpha)$, we obtain the Legendre polynomial $P_2(\alpha) = 6\alpha^2 - 6\alpha + 1$ for SPAC(0), and the Jacobi/Lobatto polynomial $L_2(\alpha) = 5\alpha^2 - 5\alpha + 1$ for SPAC(1).

Conclusion

We have proposed a first family of (Symmetric) Power balanced Adaptive collocation methods called (S)PAC that can satisfy the regularity, accuracy and power balance requirements **P1**, **P2**, **P3**. This approach has the following advantages and drawbacks

Advantages

- arbitrary high regularity order k (**P1**) and accuracy order s (**P2**) can be easily obtained by increasing the order of derivatives and the number of collocation points,
- the continuous-time power balance is exactly satisfied (when a solution exists),
- dissipation rate and orbits are more accurately tracked thanks to the power-balance **P3**.

Drawbacks Unfortunately, we also note the following important drawbacks

- the existence domain of power-balanced solutions is bounded by a maximal the dissipation rate (for real poles) or more generally by the power-balanced regions of figure 4.4 for complex poles,
- an implicit nonlinear equation has to be solved for each time-step (even for linear systems),
- polynomial parameters are implicitly defined with respect to the adaptive parameter α which does not appear linearly in the equations. This makes estimation of parameters in the case of nonlinear vector field $\boldsymbol{f}(\boldsymbol{x})$ a difficult problem [4] for which existence/uniqueness/convergence conditions remains an open subject.

To overcome these problems, we abandon the collocation approach and adopt a different strategy: we interpret the power-balance as an orthogonality condition $\rho(\boldsymbol{X}) = \left\langle \nabla H(\boldsymbol{X}) \,\middle|\, \boldsymbol{E}(\boldsymbol{X}) \right\rangle = 0$ between the vector field error $\boldsymbol{E}(\boldsymbol{X}) = \dot{\boldsymbol{X}} - \boldsymbol{f}(\boldsymbol{X})$ and the Hamiltonian gradient $\nabla H(\boldsymbol{X})$. This interpretation leads us to methods based on continuous-time functional projection [5] in chapter chapter 5.

4. A strategy consist in alternating between the fixed-point (or Newton) estimation of the implicit polynomial $\boldsymbol{X}_\alpha$ (through collocation of the vector field for a given α), and optimisation of the collocation point α for a given polynomial $\boldsymbol{X}_\alpha$. Joint optimisation of both parameters has also been investigated but is not detailed here.

5. Note that, interpolation of the vector field in collocation methods can also be interpreted as continuous-time projection in Sobolev spaces (rather than discrete inner product spaces). However, continuous-time projection alone is not sufficient to preserve the power balance. This viewpoint is detailed in chapter 5, particularly in section 5.2.7.

Chapter 5

Power-balanced projection methods

Spectral methods are like Swiss watch. They work beautifully, but a little dust in the gear stops them entirely.

Philip L. Roe, quoted by J. P. Boyd, SIAM Rev., 46(2004)

Contents

5.1 Regular Projection Methods for pH-ODE and pH-DAE 119
5.1.1 Power-balance condition 119
5.1.2 Examples of projector design 121
5.1.3 RPM for pH-ODE 122
5.1.4 RPM for pH-DAE 123
5.2 Analysis of RPM for pH-ODE 125
5.2.1 Reminder on Runge-Kutta methods 125
5.2.2 Reformulation of RPM as Continuous-Stage Runge-Kutta methods 125
5.2.3 Existence and uniqueness of solutions 127
5.2.4 Linear Stability function 127
5.2.5 Energy preservation (P3) 127
5.2.6 Order conditions and polynomial reproduction (P2) 128
5.2.7 Regularity (P1) 129
5.3 Analysis of RPM for pH-DAE 135
5.3.1 Accuracy and stage order for stiff ODE and DAE 135
5.3.2 Existence and uniqueness of solutions 135
5.4 Implementation choices 140
5.4.1 Closed-form projection results for nonlinear maps of affine functions 140
5.4.2 General purpose numerical quadratures 144
5.4.3 Representations, fixed-point and Newton iterations 146
5.5 Examples 147
5.5.1 Nonlinear Conservative LC 147
5.5.2 Diode Clipper 153

Introduction

This chapter presents one of the main results of this thesis: we establish a sufficient condition on projectors to obtain time-continuous power-balanced trajectories. Indeed, in chapter 4, we have seen that it is not possible to unconditionally satisfy the power balance functional (4.4) using (adaptive) collocation methods. In particular (see figures 4.4 and 4.5 p.114), the existence domain of power balanced solution is bounded: there is a maximal dissipation rate (or more generally a method-dependent maximal pole radius) above which power-balanced solutions cease to exist. Furthermore, the power-balance constraint led to numerical schemes whose parameter estimation is nonlinear in the parameters (even for linear ODE).

To avoid these problems, in this chapter, which is central in this thesis, we propose a continuous-time power-balanced functional projection approach.

The chapter is structured as follows [1]:

- In section 5.1, we define regular power balanced methods (RPM) of variable projection and regularity orders which satisfy properties **P1**, **P2**, **P3** (defined p.79). The main foundational results, which links functional L^2 projection and power balance are exposed in subsection 5.1.1, where we introduce the functional notion of projected conservative (Dirac) and dissipative structures over time-frames. Based on these results, RPM are first defined for pH-ODE in subsection 5.1.3, and for pH-DAE in subsection 5.1.4.

- In section 5.2, instead of jumping straight to implementation and simulation issues (see sections 5.4, 5.5), we provide a thorough analysis of RPM in the case of pH-ODE. This step is important to guide the choice of approximation spaces. In subsection 5.2.2, we reformulate RPM as continuous-stage Runge-Kutta methods. The goal is twofold: first to leverage the vast amount of results available for Runge-Kutta methods, second to bridge the functional projection and the Runge–Kutta viewpoints. Existence and uniqueness conditions are considered in subsection 5.2.3, stability functions in subsection 5.2.4, power balance in subsection 5.2.5, accuracy order conditions in subsection 5.2.6. Finally regularity analysis and Peano error kernels are detailed in subsection 5.2.7. A landmark of this section is that projection spaces that reproduce polynomials yield high-order accuracy.

- In section 5.3 we try to tackle the more difficult subject of pH-DAE. A short discussion on accuracy and stage-order and stiffness is provided in subsection 5.3.1. But most of the work is dedicated to establishing milestones towards *practical* existence and uniqueness conditions for RPM applied to pH-DAE by exploiting the particular structure of the equations.

- In section 5.4, we address the implementation of RPM: numerical computation of projections, choice of unknowns and implicit equation solving using Newton iteration.

- In section 5.5, we finally detail and illustrate RPM modelling and simulation on two examples [2]: a conservative pH-ODE and a dissipative pH-DAE. For both uses cases, we provide and compare several simulations at different projection and regularity orders. A close attention is also paid to energy preservation (up to machine precision), the quality/regularity of continuous-time orbits and to the anti-aliasing and generalized spectral bandwidth.

Finally, we conclude this chapter by analysing the strengths and weaknesses of RPM and compare with state of the art energy-preserving methods.

1. Application oriented readers, may skip numerical analysis sections 5.2 and 5.3, which are mostly theoretical, to jump straight to implementation in section 5.4 p.140 and the numerical simulations in section 5.5 p.147

2. Note that chapter 8 p.197 is dedicated to applications on real circuits, where the complete process (from circuit modelling to numerical simulation) is detailed with a finer level of details.

5.1 Regular Projection Methods for pH-ODE and pH-DAE

5.1.1 Power-balance condition

Motivation In chapter 4 we have seen that using collocation, it is not possible to unconditionally satisfy the power balance condition $\left\langle \nabla H(\boldsymbol{X}) \,\middle|\, \dot{\boldsymbol{X}} - \boldsymbol{f}(\boldsymbol{X}) \right\rangle = 0$ (see Equation 4.4). We propose, instead, to consider the weak ODE formulation over a subspace V of $L^2(\Omega, \mathbb{R}^n)$

$$\left\langle \boldsymbol{v} \,\middle|\, \dot{\boldsymbol{X}} - \boldsymbol{f}(\boldsymbol{X}) \right\rangle = 0, \quad \forall \boldsymbol{v} \in V.$$

Note that, if we had $\nabla H(\boldsymbol{X}) \in V$, this would imply the orthogonality $\left\langle \nabla H(\boldsymbol{X}) \,\middle|\, \dot{\boldsymbol{X}} - \boldsymbol{f}(\boldsymbol{X}) \right\rangle = 0$. Unfortunately, for $\dot{\boldsymbol{X}} \in V$, by integration and nonlinearity, the function $\nabla H(\boldsymbol{X}(\tau))$ belongs to a larger space. It needs to be projected on V *without loosing energy/passivity preservation.*

To this end, we propose the following definition and theorem that are applicable for both pH-ODE and pH-DAE (see corollaries 5.1-5.3).

Definition 5.1 (Projected structure)**.** Let $\mathbf{A} \in \mathbb{R}^{n \times n}$ be a matrix defining the structure

$$\mathcal{S} = \left\{ (\mathbf{f}, \mathbf{e}) \in \mathbb{R}^n \times \mathbb{R}^n \mid \mathbf{f} = \mathbf{A}\mathbf{e} \right\}.$$

Denote $F = L^2(\Omega, \mathbb{R}^n)$ the flow space of square integrable time signals over an open $\Omega \subset \mathbb{R}$. Denote $E = F^* \sim F$ the (dual) space of effort signals. Let $\boldsymbol{\mathcal{P}} : F \to F$ be a projector and denote $\boldsymbol{\mathcal{P}}^* : E \to E$ its adjoint for the L^2 duality pairing $\langle \cdot \,|\, \cdot \rangle$. We call the functional set

$$\mathcal{S}_{\boldsymbol{\mathcal{P}}} = \left\{ (\boldsymbol{f}, \boldsymbol{e}) \in F \times E \mid \boldsymbol{f} = \boldsymbol{\mathcal{P}}\mathbf{A}\boldsymbol{e} \right\}, \tag{5.1}$$

a *projected structure* over the time interval Ω.

We want that the projected structure $\mathcal{S}_{\boldsymbol{\mathcal{P}}}$ preserves (on average over Ω) passivity properties (in the sense of (1.53b) p.34) of the original structure $\mathcal{S}$. To this end, we propose

Theorem 5.1 (Projected passivity)**.** *Assume that the pair* $(\boldsymbol{\mathcal{P}}, \mathbf{A})$ *satisfies the condition*

$$\boldsymbol{\mathcal{P}}\mathbf{A} = \mathbf{A}\boldsymbol{\mathcal{P}}^*. \tag{5.2}$$

Then, the projection $\boldsymbol{\mathcal{P}}$ *preserves the passivity properties of* $\mathcal{S}$*: for all* $(\boldsymbol{f}, \boldsymbol{e}) \in \mathcal{S}_{\boldsymbol{\mathcal{P}}}$

$$\mathcal{S}_{\boldsymbol{\mathcal{P}}} \text{ is passive if } \mathbf{A} \preceq 0, \qquad \text{i.e.} \qquad \langle \boldsymbol{e} \,|\, \boldsymbol{f} \rangle \leq 0, \tag{5.3a}$$

$$\mathcal{S}_{\boldsymbol{\mathcal{P}}} \text{ is power-conserving if } \mathbf{A} = -\mathbf{A}^\mathsf{T}, \qquad \text{i.e.} \qquad \langle \boldsymbol{e} \,|\, \boldsymbol{f} \rangle = 0. \tag{5.3b}$$

When (5.3a) *(resp.* (5.3b)*) holds, we call* $\mathcal{S}_{\boldsymbol{\mathcal{P}}}$ *a projected dissipative (resp. Dirac) structure.*

Proof. The result follows from the sequence of relations

$$\langle \boldsymbol{e} \,|\, \boldsymbol{f} \rangle \overset{a}{=} \langle \boldsymbol{e} \,|\, \boldsymbol{\mathcal{P}}\mathbf{A} \,|\, \boldsymbol{e} \rangle \overset{b}{=} \left\langle \boldsymbol{e} \,\middle|\, \boldsymbol{\mathcal{P}}^2\mathbf{A} \,\middle|\, \boldsymbol{e} \right\rangle \overset{c}{=} \langle \boldsymbol{e} \,|\, \boldsymbol{\mathcal{P}}\mathbf{A}\boldsymbol{\mathcal{P}}^* \,|\, \boldsymbol{e} \rangle \overset{d}{=} \langle \boldsymbol{e} \,|\, \boldsymbol{\mathcal{P}}(\mathbf{J} - \mathbf{R})\boldsymbol{\mathcal{P}}^* \,|\, \boldsymbol{e} \rangle$$
$$\overset{e}{=} - \langle \boldsymbol{e} \,|\, \boldsymbol{\mathcal{P}}\mathbf{R}\boldsymbol{\mathcal{P}}^* \,|\, \boldsymbol{e} \rangle \overset{f}{\leq} 0.$$

using (a) projected flows $\boldsymbol{f} = \boldsymbol{\mathcal{P}}\mathbf{A}\boldsymbol{e}$ (5.1), (b) idempotency $\boldsymbol{\mathcal{P}}^2 = \boldsymbol{\mathcal{P}}$, (c) commutation (5.2) $\boldsymbol{\mathcal{P}}\mathbf{A} = \mathbf{A}\boldsymbol{\mathcal{P}}^*$, (d) equality $\mathbf{A} = \mathbf{J} - \mathbf{R}$ with $\mathbf{J} = \frac{1}{2}(\mathbf{A} - \mathbf{A}^\mathsf{T})$, $\mathbf{R} = -\frac{1}{2}(\mathbf{A} + \mathbf{A}^\mathsf{T})$, (e) skew-ajointness (5.3b) of $\boldsymbol{\mathcal{P}}\mathbf{J}\boldsymbol{\mathcal{P}}^*$ and (f) positive self-adjointness (5.3a) of $\boldsymbol{\mathcal{P}}\mathbf{R}\boldsymbol{\mathcal{P}}^*$. This yields $\langle \boldsymbol{e} \,|\, \boldsymbol{f} \rangle = 0$ when $\mathbf{R} = \mathbf{0}$. □

Theorem 5.1 has a number of important implications for PHS detailed below.

Corollary 5.1 (Projected Hamiltonian System)**.** *Let* $\dot{\mathbf{x}} = \mathbf{J}\nabla H(\mathbf{x})$ *be a Hamiltonian system and* $\boldsymbol{\mathcal{P}}$ *a projector such that* $(\boldsymbol{\mathcal{P}}, \mathbf{J})$ *satisfies* (5.2)*. Then, for* $\boldsymbol{X} \in H^1(\Omega, \mathbb{R}^n)$ *solution of*

$$\dot{\boldsymbol{X}} = \boldsymbol{\mathcal{P}}\mathbf{J}\nabla H(\boldsymbol{X}), \quad \boldsymbol{X}(t_0) = \mathbf{x}_0, \tag{5.4}$$

the energy is conserved on the boundaries of $\Omega = (t_0, t_1)$*, namely* $H\left(\boldsymbol{X}(t_1)\right) = H\left(\boldsymbol{X}(t_0)\right)$*.*

Proof. The result follows from $0 \overset{a}{=} \langle \boldsymbol{e} \,|\, \boldsymbol{f} \rangle \overset{b}{=} \left\langle \nabla H(\boldsymbol{X}) \,\middle|\, \dot{\boldsymbol{X}} \right\rangle \overset{c}{=} H(\mathbf{x}_1) - H(\mathbf{x}_0)$, using (a) Theorem 5.1 with $\mathbf{A} = \mathbf{J} = -\mathbf{J}^\mathsf{T}$, (b) $\boldsymbol{f} = \dot{\boldsymbol{X}}$, $\boldsymbol{e} = \nabla H(\boldsymbol{X})$, (c) the gradient theorem. □

Corollary 5.2 (Projected pH-ODE)**.** *Consider a projected input-state-output pH-ODE with given input* $\boldsymbol{u} \in L^2(\Omega, \mathbb{R}^{n_P})$ *and* $(\boldsymbol{\mathcal{P}}, \mathbf{J} - \mathbf{R})$ *satisfying* (5.2)

$$\begin{bmatrix} \dot{\boldsymbol{X}} \\ \boldsymbol{y} \end{bmatrix} = \boldsymbol{\mathcal{P}}\left(\mathbf{J} - \mathbf{R}\right) \begin{bmatrix} \nabla H(\boldsymbol{X}) \\ \boldsymbol{u} \end{bmatrix}, \qquad \boldsymbol{X}(t_0) = \mathbf{x}_0, \tag{5.5}$$

Then, for $\boldsymbol{X} \in H^1(\Omega, \mathbb{R}^{n_S})$*,* $\boldsymbol{y} \in L^2(\Omega, \mathbb{R}^{n_P})$ *solutions of* (5.5)*,* $\mathbf{x}_1 = \boldsymbol{X}(t_1)$*, the projected pH-ODE is passive, i.e. it satisfies the average power balance over* $\Omega = (t_0, t_1)$

$$H(\mathbf{x}_1) - H(\mathbf{x}_0) + \langle \boldsymbol{u} \,|\, \boldsymbol{y} \rangle \leq 0.$$

Proof. The result follows from $0 \overset{a}{\geq} \langle \boldsymbol{e} \,|\, \boldsymbol{f} \rangle \overset{b}{=} \left\langle \nabla H(\boldsymbol{X}) \,\middle|\, \dot{\boldsymbol{X}} \right\rangle + \langle \boldsymbol{u} \,|\, \boldsymbol{y} \rangle \overset{c}{=} H(\mathbf{x}_1) - H(\mathbf{x}_0) + \langle \boldsymbol{u} \,|\, \boldsymbol{y} \rangle$, using (a) Theorem 5.1 with $\mathbf{A} = \mathbf{J} - \mathbf{R}$, (b) $\boldsymbol{f} = \left[\begin{smallmatrix} \dot{\boldsymbol{X}} \\ \boldsymbol{y} \end{smallmatrix}\right]$, $\boldsymbol{e} = \left[\begin{smallmatrix} \nabla H(\boldsymbol{X}) \\ \boldsymbol{u} \end{smallmatrix}\right]$, (c) the gradient theorem. □

Corollary 5.3 (Projected pH-DAE)**.** *Consider the projected semi-explicit pH-DAE with given input* $\boldsymbol{u} \in L^2(\Omega, \mathbb{R}^{n_P})$ *and* $(\boldsymbol{\mathcal{P}}, \mathbf{J})$ *satisfying* (5.2)

$$\begin{bmatrix} \dot{\boldsymbol{X}} \\ \boldsymbol{w} \\ \boldsymbol{y} \end{bmatrix} = \boldsymbol{\mathcal{P}}\mathbf{J} \begin{bmatrix} \nabla H(\boldsymbol{X}) \\ \boldsymbol{z}(\boldsymbol{w}) \\ \boldsymbol{u} \end{bmatrix}, \qquad \boldsymbol{X}(t_0) = \mathbf{x}_0.\cdot \tag{5.6}$$

Then, for $\boldsymbol{X} \in H^1(\Omega, \mathbb{R}^{n_S})$*,* $\boldsymbol{w} \in L^2(\Omega, \mathbb{R}^{n_R})$ $\boldsymbol{y} \in L^2(\Omega, \mathbb{R}^{n_P})$ *solutions of* (5.6) *and* $\mathbf{x}_1 = \boldsymbol{X}(t_1)$*, the projected pH-DAE is passive, i.e. it satisfies the average power balance over* $\Omega = (t_0, t_1)$

$$H(\mathbf{x}_1) - H(\mathbf{x}_0) + \langle \boldsymbol{u} \,|\, \boldsymbol{y} \rangle = -\left\langle \boldsymbol{z}(\boldsymbol{w}) \,\middle|\, \boldsymbol{w} \right\rangle \leq 0.$$

Proof. The results follows from

$$\begin{aligned} & 0 \overset{a}{=} \langle \boldsymbol{e} \,|\, \boldsymbol{f} \rangle \overset{b}{=} \left\langle \nabla H(\boldsymbol{X}) \,\middle|\, \dot{\boldsymbol{X}} \right\rangle + \left\langle \boldsymbol{z}(\boldsymbol{w}) \,\middle|\, \boldsymbol{w} \right\rangle + \langle \boldsymbol{u} \,|\, \boldsymbol{y} \rangle \\ \overset{c}{\Longleftrightarrow} \quad & H(\mathbf{x}_1) - H(\mathbf{x}_0) + \langle \boldsymbol{u} \,|\, \boldsymbol{y} \rangle = -\left\langle \boldsymbol{z}(\boldsymbol{w}) \,\middle|\, \boldsymbol{w} \right\rangle \overset{d}{\leq} 0, \end{aligned}$$

using (a) Theorem 5.1 with $\mathbf{A} = \mathbf{J}$, (b) $\boldsymbol{f} = (\dot{\boldsymbol{X}}, \boldsymbol{w}, \boldsymbol{y})$, $\boldsymbol{e} = (\nabla H(\boldsymbol{X}), \boldsymbol{z}(\boldsymbol{w}), \boldsymbol{u})$ (c) the gradient theorem and (d) pointwise non-negativity of $\boldsymbol{z}(\mathbf{w}) \cdot \mathbf{w} \geq 0$. □

5.1.2 Examples of projector design

Theorem 5.1 allows some flexibility in the design of projectors. This can be illustrated on example 2.10 61 (Diode clipper) whose structure matrix $\mathbf{J}$ is recalled below.

$$\begin{bmatrix} v_R \\ v_D \\ i_C \\ i_S \end{bmatrix} = \begin{bmatrix} & & -1 & 1 \\ & & 1 & 0 \\ 1 & -1 & & \\ -1 & 0 & & \end{bmatrix} \begin{bmatrix} i_R(v_R) \\ i_D(v_D) \\ v_C(i_C) \\ v_S \end{bmatrix}$$

Several choices of projectors $\boldsymbol{\mathcal{P}}$ can be considered

a) The simplest choice consists in using the same scalar projector $\mathcal{P} = \mathcal{P}^*$ for each dimension by introducing $\boldsymbol{\mathcal{P}} = \mathcal{P} \otimes \mathbf{I}_4$ (by construction $\boldsymbol{\mathcal{P}}\mathbf{J} = \mathbf{J}\boldsymbol{\mathcal{P}} = \mathbf{J}\boldsymbol{\mathcal{P}}^*$). This defines the skew-adjoint operator

$$\boldsymbol{\mathcal{P}}\mathbf{J} = \begin{bmatrix} & & -\mathcal{P}^* & \mathcal{P}^* \\ & & \mathcal{P}^* & 0 \\ \mathcal{P} & -\mathcal{P} & & \\ -\mathcal{P} & 0 & & \end{bmatrix}$$

This choice is the one explored and detailed in section 5.1 to build Power-Balanced methods for pH-ODEs and pH-DAEs.

b) A natural extension, is to use a diagonal projector $\boldsymbol{\mathcal{P}} = \mathrm{diag}(\mathcal{P}_R, \mathcal{P}_D, \mathcal{P}_C, \mathcal{P}_S)$ with different (not necessarily self-adjoint) projectors for each dimension so that

$$\boldsymbol{\mathcal{P}}\mathbf{J} = \begin{bmatrix} \mathcal{P}_R & & & \\ & \mathcal{P}_D & & \\ & & \mathcal{P}_C & \\ & & & \mathcal{P}_S \end{bmatrix} \begin{bmatrix} & & -1 & 1 \\ & & 1 & 0 \\ 1 & -1 & & \\ -1 & 0 & & \end{bmatrix} = \begin{bmatrix} & & -\mathcal{P}_R & \mathcal{P}_R \\ & & \mathcal{P}_D & 0 \\ \mathcal{P}_C & -\mathcal{P}_C & & \\ -\mathcal{P}_S & 0 & & \end{bmatrix}.$$

However, note that, in order to have $\boldsymbol{\mathcal{P}}\mathbf{J}$ skew-adjoint, it is necessary to fulfil hidden constraints $P_R = \mathcal{P}_D = \mathcal{P}$ and $\mathcal{P}_C = \mathcal{P}_S = \mathcal{P}^*$ for a given projector $\mathcal{P}$ (and its adjoint $\mathcal{P}^*$). This choice is more flexible than the self-adjointness constraint (a) for partitionnable systems. In particular, canonical Hamiltonian systems could be discretized as

$$\dot{\boldsymbol{p}} = -\mathcal{P}^* \frac{\partial H}{\partial \boldsymbol{q}}(\boldsymbol{p}, \boldsymbol{q}), \qquad \dot{\boldsymbol{q}} = \mathcal{P} \frac{\partial H}{\partial \boldsymbol{p}}(\boldsymbol{p}, \boldsymbol{q}).$$

c) The most general situation arises by direct substitution of each cell of the structure matrix by projectors to obtain a skew-adjoint approximation of the structure matrix $\mathbf{J}$ (or $\mathbf{J}-\mathbf{R}$). In our example, we may choose 3 projectors $\mathcal{P}_{CR}, \mathcal{P}_{CD}, \mathcal{P}_{SR}$ such that the following functional matrix operator $\boldsymbol{\mathcal{J}}$ (approximating $\mathbf{J}$) is skew-adjoint

$$\boldsymbol{\mathcal{J}} = \begin{bmatrix} & & -\mathcal{P}_{CR}^* & \mathcal{P}_{SR}^* \\ & & \mathcal{P}_{CD}^* & 0 \\ \mathcal{P}_{CR} & -\mathcal{P}_{CD} & & \\ \mathcal{P}_{SR} & 0 & & \end{bmatrix} = -\boldsymbol{\mathcal{J}}^*.$$

Alternatively, we could define the skew-adjoint operator $\boldsymbol{\mathcal{J}} = \boldsymbol{\mathcal{P}}\mathbf{J}\boldsymbol{\mathcal{P}}^*$ from (b). This choice is not explored further in this thesis, but is left as an interesting perspective for future work.

5.1.3 RPM for pH-ODE

We propose a power-balanced method for pH-ODEs. The key ideas of the method are a) to use corollary 5.2 to obtain projected power balanced solutions (**P2**) in a subspace of L^2, b) to improve this result using multi-derivatives supplementary boundary conditions (**P1**) so that the concatenation of time frames yields globally smooth solutions in the Sobolev space H^k.

For our purposes, we rewrite input-state-output pH-ODEs from definition 1.22 p.33 as

$$\underbrace{\begin{bmatrix}\dot{\boldsymbol{x}}\\ \boldsymbol{y}\end{bmatrix}}_{\boldsymbol{a}} = (\mathbf{J}-\mathbf{R})\underbrace{\begin{bmatrix}\nabla H(\boldsymbol{x})\\ \boldsymbol{u}\end{bmatrix}}_{\boldsymbol{b}} =: \begin{bmatrix}\boldsymbol{f}(\boldsymbol{x},\boldsymbol{u})\\ \boldsymbol{g}(\boldsymbol{x},\boldsymbol{u})\end{bmatrix}, \quad \boldsymbol{x}(t_0)=\mathbf{x}_0. \tag{5.7}$$

with $\dot{\boldsymbol{x}}(t), \nabla H(\boldsymbol{x}(t)) \in \mathbb{R}^{n_{\mathbf{x}}}$ and $\boldsymbol{y}(t), \boldsymbol{u}(t) \in \mathbb{R}^{n_{\mathbf{y}}}$. In this chapter, let $[t_0, t_1]$ be a time step, $h = t_1 - t_0$ its step size, and $t(\tau) = t_0 + h\tau$, with $\tau \in \overline{\Omega} = [0,1]$ a time variable change for which we define the differential operator $\mathcal{D} := \frac{1}{h}\frac{\mathrm{d}}{\mathrm{d}\tau}$ (i.e. $\mathcal{D} \equiv \frac{\mathrm{d}}{\mathrm{d}t}$). We propose the following method

Method 5.1 (RPM for pH-ODE)**.** Denote p be the projection order, k the regularity order, $\ell = p + 2k$ and $n = n_{\mathbf{x}} + n_{\mathbf{y}}$. A *Regular Power-balanced projection Method* called RPM(p,k) for pH-ODE (5.7) is defined by steps (i)-(iii)

i) **P0** *Approximation spaces and operators*: Let $\{\phi_i\}_{i=0}^{\ell-1} \in H^k(\Omega) \subset L^2(\Omega)$ be an orthonormal basis for the L^2 inner product and define the subspaces of $L^2(\Omega)$

$$A_{\mathcal{P}} := \operatorname{span}\{\phi_i\}_{i=0}^{p-1}, \qquad A_R := \operatorname{span}\{\phi_i\}_{i=p}^{\ell-1}, \qquad A := A_{\mathcal{P}} \oplus A_R. \tag{5.8}$$

We assume that (H1) $A_{\mathcal{P}}$ is such that the orthogonal projector $\mathcal{P}$ on $A_{\mathcal{P}}$, reproduces constant functions and that (H2) the image of A_R through $\mathcal{B}$ spans $\mathbb{R}^{2k}$ where $\mathcal{B} : H^k(\Omega,\mathbb{R}) \to \mathbb{R}^{2k}$ the (multi-derivatives) boundary trace operator [Aub11, p.163] is

$$\mathcal{B} := \left(\mathcal{B}_0^0, \dots, \mathcal{B}_0^{k-1}, \mathcal{B}_1^0, \dots, \mathcal{B}_1^{k-1}\right), \qquad \text{with} \qquad \mathcal{B}_\alpha^m(u) := (\mathcal{D}^m u)(\alpha). \tag{5.9}$$

Denote $\widetilde{\boldsymbol{A}} = A^{n_{\mathbf{x}}} \times A^{n_{\mathbf{y}}}$, and $\widetilde{\boldsymbol{B}} \simeq \widetilde{\boldsymbol{A}}$ approximation spaces for dual variables $\boldsymbol{a}$, $\boldsymbol{b}$ and $\boldsymbol{\mathcal{P}} = \mathcal{P} \otimes \mathbf{I}_n$, $\boldsymbol{\mathcal{B}} = \mathcal{B} \otimes \mathbf{I}_n$ the extensions of $\mathcal{P}, \mathcal{B}$ to $L^2(\Omega)^n$ and $H^k(\Omega)^n$.

ii) **P2, P3** *Accuracy and power balance*: Denote $\boldsymbol{a}_{\mathcal{P}} = (\delta\boldsymbol{X}, \boldsymbol{Y}) \in \boldsymbol{\mathcal{P}}(\widetilde{\boldsymbol{A}})$ the unknowns of the projection step and define the time-stepping method $\Phi_h : \mathbf{x}_0 \mapsto \mathbf{x}_1$ such that

$$\begin{bmatrix}\delta\boldsymbol{X}\\ \boldsymbol{Y}\end{bmatrix} = \boldsymbol{\mathcal{P}}(\mathbf{J}-\mathbf{R})\begin{bmatrix}\nabla H(\boldsymbol{X})\\ \boldsymbol{u}\end{bmatrix}, \quad \text{where} \quad \begin{cases}\boldsymbol{X}(\tau) & := \mathbf{x}_0 + h\displaystyle\int_0^\tau \delta\boldsymbol{X}(\sigma)\,\mathrm{d}\sigma,\\ \mathbf{x}_1 & := \boldsymbol{X}(1).\end{cases} \tag{5.10}$$

iii) **P1** *Regularity*: For $k \geq 1$, denote $\widetilde{\boldsymbol{a}} = (\widetilde{\delta\boldsymbol{X}}, \widetilde{\boldsymbol{Y}}) \in \widetilde{\boldsymbol{A}}$ the unknowns of the regularisation step such that $\boldsymbol{\mathcal{P}}\widetilde{\boldsymbol{a}} = \boldsymbol{a}_{\mathcal{P}}$ and satisfying the multi-derivatives boundary conditions

$$\boldsymbol{\mathcal{B}}\begin{bmatrix}\widetilde{\delta\boldsymbol{X}}\\ \widetilde{\boldsymbol{Y}}\end{bmatrix} = \boldsymbol{\mathcal{B}}(\mathbf{J}-\mathbf{R})\begin{bmatrix}\nabla H(\widetilde{\boldsymbol{X}})\\ \boldsymbol{u}\end{bmatrix}, \quad \text{where} \quad \widetilde{\boldsymbol{X}}(\tau) := \mathbf{x}_0 + h\int_0^\tau \delta\widetilde{\boldsymbol{X}}(\sigma)\,\mathrm{d}\sigma. \tag{5.11}$$

The condition $\boldsymbol{\mathcal{P}}\widetilde{\boldsymbol{a}} = \boldsymbol{a}_{\mathcal{P}}$ ensures that the regular solution $\widetilde{\boldsymbol{a}}$ is at least as good as the projected solution $\boldsymbol{a}_{\mathcal{P}}$, (i.e. regularity is not in conflict with the power balance). Furthermore, if the

projector $\mathcal{P}$ reproduces constants (H1), then by orthogonality, $\int_0^1 \phi_n(s)\,\mathrm{d}s = 0$ for all $n \geq p$, such that by construction the projected and the regularised trajectories share the same endpoint $\mathbf{x}_1 = \boldsymbol{X}(1) = \widetilde{\boldsymbol{X}}(1)$. By consequence supplementary boundary conditions (5.11) only depend on the numerical value of vectors $\mathbf{x}_0, \mathbf{x}_1$ and on the formal derivatives of functions $\nabla H, \boldsymbol{u}$ (see section B.3 p.278 for numerical evaluation). Hypothesis (H2) ensures that steps (iii) is solvable.

5.1.4 RPM for pH-DAE

We extend the method RPM(p,k) from method 5.1 to semi-explicit pH-DAEs. The main difference comes from the appearance of memoryless algebraic constraints through the variables $\boldsymbol{w}$. For our purposes, we rewrite semi-explicit pH-DAEs from definition 1.24 p.34 as

$$\underbrace{\begin{bmatrix}\dot{\boldsymbol{x}}\\ \boldsymbol{w}\\ \boldsymbol{y}\end{bmatrix}}_{\boldsymbol{a}} = \mathbf{J}\underbrace{\begin{bmatrix}\nabla H(\boldsymbol{x})\\ \boldsymbol{z}(\boldsymbol{w})\\ \boldsymbol{u}\end{bmatrix}}_{\boldsymbol{b}} =: \begin{bmatrix}\boldsymbol{f}(\boldsymbol{x},\boldsymbol{w},\boldsymbol{u})\\ \boldsymbol{g}(\boldsymbol{x},\boldsymbol{w},\boldsymbol{u})\\ \boldsymbol{h}(\boldsymbol{x},\boldsymbol{w},\boldsymbol{u})\end{bmatrix}, \quad \boldsymbol{x}(t_0) = \mathbf{x}_0. \tag{5.12}$$

Method 5.2 (RPM for pH-DAE)**.** Let $p > 0$ be the projection order and $k \geq 0$ the regularity order and $n = n_{\mathbf{x}} + n_{\mathbf{w}} + n_{\mathbf{y}}$. A *Regular Power balanced projection Method* RPM(p,k) for pH-DAE (5.12) is defined by steps (i)-(iii)

i) **P0** *Approximation spaces*: Let $A, A_{\mathcal{P}}, A_R$ be approximations spaces from (5.8). Let $\widetilde{\boldsymbol{A}} = \mathcal{H}^{n_{\mathbf{x}}} \times \mathcal{H}^{n_{\mathbf{w}}} \times \mathcal{H}^{n_{\mathbf{y}}}$, $\widetilde{\boldsymbol{B}} \simeq \widetilde{\boldsymbol{A}}$ and denote $\boldsymbol{\mathcal{P}} = \mathcal{P} \otimes \mathbf{I}_n$, $\boldsymbol{\mathcal{B}} = \mathcal{B} \otimes \mathbf{I}_n$.

ii) **P2, P3** *Accuracy and power balance*: denote $\boldsymbol{a}_{\mathcal{P}} = (\delta\boldsymbol{X}, \boldsymbol{W}, \boldsymbol{Y}) \in \boldsymbol{\mathcal{P}}(\widetilde{\boldsymbol{A}})$ the unknowns of the projection step and define the time-stepping method $\Phi_h : \mathbf{x}_0 \mapsto \mathbf{x}_1$ such that

$$\begin{bmatrix}\delta\boldsymbol{X}\\ \boldsymbol{W}\\ \boldsymbol{Y}\end{bmatrix} = \boldsymbol{\mathcal{P}}\mathbf{J}\begin{bmatrix}\nabla H(\boldsymbol{X})\\ \boldsymbol{z}(\boldsymbol{W})\\ \boldsymbol{u}\end{bmatrix}, \qquad \begin{cases}\boldsymbol{X}(\tau) & := \mathbf{x}_0 + h\displaystyle\int_0^\tau \delta\boldsymbol{X}(\sigma)\,\mathrm{d}\sigma,\\ \mathbf{x}_1 & := \boldsymbol{X}(1).\end{cases} \tag{5.13}$$

iii) **P1** *Regularity*: denote $\widetilde{\boldsymbol{a}} = (\widetilde{\delta\boldsymbol{X}}, \widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{Y}}) \in \widetilde{\boldsymbol{A}}$ the unknowns of the regularisation step such that $\boldsymbol{\mathcal{P}}\widetilde{\boldsymbol{a}} = \boldsymbol{a}_{\mathcal{P}}$ and satisfying the multi-derivative boundary conditions

$$\boldsymbol{\mathcal{B}}\begin{bmatrix}\widetilde{\delta\boldsymbol{X}}\\ \widetilde{\boldsymbol{W}}\\ \widetilde{\boldsymbol{Y}}\end{bmatrix} = \boldsymbol{\mathcal{B}}\mathbf{J}\begin{bmatrix}\nabla H(\widetilde{\boldsymbol{X}})\\ \boldsymbol{z}(\widetilde{\boldsymbol{W}})\\ \boldsymbol{u}\end{bmatrix}, \qquad \widetilde{\boldsymbol{X}}(\tau) := \mathbf{x}_0 + h\int_0^\tau \widetilde{\delta\boldsymbol{X}}(\sigma)\,\mathrm{d}\sigma. \tag{5.14}$$

Note that solutions of equation (5.13) are only weak DAE solutions in the sense of L^2 projection. In particular, concatenation of time steps yields piecewise discontinuous solutions in step ii). The boundary values of flow and efforts are not defined in L^2: only $\boldsymbol{X}$ (but not $\delta\boldsymbol{X}$) is piecewise continuous because of integration. However step iii) restores continuity such that the concatenation of time-frames for $\widetilde{\delta\boldsymbol{X}}, \widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{Y}}$ yields globally smooth functions in the Sobolev space H^k.

Comments Note that, contrary to most numerical methods, because of our virtual analog viewpoint (see chapter 3 p.81), in practice, we are more interested in the quality of the continuous-time approximation of dual flow/efforts variables $\widetilde{\boldsymbol{a}} = (\delta\widetilde{\boldsymbol{X}}, \widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{Y}})$, $\widetilde{\boldsymbol{b}} = (\nabla H(\widetilde{\boldsymbol{X}}), \boldsymbol{z}(\widetilde{\boldsymbol{W}}), \boldsymbol{u})$ than on the sequence of values $\{\mathbf{x}_n\}$ produced by the discrete time-stepping map $\Phi_h : \mathbf{x}_n \mapsto \mathbf{x}_{n+1}$. This map remains an important object to study the numerical properties of the numerical schemes, but from a signal processing perspective, it only gives us a partial viewpoint by sampling the (non bandlimited!) trajectory $\boldsymbol{X}$ on the boundaries of each time frame.

Note that passivity propagates from time-frame to time-frame. Also note that for RPM, contrary to symplectic[3] methods [HLW06, KL19], *the exact Hamiltonian* is preserved (resp. dissipated) when it is evaluated *on the boundaries* of each time frame (see figure 5.12 p.152 for the behaviour of the energy within each time interval).

Theoretical existence and uniqueness conditions for RPM are addressed in section for 5.2 (for pH-ODE) and in section 5.3 (for pH-DAE). Accuracy analysis is detailed in subsection 5.2.6. Computational implementation details such as the computation of projections, the evaluation of boundary derivatives or implicit equation solving are considered in section 5.4.

In particular we give the following results

- RPM are energy (resp. passivity) preserving (see corollary 5.2).
- RPM are A-stable (see proposition 5.2 p.127 and section D.7 p.297).
- RPM have (pointwise) accuracy order $2p$ (on interval boundaries[4]) if the projector $\mathcal{P}$ reproduces polynomials of degree $p-1$ (see subsection 5.2.6 p.128). For this reason, in applications, we will use the shifted (L^2) orthonormal Legendre polynomial basis (defined in section C.4 p.286). For comparison, in section D.7 p.297 we provide the stability function of the orthonormal cosine basis (which only yields second order accurate time-stepping approximations).
- The regularisation step (iii) yields a secondary (non self-adjoint) projector $\mathcal{Q}$ (formalised in subsection 5.2.7 p.129). Peano error kernels of projectors $\mathcal{P}$ and $\mathcal{Q}$ are derived and shown in figures 5.4 and 5.5 p.134.
- A graphical illustration of the method and of the respective roles of *nested projectors* $\mathcal{P}$ and $\mathcal{Q}$ is shown in figure D.1 p.295.

Readers that are not interested in the theoretical or technical details, may skip directly to the examples shown in section 5.5 p.147.

3. It is known from [ZM88] that approximate symplectic algorithms cannot preserve energy for nonintegrable systems.

4. Conversely, the accuracy (in the L^2 norm) of continuous-time flow and effort trajectories *within* each time frame is proportional to the number of degrees of freedom $p + 2k$.

5.2 Analysis of RPM for pH-ODE

To analyse RPM, in order to compare with the literature and to study existence/uniquess, and accuracy conditions, it is convenient to reformulate (5.10) (def. 5.1, step ii) using the framework of *continuous-stage Runge Kutta* methods (CSRK). The main object in this section is the orthogonal projector $\mathcal{P}$ whose reproducing kernel is (see eq. (3.9) p.84)

$$K_{\mathcal{P}}(\tau,\sigma) = \sum_{i=0}^{p-1} \phi_n(\tau)\phi_n(\sigma). \tag{5.15}$$

in a chosen orthonormal basis such that $\operatorname{span}\{\phi_n\}_{n=0}^{p-1} = A_{\mathcal{P}}$.

We show in sections 5.2.2 to 5.2.6 that CSRK parameters can all be obtained from the kernel $K_{\mathcal{P}}$ and that energy-preservation, existence/uniqueness, stability function and accuracy automatically follow from the properties of $\mathcal{P}$. Then we show in section 5.2.7 that the third step of RPM (the regularisation step) yields another (oblique) projector $\mathcal{Q}$ refining $\mathcal{P}$ and we compare their respective approximation properties and Peano error kernels.

5.2.1 Reminder on Runge-Kutta methods

Definition 5.2 (Runge–Kutta method [HLW06] p.29)**.** Let b_i, $a_{i,j}$ $(i,j = 1,\dots,s)$ be real numbers and let $c_i = \sum_{j=1}^{s} a_{ij}$. An *s-stage Runge–Kutta method is given by*

$$\begin{cases} \mathbf{k}_i = \boldsymbol{f}\left(t_0 + hc_i,\ \mathbf{x}_0 + h\sum_{j=1}^{s} a_{ij}\mathbf{k}_j\right), \quad i = 1,\dots,s \\ \mathbf{x}_1 = \mathbf{x}_0 + h\sum_{i=1}^{s} b_i\mathbf{k}_i. \end{cases} \tag{5.16}$$

The slopes $\mathbf{k}_i$ do not necessarily exists, however, the implicit function theorem assures that, for sufficiently small h, the nonlinear system for the values $\mathbf{k}_1,\dots,\mathbf{k}_s$ has a locally unique solution close to $\mathbf{k}_i \approx \boldsymbol{f}(t_0,\mathbf{x}_0)$. Since Butcher's work the coefficients are usually displayed as follows

$$\begin{array}{c|ccc} c_1 & a_{11} & \dots & a_{1s} \\ \vdots & \vdots & & \vdots \\ c_s & a_{s1} & \dots & a_{ss} \\ \hline & b_1 & \dots & b_s \end{array} \quad \equiv \quad \begin{array}{c|c} \mathbf{c} & \mathbf{A} \\ \hline & \mathbf{b} \end{array}.$$

5.2.2 Reformulation of RPM as Continuous-Stage Runge-Kutta methods

The idea of CSRK was hinted by Butcher in [But72], but it had to wait until the work of Hairer in 2010 [Hai10] to understand the key role of CSRK methods to derive energy-preserving integrators. Early examples of energy–preserving CSRK method are the Average Vector Field method [QM08, CGM+12, COS14] and Hamiltonian Boundary Value methods (HBVMs) which were later interpreted as CSRK in [ABI19]. A similar thread of research arises from the use of Time Finite Elements Methods (TFEM) and (Continuous) Galerkin projection in time [TS12] based on ideas that can be traced back to [Hul92, BB93, Bot97, BS00]. For more details on CSRK methods please refer to the overview paper [Tan18].

CSRK methods are generalisations of Runge–Kutta methods (5.16) for an infinite number of stage values $\dot{\boldsymbol{X}}(\tau)$ so that the matrix $\mathbf{A}$, weights $\mathbf{b}$ and abscissae $\mathbf{c}$ in def. 5.2 are replaced by *functions* $A(\tau,\sigma), B(\sigma), C(\tau)$.

Definition 5.3 (CSRK method [Tan18])**.** A *Continuous-Stage Runge-Kutta method* is a one step method $\Phi_h : \mathbf{x}_0 \mapsto \mathbf{x}_1$ defined by a triplet $(A : \Omega \times \Omega \to \mathbb{R},\ B : \Omega \to \mathbb{R},\ C : \Omega \to \mathbb{R})$ and

$$\boldsymbol{X}(\tau) = \mathbf{x}_0 + h \int_0^1 A(\tau,\sigma) \boldsymbol{f}\left(t(\sigma), \boldsymbol{X}(\sigma)\right) \mathrm{d}\sigma, \tag{5.17a}$$

$$\mathbf{x}_1 = \mathbf{x}_0 + h \int_0^1 B(\sigma) \boldsymbol{f}\left(t(\sigma), \boldsymbol{X}(\sigma)\right) \mathrm{d}\sigma, \tag{5.17b}$$

where $\boldsymbol{X}(\tau)$ approximates $\boldsymbol{x}\left(t(\tau)\right)$ at times $t(\tau) = t_0 + hC(\tau)$ for $\tau \in \Omega = [0,1]$.

Proposition 5.1. *For RPM, the reproducing kernel $K_{\mathcal{P}}(\tau,\sigma)$ of projector $\mathcal{P}$ defined by (5.15) uniquely defines the CSRK triplet.*

$$A(\tau,\sigma) = \int_0^\tau K_{\mathcal{P}}(\xi,\sigma)\,\mathrm{d}\xi, \tag{5.18a}$$

$$B(\sigma) = \int_0^1 K_{\mathcal{P}}(\tau,\sigma)\,\mathrm{d}\tau = 1, \tag{5.18b}$$

$$C(\tau) = \int_0^\tau \int_0^1 K_{\mathcal{P}}(\xi,\sigma)\,\mathrm{d}\sigma\,\mathrm{d}\xi = \tau. \tag{5.18c}$$

Proof. The proof in detailed in section D.2 p.291. □

Remark 5.1. For consistency, it is often assumed [Tan18, 2.3] that

$$C(\tau) = \int_0^1 A(\tau,\sigma)\,\mathrm{d}\sigma.$$

For RPM, this is automatically fulfilled because of (5.18a)-(5.18c). Also note that, differentiating (5.18c) and comparing with (5.18b) yields the symmetric relation between C' and B

$$C'(\tau) = \int_0^1 K_{\mathcal{P}}(\tau,\sigma)\,\mathrm{d}\sigma = \qquad 1 \qquad = \int_0^1 K_{\mathcal{P}}(\tau,\sigma)\,\mathrm{d}\tau = B(\tau).$$

In short, the symmetry of kernel $K_{\mathcal{P}}$ (self-adjointness of $\mathcal{P}$) and the reproduction of constants ensure that the weight $B(\sigma) = 1$ is uniform (5.18b) and consistent with the (uniform density of the) measure $\mathrm{d}C(\tau) = B(\tau)\,\mathrm{d}\tau$ in the variable change $t(\tau) = t_0 + hC(\tau) \implies \mathrm{d}t(\tau) = h\,\mathrm{d}\tau$.

5.2.3 Existence and uniqueness of solutions

Here we provide existence and uniqueness conditions for a CSRK method when $\dot{\boldsymbol{X}} \in L^2(\Omega, \mathbb{R}^n)$ and $\mathcal{P}$ is an orthogonal projector. Our result and proof are different from the ones in [TS12, MB16, Tan18] because we consider convergence in the L^2 norm.

Theorem 5.2. *Let $\mathcal{P}$ be an orthogonal L^2 projector such that the associated CSRK method (def.5.3) satisfies (5.18a)–(5.18c). If $\boldsymbol{f}$ is L-Lipschitz and $hL < \frac{\pi}{2}$. Then, the method has a unique solution in L^2.*

Proof. The proof is detailed in section D.4 p.293. □

5.2.4 Linear Stability function

We consider the Dahlquist test equation $\dot{x} = \lambda x$, $x(0) = x_0$, $\lambda \in \mathbb{C}$, and a time stepping method $\Phi_\lambda : x_0 \mapsto x_1 = R(\lambda)x_0$ defined by the orthogonal projection $\dot{x} = \mathcal{P}\lambda x$ on $\Omega = (0, 1)$.

Proposition 5.2 (Stability function)**.** *Let $\{\phi_n(\tau)\}_{n=0}^{p-1}$ be an orthonormal basis of dimension p in $L^2(\Omega)$ reproducing constants. Let $\mathbf{1} = [\langle \phi_n, 1\rangle]_{n=0}^{p-1}$ and*

$$\mathbf{V} = \left[\langle \phi_m, \mathcal{V}\phi_n\rangle\right]_{m,n=0\dots p-1}, \qquad \textit{where} \qquad (\mathcal{V}u)(\tau) = \int_0^\tau u(s)\,\mathrm{d}s, \tag{5.19}$$

be the matrix representations of the constant function and of the (projected) Volterra integration operator. Then, the stability function of method Φ_λ with projection order p is given by

$$R(\lambda) = 1 + \lambda \mathbf{1}^\mathsf{T}(\mathbf{I} - \lambda\mathbf{V})^{-1}\mathbf{1} = \frac{\det(\mathbf{I} + \lambda\mathbf{V}^\mathsf{T})}{\det(\mathbf{I} - \lambda\mathbf{V})}. \tag{5.20}$$

Proof. The proof is detailed in section D.7 p.297. □

5.2.5 Energy preservation (P3)

Since $\mathcal{P}$ is an orthogonal projector, it is self adjoint ($\mathcal{P} = \mathcal{P}^*$). Furthermore, by construction it commutes with matrices, so we already know from corollary 5.2 p.120 that our method is energy, (resp. passivity) preserving for pH-ODE. Here, we provide an alternate interpretation using CSRK theory to highlight the role of the reproducing kernel $K_\mathcal{P}$.

In the context of CSRK methods, a method is energy–preserving [Tan18, thm.3.7] when

$$\left(\frac{\partial A}{\partial \tau}\right)(\tau, \sigma) = \left(\frac{\partial A}{\partial \tau}\right)(\sigma, \tau), \qquad A(0, \sigma) = 0, \qquad A(1, \sigma) = B(\sigma).$$

Reformulated with the reproducing kernel $K_\mathcal{P}$, using (5.18a), this is equivalent to the three conditions

$$K_\mathcal{P}(\tau, \sigma) \overset{a}{=} K_\mathcal{P}(\sigma, \tau), \qquad \int_0^\tau K_\mathcal{P}(\xi, \sigma)\,\mathrm{d}\xi\Big|_{\tau=0} \overset{b}{=} 0, \qquad \int_0^1 K_\mathcal{P}(\tau, \sigma)\,\mathrm{d}\tau \overset{c}{=} B(\sigma).$$

(a) The symmetry of $K_\mathcal{P}(\tau, \sigma) = \sum_{i=0}^{p-1} \phi_i(\tau)\phi_i(\sigma)$ follows from its construction. It is equivalent to $\mathcal{P}$ being self-adjoint. (b) The second condition always hold when $K_\mathcal{P} \in L^2(\Omega) \otimes L^2(\Omega)$ (i.e. $K_\mathcal{P}$ does not contain Dirac delta distributions) and (c) the third condition is fulfilled by (5.18b). This is equivalent to $\mathbf{x}_1 = \boldsymbol{X}(1)$.

5.2.6 Order conditions and polynomial reproduction (P2)

Usually, the accuracy order of one-step methods is studied using the theory of B-series [HLW06]. Here, we establish that CSRK order conditions are automatically fulfilled when the RPM projector $\mathcal{P}$ reproduces polynomials up to a given order (Strang-Fix conditions).

Definition 5.4 (Accuracy order [Tan18])**.** A CSRK method is of *accuracy order* s if for all sufficiently regular problems (5.17a)-(5.17b) its local error satisfies $\boldsymbol{x}(t_0+h)-\mathbf{x}_1=\mathcal{O}(h^{s+1})$ as $h\to 0$.

The main tool we use to study accuracy is a generalisation to CSRK methods of the *simplifying order assumptions* for Runge–Kutta methods (see [BG08, p.186] and [HNW93, p.208]). They are given by the following theorem.

Theorem 5.3 (Simplifying order assumptions [Hai10])**.** *If a CSRK method satisfies the* simplifying order assumptions *for integers* $\rho,\eta,\zeta\geq 1$.

$$\check{B}(\rho):\qquad \int_0^1 B(\tau)C(\tau)^{k-1}\,\mathrm{d}\tau=\frac{1}{k},\qquad k=1,\dots,\rho, \tag{5.21a}$$

$$\check{C}(\eta):\qquad \int_0^1 A(\tau,\sigma)C(\sigma)^{k-1}\,\mathrm{d}\sigma=\frac{C(\tau)^k}{k},\qquad k=1,\dots,\eta, \tag{5.21b}$$

$$\check{D}(\zeta):\qquad \int_0^1 B(\tau)C(\tau)^{k-1}A(\tau,\sigma)\,\mathrm{d}\tau=\frac{1}{k}B(\sigma)(1-C(\sigma)^k),\qquad k=1,\dots,\zeta. \tag{5.21c}$$

Then, its accuracy order is at least $s\geq\min(\rho,2\eta+2,\eta+\zeta+1)$.

In RPM, these conditions are greatly simplified, they are linked to the polynomial reproduction properties of the projector $\mathcal{P}$. To this end, we establish the following proposition.

Proposition 5.3. *Let* $\mathcal{P}$ *be a projector with kernel* $K_{\mathcal{P}}(\tau,\sigma)$ *such that the associated CSRK method satisfies* $B(\sigma)=1$, $C(\tau)=\tau$, $\frac{\partial A}{\partial\tau}=K_{\mathcal{P}}(\tau,\sigma)$ *(eq.* (5.18a)-(5.18c)*). Then, the simplifying order assumptions* (5.21a)-(5.21c) *are equivalent to*

$$\check{B}(\rho):\qquad \int_0^1 \tau^{k-1}\,\mathrm{d}\tau=\frac{1}{k},\qquad k=1,\dots,\rho, \tag{5.22a}$$

$$\check{C}(\eta):\qquad \mathcal{P}\tau^{k-1}=\tau^{k-1},\qquad k=1,\dots,\eta, \tag{5.22b}$$

$$\check{D}(\zeta):\qquad \mathcal{P}^*\tau^{k}=\tau^{k},\qquad k=1,\dots,\zeta. \tag{5.22c}$$

The CSRK order conditions $B(\infty)$ *always hold and* $\check{C},\check{D}$ *are equivalent to the polynomial reproduction property of* $\mathcal{P}$ *and* $\mathcal{P}^*$ *(see Strang–Fix*[a] *conditions [FS69, SF11]).*

a. Also refer to [Lig91, Uns96, BU99, DVB07] for the importance of Strang–Fix conditions in approximation, wavelet and generalized sampling theories.

Proof. The proof is detailed in section D.3 p.292. □

Accuracy order For RPM, the projector $\mathcal{P}$ is self-adjoint. Then, condition $\check{C}(\eta=p)$ implies $\check{D}(\zeta=p-1)$. By consequence, if the RPM projection reproduces polynomials of order (def. 5.1) p. Then, by theorem 5.3, the accuracy order s of its local truncation error (def. 5.4) is at least

$$\boxed{s\geq 2p.} \tag{5.23}$$

5.2.7 Regularity (P1)

The main drawback of piecewise L^2 projection is that the resulting approximations are piecewise discontinuous (blue curves on figure 5.2). We show that step iii) of method 5.1 induces a projector $\mathcal{Q}$ which both restores piecewise continuity and improves the accuracy (a graphical illustration of the method is shown in figure D.1 p.295). Then we compare the approximation properties of $\mathcal{P}$ and $\mathcal{Q}$ in the Hilbert space L^2 (see figures 5.2 and 5.3).

First, we give an explicit construction of the inverse boundary operator $\mathcal{B}^{-1}$ in A_R (i.e. the continuous reconstruction operator complimentary to the multi-derivative boundary analysis functionals $\mathcal{B}^m_\alpha(\cdot)$ used to obtain regularity in the Sobolev space H^k).

Proposition 5.4. *Let $\left\{\psi^m_\alpha(\tau)\right\}$ for $m = 0, \dots, k-1$, $\alpha \in \{0,1\}$ be linear combinations of $\{\phi_n\}_{n=p}^{p+2k-1}$ (spanning the space A_R in (5.8)) satisfying the biorthogonality conditions* [a]

$$\mathcal{B}^m_\alpha(\psi^{m'}_{\alpha'}) = \begin{cases} 1 & \alpha = \alpha' \text{ and } m = m' \\ 0 & \text{otherwise,} \end{cases}, \quad \forall \alpha \in \{0,1\}, \forall m \in \{0, \dots, k-1\}, \tag{5.24}$$

then the synthesis operator $\mathcal{B}^{-1} : \mathbb{R}^{2k} \mapsto A_R$ satisfying $\mathcal{B}\mathcal{B}^{-1} = \mathbf{I}_{2k}$ and $\mathcal{B}^{-1}\mathcal{B} = \mathcal{I}_{A_R}$, is

$$(\mathcal{B}^{-1}\mathbf{u})(\tau) = \sum_{\alpha=0}^{1}\sum_{m=0}^{k-1} \psi^m_\alpha(\tau)\mathbf{u}_{\alpha k+m}, \quad \forall \mathbf{u} \in \mathbb{R}^{2k}. \tag{5.25}$$

a. Two sequences $\{f_m\}$, $\{g_n\}$ are said to be biorthogonal if $\langle f_m \,|\, g_n \rangle = \delta_{mn}$.

Example Let $\{\phi_n\}$ be the orthonormal Legendre polynomials ((C.16) p.286). The corresponding synthesis functions $\left\{\psi^m_\alpha(\tau)\right\}$ are shown on figure 5.1 for projection orders $p \in \{0,1,2\}$ and regularity orders $k \in \{1,2,3\}$. Note that the right boundary functions ($\alpha = 1$) are drawn shifted on $[-1,0]$ to emphasize the global continuity and limited support of boundary functions on $[-1,1]$.

Proposition 5.5. *Step iii) of RPM, def. 5.1, induces a projector $\mathcal{Q} : H^k(\Omega) \to A$, satisfying*

$$\mathcal{Q} = \mathcal{P} \oplus \mathcal{R}, \qquad \text{where} \qquad \mathcal{R} = \mathcal{B}^{-1}\mathcal{B}(\mathcal{I} - \mathcal{P}). \tag{5.26}$$

Proof. The proof is detailed in section D.5 p.294. □

According to (5.24) and (5.26), operator $\mathcal{Q}$ can be written as an integral operator $(\mathcal{Q}f)(\tau) = \int_0^1 K_\mathcal{Q}(\tau,\sigma)f(\sigma)\,\mathrm{d}\sigma$ with kernel

$$K_\mathcal{Q}(\tau,\sigma) = K_\mathcal{P}(\tau,\sigma) + \sum_{\alpha=0}^{1}\sum_{m=0}^{k-1} \psi^m_\alpha(\tau)\left(\delta^{(m)}(\sigma-\alpha) - \frac{\partial^m K_\mathcal{P}}{\partial \tau^m}(\alpha,\sigma)\right). \tag{5.27}$$

where $K_\mathcal{P}$ is defined according to (3.9) p.84 (see table 5.1 p.132 for some examples).

Approximation properties A qualitative study of the approximation properties of operator $\mathcal{Q}$ is shown on figure 5.2. The function to approximate, $\exp(-8\tau)$, is chosen such that, from an ODE viewpoint, the system is both smooth and stiff (with a time constant height times larger than the step-size). On this example, a numerical study of the convergence rate of $\mathcal{Q}$, according to projection order p, and regularity order k, is also shown on figure 5.3.

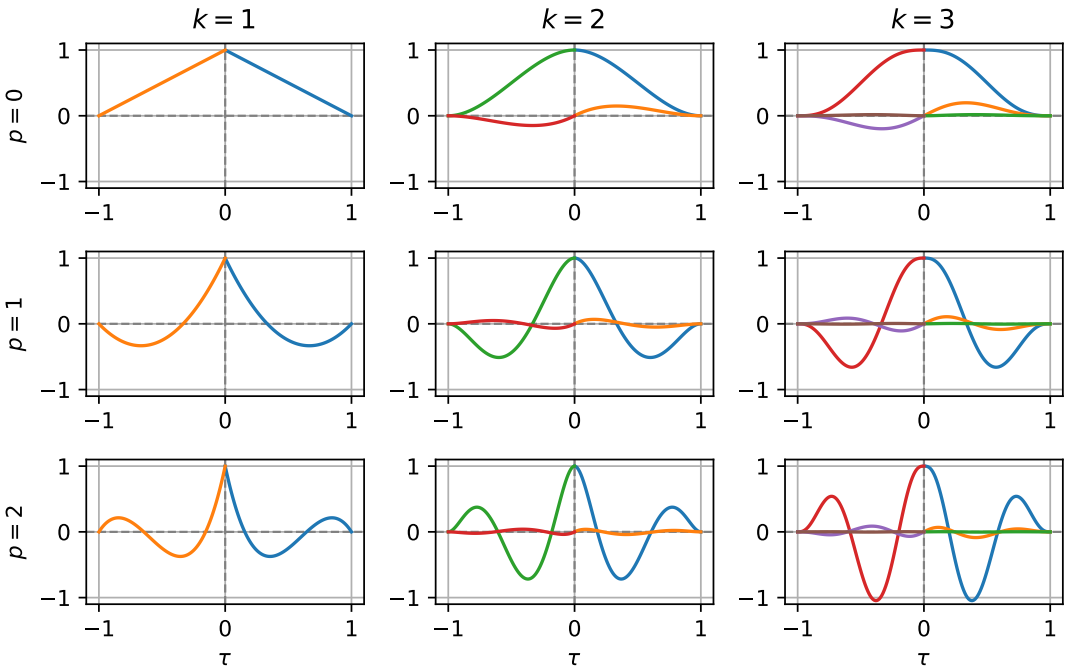
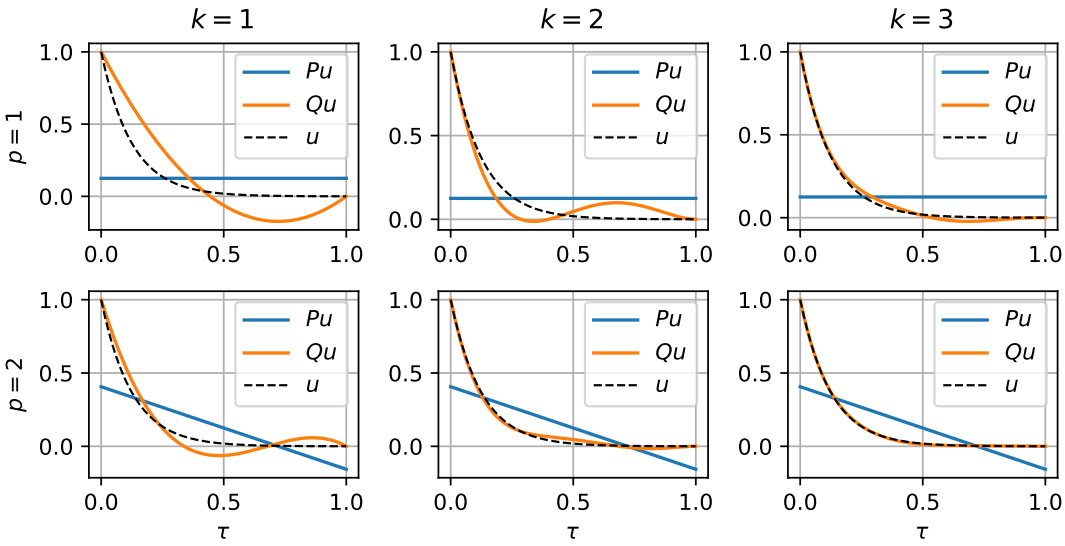
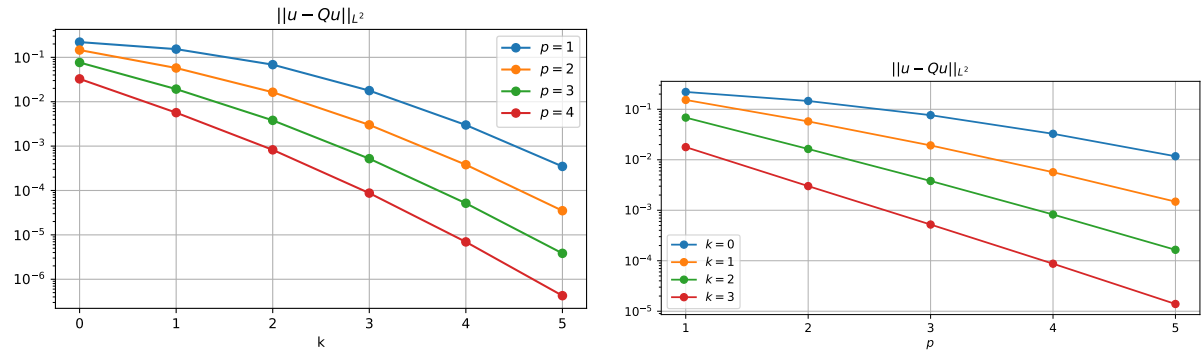


Figure 5.1 – (Polynomial supplementary boundary functions). The basis $\{\psi_\alpha^m(\tau)\}$ is shown for projection order $p \in \{0, 1, 2\}$, and regularity order $k \in \{1, 2, 3\}$. The case $p = 0$ (which corresponds to Hermite splines) is not used in this thesis as the consistency of the time stepping method requires that $p \geq 1$. By construction, these boundary functions act as continuous regularisations of the Dirac delta distributions $\delta^{(m)}$.



(a) Approximated functions.

Figure 5.2 – Comparison of operators $\mathcal{P}$ and $\mathcal{Q}$ to approximate $u(\tau) = \exp(-8\tau)$ for projection order $p \in \{1, 2\}$ and regularity order $k \in \{1, 2, 3\}$. On this example, we clearly see that L^2 projection $\mathcal{P}u$ (in blue) is discontinuous and a crude approximation of the function u. The projection $\mathcal{Q}u$ preserves and refines the approximation $\mathcal{P}u$. It restores $\mathcal{C}^{k-1}$-continuity by interpolating u and its derivatives on the boundary of the interval. We also observe that it reduces the pointwise approximation error: the amplitude of oscillations decreases with increasing k.



(a) k-convergence **(b)** p-convergence

Figure 5.3 – Comparison of k-convergence and p-convergence of RPM(p, k) on the approximation error $\|u - \mathcal{Q}u\|_{L^2}$ for $u(\tau) = \exp(-8\tau)$. We remark on figure 5.3a (k-convergence) that the error for $k = 5$ (for all values of p) is systematically smaller than the error on figure 5.3b (p-convergence) for $p = 5$ (for all values of k).

Relation between projection order p and continuity order k:

We ask the following question:

For a given projection order p, what is the maximal regularity order k such that multi-derivative supplementary boundary conditions (5.11) yield a consistent ODE approximation?

To answer that question, let $\boldsymbol{z}(t = t_0 + h\tau) := \boldsymbol{X}(\tau)$ be an approximate ODE solution and $\boldsymbol{x}(t)$ the exact solution. We remark from CSRK order conditions that we have the local truncation error (see (5.23))

$$\boldsymbol{x}(t_0 + h) = \boldsymbol{z}(t_0 + h) + \mathcal{O}(h^{2p+1}).$$

Then, according backward error analysis theory [HLW06, thm 1.2, p.340], there exists for each time-step a *modified vector field* $\boldsymbol{f}_h$ such that $\boldsymbol{z}$ is locally the exact solution of the modified ODE $\dot{\boldsymbol{z}} = \boldsymbol{f}_h(\boldsymbol{z})$ with

$$\boldsymbol{f}(\boldsymbol{z}) = \boldsymbol{f}_h(\boldsymbol{z}) + \mathcal{O}(h^{2p}).$$

Since $\mathcal{D}^m = \left(\frac{1}{h}\frac{\mathrm{d}}{\mathrm{d}\tau}\right)^m$ and $\mathcal{B}_\alpha^m u = (\mathcal{D}^m u)(\alpha)$, it follows that supplementary boundary conditions yields the approximation

$$\mathcal{B}_\alpha^m\left(\boldsymbol{f}(\boldsymbol{z})\right) = \mathcal{B}_\alpha^m\left(\boldsymbol{f}_h(\boldsymbol{z})\right) + \mathcal{O}(h^{2p-m}).$$

We conclude that, for a small enough step size h, as long as $k \leq 2p$, multi-derivative boundary conditions (5.11) are consistent with the projected vector field up to order $2p - k$.

Peano error kernels and pointwise error

To study the approximation error of operators $\mathcal{P}$ and $\mathcal{Q}$, we use the Peano kernel theorem 5.4 to obtain their respective *Peano error kernels* 5.30 from which numerical bounds and qualitative information can be obtained. Let g be a function sufficiently differentiable such that its Taylor polynomial expansion with remainder may be written for $\tau \in [a, b]$ in the form [5]

$$g(\tau) = g(a) + g'(a)(\tau - a) + \ldots + g^{(d)}(a)\frac{(\tau - a)^d}{d!} + R_d^\tau[g], \quad R_d^\tau[g] = \int_a^b \frac{(\tau - s)_+^d}{d!} g^{(d+1)}(s)\,\mathrm{d}s.$$

5. Using the common notation $(\cdot)_+ = \max(0, \cdot)$.

If an approximation $\mathcal{Q}g$ reproduces polynomials up to degree d, then the residual $g-\mathcal{Q}g = \mathcal{O}(R_d^\tau[g])$ is governed by $g^{(d+1)}$ and an error kernel E which is given by the following theorem.

Theorem 5.4 (Peano kernel theorem [Ise09]). *Let $\Omega = [a,b]$, let L be a linear functional that commutes with the operation of integration, and such that $L[u] = 0$, $\forall u \in \mathbb{P}^d(\Omega)$. Then, for all $g \in \mathcal{C}^{d+1}(\Omega)$*

$$L[g] = \int_a^b E(\sigma) g^{(d+1)}(\sigma)\,\mathrm{d}\sigma, \qquad E(\sigma) := L\left[\frac{(\tau-\sigma)_+^d}{d!}\right] \tag{5.28}$$

and E is called the Peano error kernel *of L.*

In RPM, for projectors $\mathcal{P}$ (resp. $\mathcal{Q}$), we propose to use the continuous family of functionals

$$L_{\mathcal{P}}^\tau[f] := \big((\mathcal{I}-\mathcal{P})f\big)(\tau), \tag{5.29}$$

to measure the pointwise approximation error $e_{\mathcal{P}}(\tau) = f(\tau) - (\mathcal{P}f)(\tau)$ for all $\tau \in [0,1]$.

Definition 5.5. Let $\mathcal{P}$ be a projector with kernel $K_{\mathcal{P}}(\tau,\sigma)$ reproducing polynomials up to degree d. Then, by definition of $\mathcal{P}$, functionals (5.29) satisfy the conditions of theorem 5.4. The associated Peano error kernel is

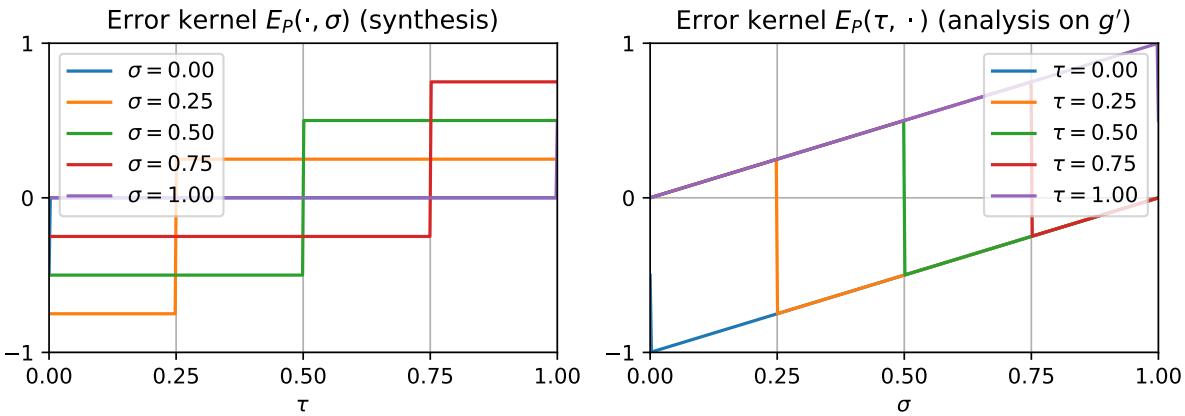
$$E_{\mathcal{P}}(\tau,\sigma) := L_{\mathcal{P}}^\tau\left[\frac{(\tau-\sigma)_+^d}{(d)!}\right] = \frac{(\tau-\sigma)_+^d}{(d)!} - \int_0^1 K_{\mathcal{P}}(\tau,\xi)\frac{(\xi-\sigma)_+^d}{(d)!}\,\mathrm{d}\xi. \tag{5.30}$$

Peano kernels for $\mathcal{P}$ Kernels for projection orders $p \in \{1,2,3\}$ in the Legendre basis are shown on figure 5.4 and Table 5.1 ($\mathcal{P}$ reproduces polynomials up to degree $d = p-1$). We note that the synthesis error kernel $E_{\mathcal{P}}(t,\cdot)$ is always non zero on the boundary $\partial\Omega = \{0,1\}$, this confirms that L^2 projection is always discontinuous on boundaries when $g^{(p)} \neq 0$. Conversely, the analysis error kernel $E_{\mathcal{P}}(\cdot,s)$ always vanishes on the boundary, this means that Legendre projection is blind to the boundary values of the residual term $g^{(p)}$.

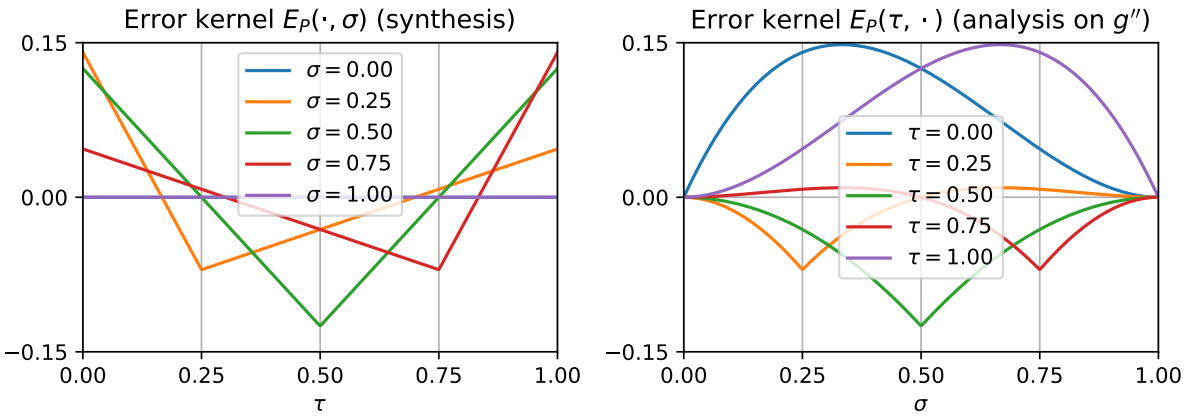
Peano kernels for $\mathcal{Q}$ Corresponding Peano error kernels for operator $\mathcal{Q}$ with $p=1$, $k=1,2$ are shown on figure 5.5 ($\mathcal{Q}$ reproduces polynomials up to degree $d = p+2k-1$). See (5.27) for the definition of $K_{\mathcal{Q}}$. As expected, the error and its derivatives vanishes on the boundary $\partial\Omega$, i.e. the error belongs to the Sobolev space $H_0^k(\Omega) = \Big\{u \in H^k(\Omega) \mid \mathcal{B}u = 0\Big\}$. We also note that the maximum norm of the kernel $E_{\mathcal{Q}}$ is an order of magnitude lower than $E_{\mathcal{P}}$.

Kernels	$K_{\mathcal{P}}(\tau,\sigma)$	$E_{\mathcal{P}}(\tau,\sigma)$
$p=1$	1	$(\tau-\sigma)_+^0 - (1-\sigma)$
$p=2$	$1+P_1(\tau)P_1(\sigma)$	$(\tau-\sigma)_+ - \big(1+(2\tau-1)(2\sigma+1)\big)\frac{1}{2}(1-\sigma)^2$
$p=3$	$\sum_{i=0}^{p-1} P_i(\tau)P_i(\sigma)$	$\frac{(\tau-\sigma)_+^2}{2!} + \left(1+\frac{\sqrt{3}}{2}P_1(\tau)(1+\sigma)+\frac{\sqrt{5}}{10}P_2(\tau)(6\sigma^2+3\sigma+1)\right)\frac{(\sigma-1)^3}{3!}$

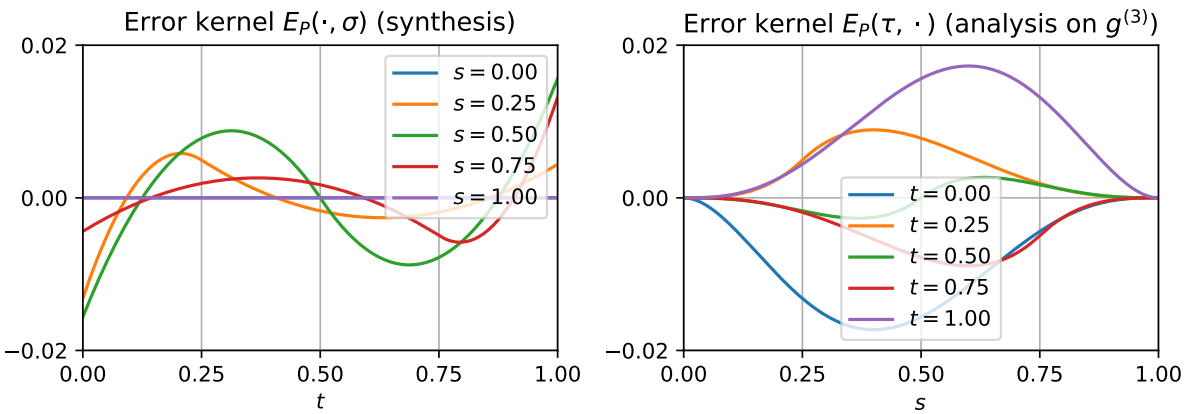
Table 5.1 – Reproducing kernel $K_{\mathcal{P}}$ and Peano error kernel $E_{\mathcal{P}}$ of Legendre orthogonal projector.



(a) Projection order $p = 1$

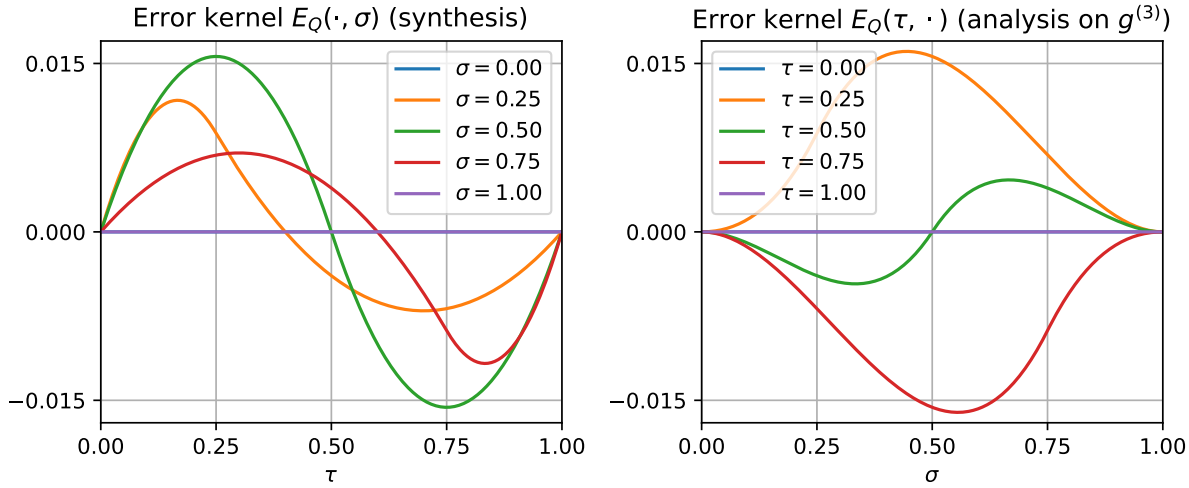


(b) Projection order $p = 2$

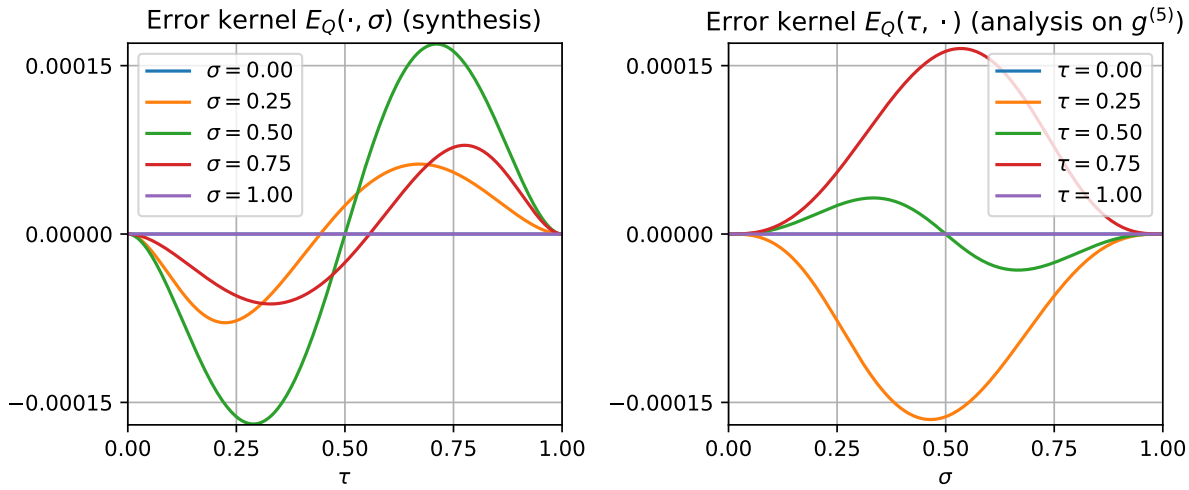


(c) Projection order $p = 3$

Figure 5.4 – Peano error kernels $E_{\mathcal{P}}(\tau, \sigma)$ for projector $\mathcal{P}$ with projection order $p \in \{1, 2, 3\}$. As $\mathcal{P}$ does not handle regularity, expected discontinuities of kernels appear at $\tau = \sigma$ (the Sobolev regularity of $E_{\mathcal{P}}(\cdot, \sigma)$ $E_{\mathcal{P}}(\tau, \cdot)$ is $p - 1$). We notice in the synthesis column that the largest approximation errors are more likely to appear towards the interval boundaries. The maximal error decreases by an order of magnitude as p is incremented. We notice, on the analysis column, that kernels are all zero on boundaries, meaning that, at these points, the error might be arbitrarily high (which is confirmed on the synthesis kernels). Conversely they have maximal weight towards the center of the interval. These observations show that projector $\mathcal{P}$ is biased towards reducing errors close to the center of the interval.



(a) Regularity order $k = 1$, projection order $p = 1$



(b) Regularity order $k = 2$, projection order $p = 1$

Figure 5.5 – Peano error kernels $E_{\mathcal{Q}}(\tau, \sigma)$ for operator $\mathcal{Q}$ with $p = 1$ and regularity $k \in \{1, 2\}$. Comparing these error kernels to those of projection $\mathcal{P}$ in figure 5.4, we notice that (in the synthesis column) the error (and it derivatives when increasing k) now vanishes on the boundaries and that the magnitude order of the error is also much smaller. However, in the analysis column, we notice that the maximal weight is still towards the center of the interval. Although projection $\mathcal{Q}$ reduces the boundary error, this means that the error might still become high near the boundaries. A more uniform handling of the point-wise error would require the use of a different basis, for example Chebyshev polynomials. Unfortunately, this choice is not an option since the uniform weight of the L^2 inner product is already dictated by the power-balance.

5.3 Analysis of RPM for pH-DAE

In this section, we consider existence, uniqueness and accuracy of solutions for pH-DAE of index 1 ((1.16) p.14) discretized using RPM (def. 5.2 p.123). A general theory is still missing. Results below are preliminary steps towards this goal. In subsection 5.3.1, we recall order reduction for stiff ODE, while subsection 5.3.2 is dedicated to existence and uniqueness of solutions.

5.3.1 Accuracy and stage order for stiff ODE and DAE

In the theory of Runge-Kutta methods applied to stiff ODE and DAE (i.e. when the time-constants of the vector field are much smaller than the step size h), it is known [HW96, thm 1.1 p.380] that point-wise super-convergence on the time stepping grid $x(t_n = hn)$ is lost. We recall that for RPM(p, k) the local truncation error accuracy is in $\mathcal{O}(h^{2p})$, see (5.23) p.128. In the case of RPM(p, k) for pH-DAE, the stiff accuracy falls back to the level of stage order conditions $C(\eta)$ (see eq. (5.22b) p.128) which reduces to $\mathcal{O}(h^r)$ with $r = \min(2p - 1, p) = p$. This corresponds to the polynomial reproduction property of the projector and thus to the accuracy for all values of the solutions between time-stepping instants, not just on the boundaries of each time interval.

5.3.2 Existence and uniqueness of solutions

First, we establish (naive) existence and uniqueness conditions for solving DAE using fixed-point iteration. These conditions are tractable, but usually too restrictive. However we know that if the fixed-point converges, then Newton iteration also converges. Second, we establish pH-specific conditions to ensure a DAE is of index-1. Finally, we propose partial results for the resolution of Newton iteration in the case of projected pH-DAE.

Fixed-point convergence

We consider the semi-explicit Hybrid Dirac structure formulation (2.18) p.55 of pH-DAE, parameterized by tree currents $\boldsymbol{i}_T$ and link voltaged $\boldsymbol{v}_L$ rewritten as a fixed-point map $\boldsymbol{G} : F \to F$,

$$\begin{bmatrix} \boldsymbol{i}_T \\ \boldsymbol{v}_L \end{bmatrix} = \boldsymbol{G}\left(\begin{bmatrix} \boldsymbol{i}_T \\ \boldsymbol{v}_L \end{bmatrix}\right) := \begin{bmatrix} \mathbf{0} & -\mathbf{C}^\mathsf{T} \\ \mathbf{C} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{V}(\boldsymbol{i}_T) \\ \boldsymbol{I}(\boldsymbol{v}_L) \end{bmatrix}, \tag{5.31}$$

where $F \subset L^2(\Omega, \mathbb{R}^{(n_T+n_L)})$ is the projection space (see def. 5.2 p.123) and $\boldsymbol{V}, \boldsymbol{I}$ are operators on F standing for *projected component laws* which yield tree voltages $\boldsymbol{v}_T$ and cotree currents $\boldsymbol{i}_L$.

A sufficient condition for existence and uniqueness of solutions is given by

Theorem 5.5. *Let $(L_{\boldsymbol{V}}, L_{\boldsymbol{I}}, L_{\mathbf{C}})$ be the Lipschitz constants of operators $(\boldsymbol{V}, \boldsymbol{I}, \mathbf{C}^\mathsf{T}\mathbf{C})$ for the L^2 norm. If $L_{\boldsymbol{V}}\, L_{\boldsymbol{I}}\, L_{\mathbf{C}} < 1$, then the fixed-point (5.31) converges to a unique solution.*

Proof. Rewrite the iterated map $\boldsymbol{G}^2$ in separated variables by composing operators as

$$\begin{aligned} \boldsymbol{i}_T &= \boldsymbol{G}_I(\boldsymbol{i}_T) = (-\mathbf{C}^\mathsf{T} \circ \boldsymbol{I} \circ \mathbf{C} \circ \boldsymbol{V})(\boldsymbol{i}_T), \\ \boldsymbol{v}_L &= \boldsymbol{G}_V(\boldsymbol{v}_L) = (\mathbf{C} \circ \boldsymbol{V} \circ (-\mathbf{C}^\mathsf{T}) \circ \boldsymbol{I})(\boldsymbol{v}_L). \end{aligned}$$

It follows that we have the Lipschitz bounds $\left\|\boldsymbol{G}_I(\boldsymbol{i}_1) - \boldsymbol{G}_I(\boldsymbol{i}_2)\right\| \leq L_{\boldsymbol{I}} L_{\boldsymbol{V}} \left\|\mathbf{C}^\mathsf{T}\mathbf{C}\right\| \|\boldsymbol{i}_1 - \boldsymbol{i}_2\|$, and $\left\|\boldsymbol{G}_V(\boldsymbol{v}_1) - \boldsymbol{G}_V(\boldsymbol{v}_2)\right\| \leq L_{\boldsymbol{I}} L_{\boldsymbol{V}} \left\|\mathbf{C}\mathbf{C}^\mathsf{T}\right\| \|\boldsymbol{v}_1 - \boldsymbol{v}_2\|$ (where $\mathbf{C}^\mathsf{T}\mathbf{C} \equiv \mathbf{C}^\mathsf{T} \circ \mathbf{C}$). Finally, since $L_{\mathbf{C}} = \left\|\mathbf{C}\mathbf{C}^\mathsf{T}\right\| = \left\|\mathbf{C}^\mathsf{T}\mathbf{C}\right\|$, then, convergence of the map $\boldsymbol{G}^2$ to a unique fixed point follows from the Banach fixed-point theorem under the contractivity condition $L_{\boldsymbol{V}}\, L_{\boldsymbol{I}}\, L_{\mathbf{C}} < 1$. □

Example 5.1 (parallel RLC)**.** We consider a parallel RLC with orthogonal projector $\mathcal{P} : L^2(\Omega) \to \mathbb{P}^0(\Omega)$, $\mathcal{P} = |\,1\rangle\langle 1\,|$ and $(i_C, v_L, v_R) \in \mathbb{P}^0(\Omega)^3$, governed by

$$\begin{bmatrix} i_C \\ v_L \\ v_R \end{bmatrix} = \begin{bmatrix} 0 & -1 & -1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} V_C(i_C) \\ I_L(v_L) \\ I_R(v_R) \end{bmatrix}, \quad \text{where} \quad \begin{cases} V_C(i_C) = \mathcal{P}\left(t \mapsto \frac{1}{C}\left(q_0 + h\int_0^t i_C(s)\,\mathrm{d}s\right)\right), \\ I_L(v_L) = \mathcal{P}\left(t \mapsto \frac{1}{L}\left(\phi_0 + h\int_0^t v_L(s)\,\mathrm{d}s\right)\right), \\ I_R(v_R) = \mathcal{P}\left(t \mapsto v_R/R\right). \end{cases}$$

We can show that $\|V_C\| = \frac{h}{2C}$, $\|I_L\| = \frac{h}{2L}$, $L_R = \frac{1}{R}$, $\left\|\mathbf{C}^\mathsf{T}\mathbf{C}\right\| = 2$ so that a sufficient convergence condition is given by $\frac{h}{2C}\max\left(\frac{h}{2L}, \frac{1}{R}\right) 2 < 1$ i.e.

$$\max\left(\frac{h^2}{2LC}, \frac{h}{RC}\right) < 1.$$

Nonlinear extensions of this example follow by replacing the linear conductance law $I_R(\cdot)$ of the resistor by a nonlinear one where the Lipschitz constant becomes $L_R = \sup\left|I_R'\right|$.

These convergence conditions are easy to obtain but unfortunately, they are not tight. As soon as algebraic components are present in both tree and link branches, convergence conditions are dominated by algebraic components (for which Lipschitz constant do not depend on the step size h): it is not possible to adapt h anymore to obtain convergence. For example adding a serial resistor R_2 to the parallel RLC leads to the condition

$$\max\left(\frac{h}{2C}, R_2\right) \cdot \max\left(\frac{h}{2L}, \frac{1}{R}\right) \cdot 2 < 1.$$

Then, if $R_2 > h/2C$, $1/R > h/2L$ and $2R_2/R > 1$, this condition does not guarantee the convergence of the fixed-point (although for linear systems a solution always exists).

Remark 5.2 (Fixed point vs Newton)**.** Clearly, we need a better alternative to the fixed-point method. As noted by [Deu11, p.289] (see also [Deu87]), for stiff and DAE systems, the use of implicit discretization methods solves only one half of the problem, the choice of iterative scheme is at least equally important. Proofs based on the Newton-Kantorovich theorem rather than the Banach fixed-point theorem are more difficult but yield tighter estimates (see [Deu11, thm 6.3, p.297] and [HW96, thm 3.5, p.397]). Indeed, classical existence and uniqueness theory (based on fixed-point iteration) is bounded by the Lipschitz constant of the vector field whereas Newton iteration converges in one iteration for linear systems and restores the full existence domain $h \in [0, \infty)$ for A-stable and L-stable methods.

Index-1 DAE

In this section, we consider the index-1 DAE hypothesis (see (1.16) p.14 and remark 1.9 p.34). In the semi-explicit pH-DAE formulation (1.52) p.34, the algebraic function

$$\boldsymbol{g}_{\mathbf{w}}(\mathbf{w}) = \mathbf{w} - \mathbf{J}_{\mathbf{w}}\boldsymbol{z}(\mathbf{w}), \tag{5.32}$$

is assumed to be invertible, where $\mathbf{J}_{\mathbf{w}}$ is a skew symmetric matrix and $\boldsymbol{z}$ a passive law ($\boldsymbol{z}(\mathbf{w})\cdot\mathbf{w} \geq 0$) so that existence and uniqueness of solutions follows from classical ODE theory (see thm 1.1 p.8). Exploiting the particular structure of semi-explicit pH-DAE, we establish the following sufficient conditions for the invertibility of $\boldsymbol{g}_{\mathbf{w}}$ in the following lemma

Lemma 5.1. *If either of the following conditions is satisfied in equation* (5.32)

C1. $\mathbf{J_w} = 0$, *or*

C2. $\boldsymbol{z}'(\mathbf{w})$ *is symmetric positive definite* ($\boldsymbol{z}'(\mathbf{w}) = \boldsymbol{z}'(\mathbf{w})^\mathsf{T} \succ 0$), *or*

C3. $\mathbf{J_w}\boldsymbol{z}'$ *satisfies conditions (C2) of lemma 5.3.*

Then, $\boldsymbol{g}_\mathbf{w}$ *is invertible and the associated pH-DAE* (1.52) *p.34 has differential index*-1.

Proof. If condition (C1) is satisfied, then $\boldsymbol{g}_\mathbf{w}$ reduces to the identity function which is obviously invertible. If condition (C2) is satisfied, denote $\mathbf{Q} = \mathbf{Q}^\mathsf{T} = \mathbf{z}'(\mathbf{w}) \succ 0$, and $\mathbf{A} = \boldsymbol{g}'_\mathbf{w} = \mathbf{I} - \mathbf{J_w}\mathbf{Q}$. Invertiblity of $\boldsymbol{g}'_\mathbf{w}$ follows from lemma 5.2 below. If condition (C3) is satisfied then invertibility of $\boldsymbol{g}'_\mathbf{w}$ follows from lemma 5.3 below with $\mathbf{M} = \mathbf{J}_w\boldsymbol{z}'$. Then, the invertibility of function $\boldsymbol{g}_\mathbf{w}$ follows from the invertibility of its Jacobian $\boldsymbol{g}'_\mathbf{w}$ using the implicit function theorem. □

We note some common cases where the conditions of lemma 5.1 are satisfied:

- Condition (C1) is often naturally satisfied because of the circuit topology. Note that it is possible to decouple instantaneous algebraic loops (forcing $\mathbf{J_w} = 0$) by adding (topologically well chosen) parasitic capacitances and inductances in the network.

- Condition (C2) is satisfied when algebraic components are one-port elements ($\boldsymbol{z}'$ is diagonal) and each component is incrementally passive (i.e. monotonically increasing $\boldsymbol{z}' \succ 0$). In particular this is the case for resistors and diodes.

The following lemmas are used in the proof of lemma 5.1.

Lemma 5.2. *Let* $\mathbf{A} = \mathbf{I} - \mathbf{JQ}$ *with* $\mathbf{J}$ *a real skew symmetric matrix and* $\mathbf{Q} = \mathbf{Q}^\mathsf{T} \succ 0$ *real positive definite. Then* $\mathbf{A}$ *is invertible with positive determinant* $\det \mathbf{A} > 0$.

Note that the form $\mathbf{I} - \mathbf{JQ}$ also appears when solving projected pH-DAE using Newton iteration.

Proof. Since $\mathbf{Q} = \mathbf{Q}^\mathsf{T} \succ 0$, there exists a real invertible upper triangular matrix $\mathbf{M}$ with positive diagonal such that $\mathbf{Q} = \mathbf{M}^\mathsf{T}\mathbf{M}$ (Cholesky factorization). Denote $\mathbf{A_M} = \mathbf{MAM}^{-1}$ the similarity transform of $\mathbf{A}$ ($\det \mathbf{A} = \det \mathbf{A_M}$) and $\mathbf{J_M} = \mathbf{MJM}^\mathsf{T}$. The result follows from the relations

$$\det \mathbf{A} = \det \mathbf{A_M} = \det(\mathbf{MM}^{-1} - \mathbf{MJM}^\mathsf{T}\mathbf{MM}^{-1}) = \det(\mathbf{I} - \mathbf{J_M}) > 0,$$

where the last inequality follows from skew-symmetry of $\mathbf{J_M}$. □

Lemma 5.3. *Let* $\mathbf{A} = \mathbf{I} - \mathbf{M}$ *with* $\mathbf{M}$ *a diagonalizable real square matrix whose real spectrum is denoted by* $\sigma_R(\mathbf{M})$ *and complex spectrum* $\sigma_C(\mathbf{M})$. *Then the following results holds*

C1. If $\lambda < 1$, $\forall \lambda \in \sigma_R(\mathbf{M})$, *and if* $\lambda < \frac{1+|\lambda|^2}{2}$, $\forall \lambda \in \sigma_C(\mathbf{M})$ *then* $\det \mathbf{A} > 0$,

C2. If $\lambda \neq 1$, $\forall \lambda \in \sigma_R(\mathbf{M})$ *and if* $\lambda \neq \frac{1+|\lambda|^2}{2}$, $\forall \lambda \in \sigma_C(\mathbf{M})$, *then* $\mathbf{A}$ *is invertible.*

Proof. Let $\mathbf{M} = \mathbf{U}^{-1}\Lambda\mathbf{U}$ be the eigenvalue decomposition of $\mathbf{M}$. Denote $\mathbf{A_U} = \mathbf{UAU}^{-1} = \mathbf{I} - \Lambda$ the similarity transform of $\mathbf{A}$. The determinant of $\mathbf{A}$ is given by the product of the eigenvalues

$$\det \mathbf{A} = \det \mathbf{A_U} = \prod_{\lambda \in \sigma_C(\Lambda)} \underbrace{(1-\lambda)(1-\overline{\lambda})} \prod_{\lambda_R \in \sigma_R(\Lambda)} (1-\lambda_R).$$

If condition (C1) is satisfied then, the first term $(1-\lambda)(1-\overline{\lambda}) = 1 - 2\Re\mathrm{e}(\lambda) + |\lambda|^2$ and the second term $(1-\lambda_R)$ are positive so $\det \mathbf{A} > 0$. If condition (C2) is satisfied then, since both terms are nonzero $\det \mathbf{A}$ is non zero and $\mathbf{A}$ is invertible. □

Newton iteration for pH-DAE with projection order $p = 1$

We investigate the implementation of step ii) of RPM 5.2 p.123 using Newton iteration for the simplest case ($p = 1$, $k = 0$). We look for ways to obtain practical existence/uniqueness conditions. We consider autonomous pH-DAE discretized using the projector $\mathcal{P} = |\,1\rangle\,\langle 1\,|$ (Since $\mathcal{P}$ is an averaging projector, we use the notation $\bar{\mathbf{f}} := \mathcal{P}\boldsymbol{f}$ for all variables) with $\mathbf{J}$ skew-symmetric

$$\begin{bmatrix} \bar{\mathbf{f}} \\ \bar{\mathbf{w}} \end{bmatrix} = \mathbf{J} \begin{bmatrix} \bar{\mathbf{e}}(\bar{\mathbf{f}}) \\ \bar{\mathbf{z}}(\bar{\mathbf{w}}) \end{bmatrix}, \qquad \begin{cases} \bar{\mathbf{e}}(\bar{\mathbf{f}}) = \displaystyle\int_0^1 \nabla H\left(\mathbf{x}_0 + h\int_0^\tau \bar{\mathbf{f}}\,\mathrm{d}s\right)\mathrm{d}\tau = \overline{\nabla} H(\mathbf{x}_0, h\bar{\mathbf{f}}) \\ \bar{\mathbf{z}}(\bar{\mathbf{w}}) = \displaystyle\int_0^1 \boldsymbol{z}(\bar{\mathbf{w}})\,\mathrm{d}\tau = \boldsymbol{z}(\bar{\mathbf{w}}). \end{cases} \tag{5.33}$$

We look for a solution $\mathbf{a}^\star$ of the algebraic equation $\boldsymbol{F}(\mathbf{a}^\star) = \mathbf{0}$, defined by the Newton function

$$\boldsymbol{F}(\mathbf{a}) := \mathbf{a} - \mathbf{J}\boldsymbol{b}(\mathbf{a}), \qquad \text{where} \qquad \mathbf{a} := \begin{bmatrix} \bar{\mathbf{f}} \\ \bar{\mathbf{w}} \end{bmatrix}, \qquad \boldsymbol{b}(\mathbf{a}) := \begin{bmatrix} \bar{\mathbf{e}}(\bar{\mathbf{f}}) \\ \bar{\mathbf{z}}(\bar{\mathbf{w}}) \end{bmatrix}. \tag{5.34}$$

To this end, we use the simplified Newton iteration

$$\Delta\mathbf{a}^k = -(\boldsymbol{F}_0')^{-1}\boldsymbol{F}(\mathbf{a}^k), \qquad \mathbf{a}^{k+1} = \mathbf{a}^k + \Delta\mathbf{a}^k, \tag{5.35}$$

where the Jacobian of $\boldsymbol{F}$ evaluated at $\mathbf{a}_0 = (\mathbf{0}, \mathbf{w}_0)$ [6] is denoted

$$\boldsymbol{F}_0' := \boldsymbol{F}(\mathbf{a}_0) = \mathbf{I} - \mathbf{J}\mathbf{Q}, \qquad \text{with} \qquad \mathbf{Q} = \begin{bmatrix} \frac{h}{2}\nabla^2 H(\mathbf{x}_0) & \mathbf{0} \\ \mathbf{0} & \boldsymbol{z}'(\mathbf{w}_0) \end{bmatrix}. \tag{5.36}$$

Existence and uniqueness conditions for simplified Newton iteration (i.e. when the Jacobian $\boldsymbol{F}'(\mathbf{x}_k)$ is approximated by $\boldsymbol{F}'(\mathbf{x}_0)$) are given by the following theorem

Theorem 5.6 (Newton-Kantorovich theorem for simplified Newton iteration [Deu11])**.** *Let $\boldsymbol{F} : D \to \mathbb{R}^n$ be a continuously differentiable mapping with $D \subset \mathbb{R}^n$ open and convex. Let $\mathbf{x}_0 \in D$ denote a given starting point. Assume that*

$$\boldsymbol{F}'(\mathbf{x}_0) \text{ is invertible with } \boldsymbol{\Gamma}_0 := \boldsymbol{F}'(\mathbf{x}_0)^{-1}, \tag{5.37a}$$

$$\left\|\boldsymbol{\Gamma}_0(\boldsymbol{F}'(\mathbf{x}) - \boldsymbol{F}'(\mathbf{x}_0))\right\| \le \omega_0\|\mathbf{x} - \mathbf{x}_0\| \text{ for all } \mathbf{x} \in D. \tag{5.37b}$$

$$h_0 := \omega_0\left\|\Delta\mathbf{x}^0\right\| \le 1/2, \text{ with } \Delta\mathbf{x}^0 = -\boldsymbol{\Gamma}_0\boldsymbol{F}(\mathbf{x}_0). \tag{5.37c}$$

Define $t^- = 1 - \sqrt{1 - 2h_0}$, $\rho = t^-/\omega_0$. Moreover, assume that $S(\mathbf{x}, \rho) = \{\mathbf{x} \,|\, \|\mathbf{x} - \mathbf{x}_0\| \le \rho\} \subset D$. Then the simplified Newton iterates $\{\mathbf{x}_k\}$ remain in the ball $S(\mathbf{x}, \rho)$ and converge to some $\mathbf{x}^\star$ with $\boldsymbol{F}(\mathbf{x}^\star) = 0$.

Towards existence and uniqueness (a sketch of proof) Our goal is to obtain simple conditions on the projected pH-DAE (5.33) so that conditions (5.37a)-(5.37c) are satisfied in order to make Newton iteration convergent. We restrict the study to the frequent case where nonlinearities are separable and monotone by assuming that

$$\nabla^2 H \text{ and } \boldsymbol{z}' \text{ are diagonal positive definite.} \tag{5.38}$$

6. Assuming the consistent initial condition $\mathbf{w}_0 = \mathbf{J}_{\mathbf{wx}}\nabla H(\mathbf{x}_0) + \mathbf{J}_{\mathbf{w}}\boldsymbol{z}(\mathbf{w}_0)$.

Under these hypotheses, we show that (5.37a) is satisfied. However further work is required to establish a proportionality relation between ω_0 and the step size h in (5.37b) so that Newton iteration (5.35) is contractive for sufficiently small h. A sketch of proof is reproduced thereafter.
Sketch of proof.

1. Since $\nabla^2 H$ and $\boldsymbol{z}'$ are positive definite. Then, according to lemma 5.2, $\boldsymbol{F}_0'$ is invertible with positive determinant $\det \boldsymbol{F}_0' > 0$, so that (5.37a) is always satisfied.
2. Since $\nabla^2 H$ and $\boldsymbol{z}'$ are diagonal. Denote $\mathbf{M} = \sqrt{\mathbf{Q}}$ in (5.36). Define $\tilde{\mathbf{a}} = \mathbf{M}\mathbf{a}$ and introduce the *affine similarity transform* $\boldsymbol{G}(\tilde{\mathbf{a}}) = \mathbf{M}\boldsymbol{F}(\mathbf{M}^{-1}\tilde{\mathbf{a}})$. For the transformed problem $\boldsymbol{G}(\tilde{\mathbf{a}}) = 0$, we have the jacobian

$$\boldsymbol{G}_0' := \boldsymbol{G}_0'(\tilde{\mathbf{a}}_0) = \mathbf{I} - \mathbf{J}_{\mathbf{M}}, \qquad \text{where} \qquad \mathbf{J}_{\mathbf{M}} = \mathbf{M}\mathbf{J}\mathbf{M}^{\mathsf{T}}.$$

Denote $\mathbf{D}_{\mathbf{x}} = \sqrt{\nabla^2 H(\mathbf{x}_0)}$ and $\mathbf{D}_{\mathbf{w}} = \sqrt{\boldsymbol{z}'(\mathbf{w}_0)}$, so that

$$\boldsymbol{G}_0' = \begin{bmatrix} \mathbf{I} - \frac{h}{2}\mathbf{D}_{\mathbf{x}}\mathbf{J}_{\mathbf{x}}\mathbf{D}_{\mathbf{x}}^{\mathsf{T}} & +\sqrt{h}\mathbf{D}_{\mathbf{x}}\mathbf{J}_{\mathbf{wx}}^{\mathsf{T}}\mathbf{D}_{\mathbf{w}}^{\mathsf{T}} \\ -\sqrt{h}\mathbf{D}_{\mathbf{w}}\mathbf{J}_{\mathbf{wx}}\mathbf{D}_{\mathbf{x}}^{\mathsf{T}} & \mathbf{I} - \mathbf{D}_{\mathbf{w}}\mathbf{J}_{\mathbf{w}}\mathbf{D}_{\mathbf{w}}^{\mathsf{T}} \end{bmatrix} = \begin{bmatrix} \mathbf{I} - \mathcal{O}(h) & \mathcal{O}(\sqrt{h}) \\ -\mathcal{O}(\sqrt{h}) & \mathcal{O}(1) \end{bmatrix}.$$

Using the determinant identity for block matrices

$$\det \left[\begin{smallmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{smallmatrix}\right] = \det(\mathbf{D}) \det\left(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}\right),$$

it follows that

$$\begin{aligned} \det \mathbf{G}_0' &= \det\left(\mathbf{I} - \mathbf{D}_{\mathbf{w}}\mathbf{J}_{\mathbf{w}}\mathbf{D}_{\mathbf{w}}^{\mathsf{T}}\right) \det\left(\mathbf{I} - \frac{h}{2}\mathbf{D}_{\mathbf{x}}\mathbf{J}_{\mathbf{x}}\mathbf{D}_{\mathbf{x}}^{\mathsf{T}} + h\mathbf{D}_{\mathbf{x}}\mathbf{J}_{\mathbf{wx}}^{\mathsf{T}}\mathbf{D}_{\mathbf{w}}^{\mathsf{T}}\left(\mathbf{I} - \mathbf{D}_{\mathbf{w}}\mathbf{J}_{\mathbf{w}}\mathbf{D}_{\mathbf{w}}^{\mathsf{T}}\right)^{-1}\mathbf{D}_{\mathbf{w}}\mathbf{J}_{\mathbf{wx}}\mathbf{D}_{\mathbf{x}}^{\mathsf{T}}\right) \\ &= \mathcal{O}(1)\det\left(\mathbf{I} + h\bar{\mathbf{A}}\right). \end{aligned}$$

with $\bar{\mathbf{A}} = \frac{1}{2}\mathbf{D}_{\mathbf{x}}\mathbf{J}_{\mathbf{x}}\mathbf{D}_{\mathbf{x}}^{\mathsf{T}} + \mathbf{D}_{\mathbf{x}}\mathbf{J}_{\mathbf{wx}}^{\mathsf{T}}\mathbf{D}_{\mathbf{w}}^{\mathsf{T}}\left(\mathbf{I} - \mathbf{D}_{\mathbf{w}}\mathbf{J}_{\mathbf{w}}\mathbf{D}_{\mathbf{w}}^{\mathsf{T}}\right)^{-1}\mathbf{D}_{\mathbf{w}}\mathbf{J}_{\mathbf{wx}}\mathbf{D}_{\mathbf{x}}^{\mathsf{T}}$.
Note that, for h sufficiently small, $\det(\mathbf{I} + h\bar{\mathbf{A}}) \approx 1 + h \operatorname{tr}\bar{\mathbf{A}} + \mathcal{O}(h^2)$, so that

$$\det \mathbf{G}_0' \approx \mathcal{O}(1)(1 + h\operatorname{tr}\bar{\mathbf{A}} + \mathcal{O}(h^2)).$$

Unfortunately, this approach is not sufficient to make ω_0 proportional to h in (5.37b). ■

5.4 Implementation choices

In order to make RPM(p, k) from methods 5.1, 5.2 p.122-123, a practical numerical method (implemented on a computer with finite memory and computation time), we need to adress the following three subproblems:

a) Numerical methods in step (ii) to compute projection coefficients (see (5.7) p.122) such as

$$\widehat{\mathbf{f}}_i = \left\langle \phi_i, \boldsymbol{f}(\boldsymbol{X}, \boldsymbol{u}) \right\rangle,$$

b) A numerical solver in step (ii) for implicit equations of the form (see (5.13) p.123) (for a given $\boldsymbol{u}$)

$$\delta \boldsymbol{X} = \mathcal{P} \boldsymbol{f}(\boldsymbol{X}, \boldsymbol{u}),$$

c) A procedure to compute multiderivatives in step (iii) (see (5.11) p.122) such as

$$\mathcal{B}\widetilde{\delta \boldsymbol{X}} = \mathcal{B}\boldsymbol{f}(\widetilde{\boldsymbol{X}}, \boldsymbol{u}).$$

In this section, we detail problems (a) and (b). For problem (a), we propose both particular efficient closed-form projections results in subsection 5.4.1 and general-purpose projections based on numerical quadratures in subsection 5.4.2. Problem b) is addressed in subsection 5.4.3. For problem (c) we use symbolic differentiation: the computation of multi-variate derivatives and elementary differentials is detailed in appendix B.3 p.278.

Exact solution	$\dot{\boldsymbol{x}}(\tau) = \boldsymbol{f}(\boldsymbol{x}(\tau))$
$\downarrow$	$\downarrow$
Projection space	$\delta \boldsymbol{X}(\tau) = \mathcal{P}\boldsymbol{f}(\boldsymbol{X}(\tau))$
$\downarrow$	$\downarrow$
Coefficient space	$\widehat{\delta \boldsymbol{X}}_i = \{\widehat{\boldsymbol{f} \circ \boldsymbol{X}}\}_i = \langle \phi_i, \boldsymbol{f} \circ \boldsymbol{X} \rangle$

Table 5.2 – (RPM) principle of the time discretisation approach.

Hypothesis For problem (a), in this thesis, input functions $\boldsymbol{u}$ are assumed to belong to a space such that projection coefficients $\widehat{\mathbf{u}}_i = \langle \phi_i, \boldsymbol{u} \rangle$ are exactly computable. Moreover $\boldsymbol{f}, \boldsymbol{g}$ are most of the time separable functions of $\mathbf{x}, \mathbf{u}$. By consequence, we only present computational methods to find the projection coefficients $\widehat{\mathbf{f}}_i = \left\langle \phi_i, \boldsymbol{f}(\boldsymbol{x}) \right\rangle$.

5.4.1 Closed-form projection results for nonlinear maps of affine functions

Here, we give an explicit formula to compute polynomial projection coefficients (e.g. Legendre expansions) of $\boldsymbol{f} \circ \boldsymbol{x}$ when $\boldsymbol{f}$ is nonlinear and $\boldsymbol{x}$ is affine. We assume that $\boldsymbol{f}(\mathbf{x})$ is a *separable function*[7] of $x_1, \ldots, x_n$ with known anti-derivatives so that we only need to consider the scalar case $\widehat{f}_i = \left\langle \phi_i, f(x(\tau)) \right\rangle$.

7. generalisation to multivariate

Typical usage for PHS concerns *both* differential and algebraic component laws of one-port elements

- $\nabla H(\mathbf{x}) = [H_1'(x_1), \dots, H_n'(x_n)]^\mathsf{T}$ for separable Hamiltonians $H(\mathbf{x}) = \sum_i H_i(x_i)$,
- and $\boldsymbol{z}(\mathbf{w}) = [z_1(w_i), \dots, z_n(w_n)]^\mathsf{T}$ for separable nonlinear algebraic constraints

Theorem 5.7 (Polynomial expansion)**.** *Let $\Omega = [0,1]$, let $x(\tau) = x_0 + \tau\delta x \in \mathbb{P}^1(\Omega)$, let $f : \mathbb{R} \mapsto \mathbb{R}$ be a function with anti-derivatives $f^{[m]}$ known up to order n and let $\{L_n\}$ be a sequence of polynomials with $\deg L_n = n$ and $\langle L_n, 1\rangle = 0$ for all $n > 0$. Then, the projection coefficients of $f \circ x$ noted $\widehat{\{f \circ x\}}_n$ and defined by*

$$\widehat{\{f \circ x\}}_n := \int_0^1 L_n(\tau) f\left(x(\tau)\right) \mathrm{d}\tau, \qquad n \in \mathbb{N}. \tag{5.39}$$

have the following finite closed-form expressions using the (known) anti-derivatives of f

$$\widehat{\{f \circ x\}}_n = \begin{cases} \displaystyle\sum_{k=0}^{n} \frac{(-1)^k}{(\delta x)^{k+1}} \left[L_n^{(k)}(s) f^{[k+1]}(x(s))\right]_0^1, & \delta x \neq 0, \\ f(x_0) & \delta x = 0, n = 0, \\ 0 & \delta x = 0, n > 0. \end{cases} \tag{5.40}$$

Proof. The proof is shown in appendix D.6. □

Some applications of this theorem are illustrated by the following two examples.

Example 5.2 (Average discrete gradient)**.** Note that using $f = \nabla H$ and projecting on $L_0(\tau) = 1$, the first coefficient of $\nabla H \circ x$ corresponds to the definition of the average discrete gradient from the Average Vector Field (AVF) method [QM08, CGM+12, COS14]. According to theorem 5.7, its closed-form expression is

$$\{\nabla H \circ x\}_0 = \int_0^1 \nabla H(x(\tau))\, \mathrm{d}\tau = \begin{cases} \dfrac{H(x_0 + \delta x) - H(x_0)}{\delta x} & \delta x \neq 0, \\ \nabla H(x_0) & \delta x = 0. \end{cases} =: \overline{\nabla} H(x_0, \delta x) \tag{5.41}$$

We note that the Average Discrete gradient has a regularisation effect shown on figure 5.6. In numerical applications, it can reduce the Lipschitz constant. For example, when applied to discontinuous functions, the averaged function is continuous everywhere except for $\delta x = 0$. We proved in [MH18] that the derivative of the Average Discrete Gradient $\overline{\nabla} H$ with respect to the unknown variable δx has the closed form expression

$$\frac{\partial}{\partial \delta x} \overline{\nabla} H(x_0, \delta x) = \begin{cases} \dfrac{\nabla H(x_0 + \delta x) - \overline{\nabla} H(x_0, \delta x)}{\delta x} & \delta x \neq 0, \\ \dfrac{1}{2} \dfrac{\partial^2 H}{\partial x^2}(x_0) & \delta x = 0. \end{cases} \tag{5.42}$$

This quantity plays the role of a "discrete Hessian" of H in the implementation of Newton iteration. We proposed an extension of this result to semi-continuous functions in [MH20].

Example 5.3 (AVF error estimation)**.** Still using $f = \nabla H$ and projecting on the next legendre polynomial $L_1(s) = 2s - 1$, after factorisation, we obtain in closed-form

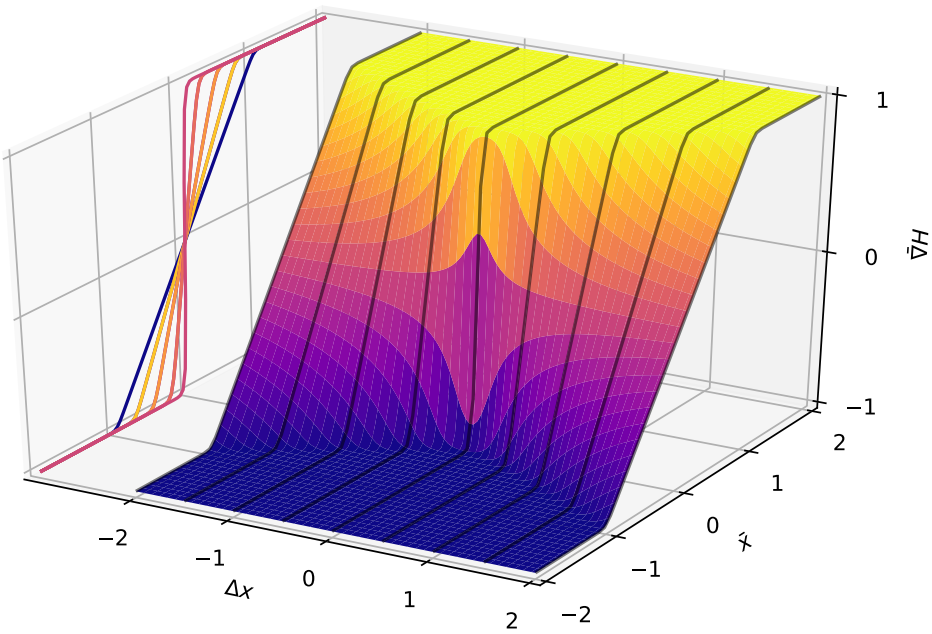
$$\{\nabla H \circ x\}_1 = \begin{cases} \dfrac{2}{x_1 - x_0}\left(\dfrac{H(x_1) + H(x_0)}{2} - \dfrac{H^{[1]}(x_1) - H^{[1]}(x_0)}{x_1 - x_0}\right) & x_1 \neq x_0, \\ 0 & x_1 = x_0. \end{cases} \tag{5.43}$$

The first term is the trapezoidal average and the second one is the continuous average of $H \circ x$ (i.e. the Average Discrete Gradient of the antiderivative $H^{[1]}(x)$). In other words, projection on L_1 is proportional to *the difference between the trapezoidal and the continuous average* of $H \circ x$. This result can be used to obtain the first coefficient of the Average Vector Field approximation error (projection order $p = 0$).

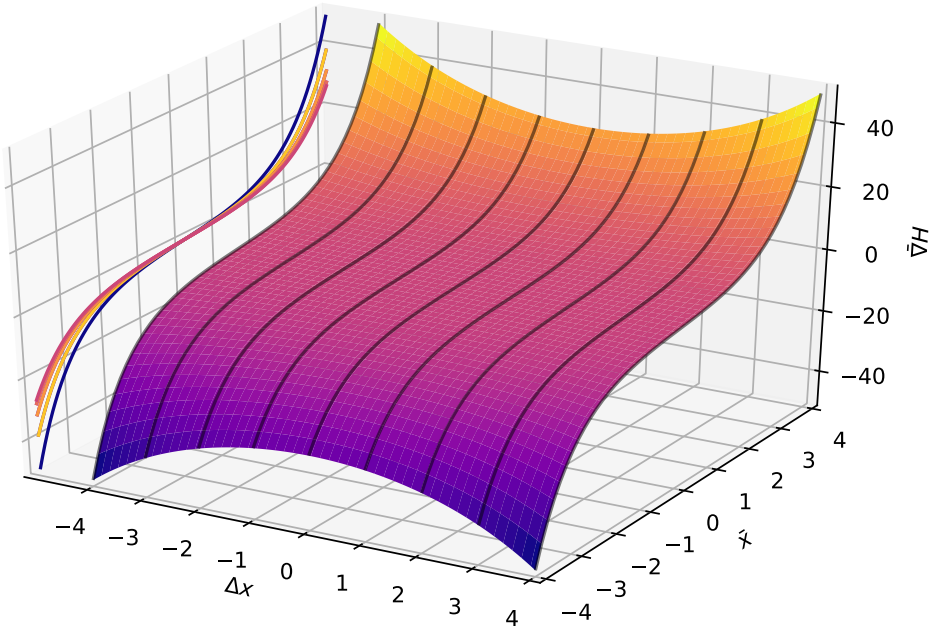
Anti-derivative anti-aliasing and spectral projection In the digital audio literature [PZLB16, BEPV17, BEV17, MH17, Hol20, Alb20, Car20], there is a growing interest for anti-derivative based anti-aliasing methods. They greatly improve the audible quality of audio simulations for a small additional cost. We note that spectral projection on polynomials can be interpreted as anti-aliasing since it truncates higher order spectral terms that cannot be represented in the approximation basis. Finally, as shown by Theorem 5.7, partial integration on the projection coefficients automatically involves anti-derivatives of the function of interest. Interesting connections between the Average Vector Field method and anti-derivative anti-aliasing have been discussed by the author in [MH17, MH18, MH19, MH20] and a partial form of Theorem 5.7 is published in [MH20].

Application to memoryless nonlinearities Note that the results from examples 5.2 and 5.3 are directly applicable to the projection of memoryless nonlinearities by using $\boldsymbol{f}(\boldsymbol{w}) = \boldsymbol{z}(\boldsymbol{w})$ (assuming dissipative potentials $\boldsymbol{Z}(\boldsymbol{w})$ are known, see 1.40 p.29). For pH-DAE this means that projection of memoryless non-linearites [8] are still computable in closed-form for projection order $p = 1$ (i.e. $\boldsymbol{w} \in \mathbb{P}^1$). This property has been exploited in [MH18, MH19, MH20].

8. We also note that a common situation in electronics is to have linear storage components (i.e. projections are exactly computable in closed form for any order $p \geq 0$) and nonlinear memoryless nonlinearities.



(a) $\nabla H(x) = \tanh(Kx)$, $K = 20$



(b) $\nabla H(x) = \sinh(Kx)$, $K = 1$

Figure 5.6 – Smoothing effect of the Average Discrete Gradient for $\nabla H(x) = \tanh(Kx)$, (i.e. $H(x) = \frac{1}{K}\ln\cosh(Kx)$ (top plot). When $K \to \infty$, it converges to the discontinuous sign function (discontinuous at the origin). The greater δx, the higher the regularisation effect. For symmetry reasons, the graph is drawn for the centered coordinates $\bar{x} = \frac{x_0+x_1}{2} = x_0 + \delta x/2$, and δx. Note that for hardening laws (bottom plot) $\nabla H(x) = \sinh(Kx)$, the ADG has the opposite effect, it increases the Lipschitz constant. To avoid this issue, we have shown in [MH20], using implicit parametrisations, that we can avoid the stiffening behaviour and improve convergence.

5.4.2 General purpose numerical quadratures

When higher order accuracy is seeked, for general functions, no exact integration formula can be used. Numerical quadratures are required to estimate projection coefficients

$$\widehat{\mathbf{f}}_n = \int_0^1 \phi_n(\tau)\boldsymbol{f}(\boldsymbol{X}(\tau))\,\mathrm{d}\tau \approx \sum_{i=1}^{L} w_i\phi_n(\tau_i)\boldsymbol{f}(\boldsymbol{X}(\tau_i)).$$

where abscissae $\tau_1,\dots,\tau_L$ and weights $w_1,\dots,w_L$ are chosen such that the integral is exact when the integrand belongs to a given functional subspace (typically polynomial or trigonometric functions). The mathematical literature on numerical quadrature formulas is huge. We forward the reader to the survey in reference [Gau81]. In this thesis we focus on Gauss-Legendre quadrature rules (see also [CMM+09, Hai10, BFCI14, CH17]).

Theorem 5.8 (Gauss-Legendre quadrature [SB13])**.** *Let $\{\tau_k\}_{k=1}^n$ be the roots of the n-th shifted orthonormal Legendre polynomial $P_n(\tau)$ and let $\{w_k\}_{k=1}^n$ be the solution of the (nonsingular) system of equations*

$$\sum_{i=1}^{n} P_j(\tau_i)w_i = \begin{cases} 1, & \text{if } j=0, \\ 0, & \text{if } j=1,\dots,n-1. \end{cases} \tag{5.44}$$

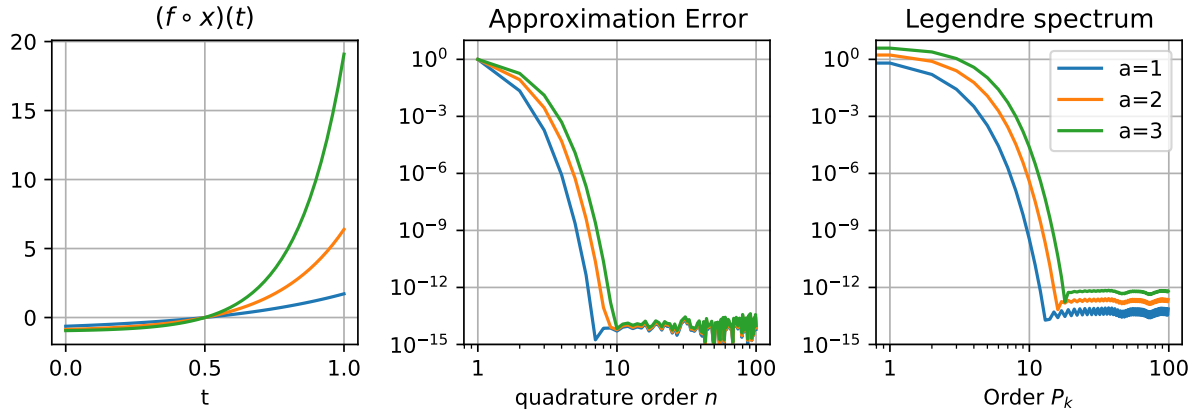
Then $w_i > 0$, for $i=1,\dots,n$ and $\int_0^1 p(\tau)\,\mathrm{d}\tau = \sum_{k=1}^n w_k f(\tau_k)$ holds $\forall p \in \mathbb{P}^{2n-1}([0,1])$.

Many proofs of this theorem exists. In this PhD, a proof highlighting the role of the reproducing kernel with explicit formulas for the weights w_k is detailed in appendix D.8 p.300.

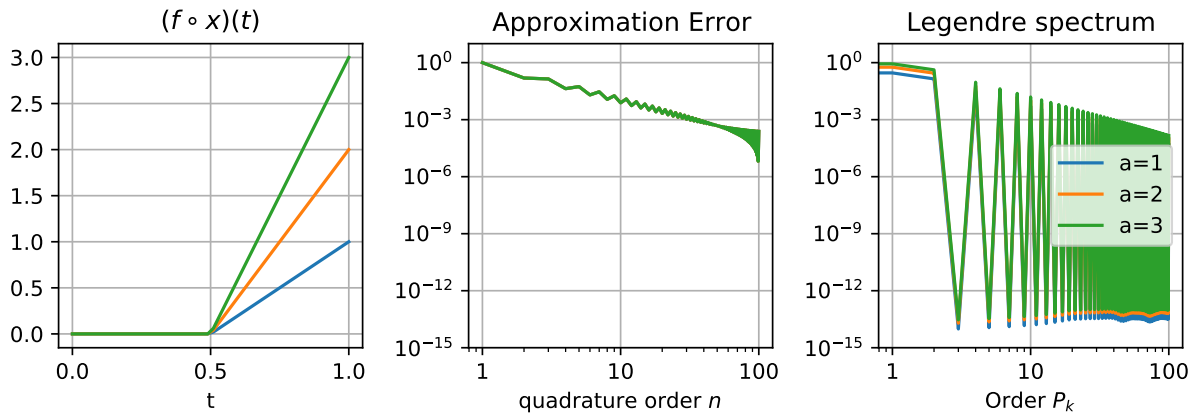
Exact projection results for polynomial nonlinearities From a practical point of view, if $\boldsymbol{f}$ is polynomial [9] with degree $d_{\boldsymbol{f}}$ and $\boldsymbol{X}$ is also polynomial [10] with degree $d_{\boldsymbol{X}}$, then $\boldsymbol{f}\circ\boldsymbol{X}$ is polynomial with degree $d = d_{\boldsymbol{f}}\cdot d_{\boldsymbol{X}}$. In other words, the "polynomial spectrum" of $\boldsymbol{f}\circ\boldsymbol{X}$ is band-limited (in the Legendre basis). By consequence, if a quadrature rule is exact for polynomials of degree d, its use in methods RPM to compute projections, makes energy and passivity preservation guaranteed (see [CH17]).

Approximation up to machine accuracy for nonlinearities with infinite spectrum In many interesting cases, $\boldsymbol{f}\circ\boldsymbol{X}$ has an infinite spectrum in the chosen basis [11] A naive implementation would require an infinite number of evaluation points. Fortunately, the situation is not desperate: if $\boldsymbol{f}$ and $\boldsymbol{X}$ are sufficiently smooth, the spectrum of $\boldsymbol{f}\circ\boldsymbol{X}$ has a fast decay rate ([WX12]) so that, exact integration (up to machine accuracy) can be reached with a finite number of evaluation points. This approach has been studied in [BFCI14] where machine accuracy is reached with few evaluation points. If however $\boldsymbol{f}$ is not smooth, then low projection orders and smaller time-steps should be used together with Theorem 5.7. Indeed, in this case, the fast convergence property of spectral methods is lost and the additional quality of higher orders methods is no longer worth the increase in numerical computation cost (cf [Boy01]). The Legendre spectrum and the convergence of Gauss-Legendre quadrature are illustrated on figure 5.7 for the cases of $\mathcal{C}^\infty$ and $\mathcal{C}^0$ functions.

9. Example: the Duffing and Van der Pol oscillators are cubic, the Lotka-Volterra equations are quadratic.
10. The spectrum of $\boldsymbol{f}\circ\boldsymbol{X}$ is also finite when $\boldsymbol{X}$ is trigonometric and $\boldsymbol{f}$ is polynomial.
11. For example $\boldsymbol{f} \in \{\sin,\cos,\sinh,\cosh,\exp,\min,\max,\dots\}$.



(a) $\mathcal{C}^\infty$ case: $f(x) = \exp(ax) - 1$



(b) $\mathcal{C}^0$ case: $f(x) = \max(ax, 0)$

Figure 5.7 – (Convergence of Gauss–Legendre quadrature). The graph (left), quadrature approximation error (middle) and Legendre spectrum $\widehat{\{f \circ x\}}_k$ (right) are plotted for the composition of functions $(f \circ x)(t)$ where $x(t) = x_0 + t(x_1 - x_0)$, $x_0 = -1, x_1 = 1$ is an affine trajectory and for two nonlinearities: (top) A $\mathcal{C}^\infty$ function $f(x) = \exp(ax) - 1$ (like a diode law) and (bottom) a piecewise linear $\mathcal{C}^0$ ReLU function $f(x) = \max(ax, 0)$ (used in opamp clipping) both for parameters $a = 1, 2, 3$. We can clearly see that for $\mathcal{C}^\infty$ functions (top), both the approximation error and the (Legendre) spectrum decay very fast. The error reaches the machine epsilon after a finite number of quadrature nodes. By contrast, for C^0 functions (bottom), both the approximation error and the Legendre spectrum decay much more slowly: the quadrature order and the number of Legendre coefficients have been increased to 100 but the quadrature error remains significant (about 10^{-4}) which is more than 10 orders of magnitude above the machine epsilon. The spectrum is shown in log-log scale to emphasize its slow linear decay (due to the discontinuity of the first derivative).

5.4.3 Representations, fixed-point and Newton iterations

Choice of representation Until now, to design RPM, we have worked with abstract functional spaces and projections, but the choice of functional space and its representation has remained open. To actually implement the method on a computer, we need finite-dimensional representations of functions for each time step (finite rate of innovation).

Questions To this end, several questions must be addressed, in particular:

- Should we use trajectories $\boldsymbol{X}(\tau)$ or their derivative $\delta\boldsymbol{X}(\tau)$ as primary representation? (i.e. should we use the state space or the space of flows and efforts as primary space?)
- Should we use use nodal representations (as in FEM and Runge–Kutta methods) or spectral representations (as in modal and spectral elements methods)? See [Boy01].
- Is it easier to work with orthogonal (as in spectral methods) or non-orthogonal (as in FEM) representations of functions? For which computational cost and numerical conditioning ?

Choices In this chapter, we make the following choices:

A. We use the projected flows $\boldsymbol{f}$ and efforts $\boldsymbol{e}$ (in L^2) as the approximated objects, rather than the state $\boldsymbol{X}$. Indeed, since we are not only interested in solving autonomous ODEs, but on manipulating PHS (with ports), it is more natural and consistent to have a common representation for all components [12]. The state $\boldsymbol{x}(t) = \boldsymbol{x}(t_0) + \int_{t_0}^{t} \boldsymbol{f}(s)\,\mathrm{d}s$ is treated as an internal construct of energy-storing components (treated as a hidden variable).

B. We use (spectral) orthogonal basis coefficients. Indeed, this is a natural fit for a projection method, furthermore they have optimal conditioning, require less computations [13] and their coefficients decay quickly for smooth functions (see fig. 5.7a).

Choice (A) is different from the standard formulation of CSRK methods (i.e. we emphasize the role of the reproducing kernel $k_{\mathcal{P}}(\tau,\sigma)$ rather than the integrated kernel $A(\tau,\sigma)$). More precisely (for an autonomous PHS) we solve the equation

$$\delta\boldsymbol{X}(\tau) = \mathcal{P}(\mathbf{J}-\mathbf{R})\nabla H\left(\mathbf{x}_0 + h\int_0^{\tau}\delta\boldsymbol{X}(s)\,\mathrm{d}s\right), \qquad \text{where} \qquad \delta\boldsymbol{X}(\tau) = \sum_{k=0}^{p-1}\phi_k(\tau)\widehat{\delta\mathbf{X}}_k,$$

for the coefficients $\widehat{\delta\mathbf{X}}_k$ (the *true unknowns* in the projection space) rather than

$$\boldsymbol{X}(\tau) = \mathbf{x}_0 + h\int_0^{\tau}\mathcal{P}\left[(\mathbf{J}-\mathbf{R})\nabla H\left(\boldsymbol{X}\right)\right](s)\,\mathrm{d}s, \qquad \boldsymbol{X}(\tau) = \mathbf{x}_0 + \sum_{k=0}^{p-1}\left(\int_0^{\tau}\phi_k(s)\,\mathrm{d}s\right)\widehat{\mathbf{X}}_{k+1}.$$

with respect to coefficients $\widehat{\mathbf{X}}_k$ (where the initial condition $\mathbf{x}_0$ is given by the problem). Note that our choice is closely related to the W-transformation of Runge-Kutta methods [BG08, p.267].

Fixed-point and Newton iteration For pH-ODE, we have seen (theorem 5.2 p.127) that the fixed-point iteration is contracting for $hL < \frac{\pi}{2}$ where L is the Lipschitz constant of the vector field. However the existence domain of solutions can be larger than predicted by Lipschitz conditions [14] and the fixed-point convergence is often too slow. For these reasons it is often advantageous to use (simplified) Newton iteration and we know that if the fixed-point converges, then Newton converges too. Newton iteration for pH-DAE is also discussed in subsection 5.3.2 p.138.

12. energy storing: $\boldsymbol{f} = \dot{\boldsymbol{x}}$, $\boldsymbol{e} = \nabla H(\boldsymbol{x})$, memoryless: $\boldsymbol{f} = \boldsymbol{w}$, $\boldsymbol{e} = \boldsymbol{z}(\boldsymbol{w})$, ports: $\boldsymbol{f} = \boldsymbol{y}$, $\boldsymbol{e} = \boldsymbol{u}$.

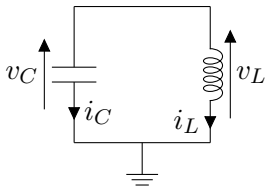
13. For example, the operational matrix of the Volterra integration operator is tri-diagonal and almost skew-symmetric in the Legendre Basis (see appendix C.4 p.286).

14. For example $\dot{x} = \lambda x$ has solution $\exp(\lambda t)x_0$ independently of the stiffness of its Lipschitz constant $L = |\lambda|$ and Newton iteration converges in one iteration for linear problems.

5.5 Examples

5.5.1 Nonlinear Conservative LC

We consider a nonlinear LC oscillator described by the schematics (left), Dirac structure (middle) and its Hamiltonian formulation (right)



$$\equiv \begin{bmatrix} i_C \\ v_L \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} v_C \\ i_L \end{bmatrix} \iff \begin{bmatrix} \dot{q} \\ \dot{\phi} \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \nabla H_C(q) \\ \nabla H_L(\phi) \end{bmatrix}.$$

The flows are $i_C = \dot{q}$, $v_L = \dot{\phi}$, the effort laws and associated Hamiltonian are given by

$$v_C(q) = \nabla H_C(q) = \frac{q}{C}, \qquad H_C(q) = \frac{q^2}{2C},$$
$$i_L(\phi) = \nabla H_L(\phi) = I_S \tanh\left(\frac{\phi}{LI_S}\right), \qquad H_L(\phi) = LI_S^2 \ln\cosh\left(\frac{\phi}{LI_S}\right),$$

where I_S denote the saturation current of the inductor [15]. For simplicity, we take $L = C = \omega^{-1}$ and $I_S = 1$ such that for small values [16] of ϕ the oscillator has pulsation $\omega = 1/\sqrt{LC}$ rad s^{-1}.

- Step i) We use the orthonormal Legendre basis $\left[P_i(\tau)\right]_{i=0}^{p-1}$ and use as unknowns the vector of Legendre coefficients

$$\delta\vec{\mathbf{q}} := \left[\langle P_i \,|\, i_C\rangle\right]_{i=0}^{p-1}, \qquad \delta\vec{\Phi} := \left[\langle P_i \,|\, v_L\rangle\right]_{i=0}^{p-1},$$

- Step ii) for any scalar function $H(x)$, we define its (Legendre) projected gradient by

$$\vec{\nabla} H(x_0; \delta\vec{\mathbf{x}}) := \left[\left\langle P_i \,\middle|\, \nabla H\left(x_0 + h\int_0^\tau \sum_{j=0}^{p-1} P_j(\sigma)\delta\vec{\mathbf{x}}_j \,\mathrm{d}\sigma\right)\right\rangle\right]_{i=0}^{p-1}. \tag{5.45}$$

  Substituting functions of time by their projection coefficients (computed according to the results of sections 5.4.1 5.4.2), we obtain an algebraic system of dimension $2p$ (projected Hamitonian system) which is solved using Newton iteration.

$$\begin{bmatrix} \delta\vec{\mathbf{q}} \\ \delta\vec{\Phi} \end{bmatrix} = \begin{bmatrix} \mathbf{0} & -\mathbf{I}_p \\ \mathbf{I}_p & \mathbf{0} \end{bmatrix} \begin{bmatrix} \vec{\nabla} H_C(q_0; \delta\vec{\mathbf{q}}) \\ \vec{\nabla} H_L(\phi_0; \delta\vec{\Phi}) \end{bmatrix}. \tag{5.46}$$

- Step iii) For $\alpha \in \{0, 1\}$ we evaluate the boundary conditions (according to B.3 p.278)

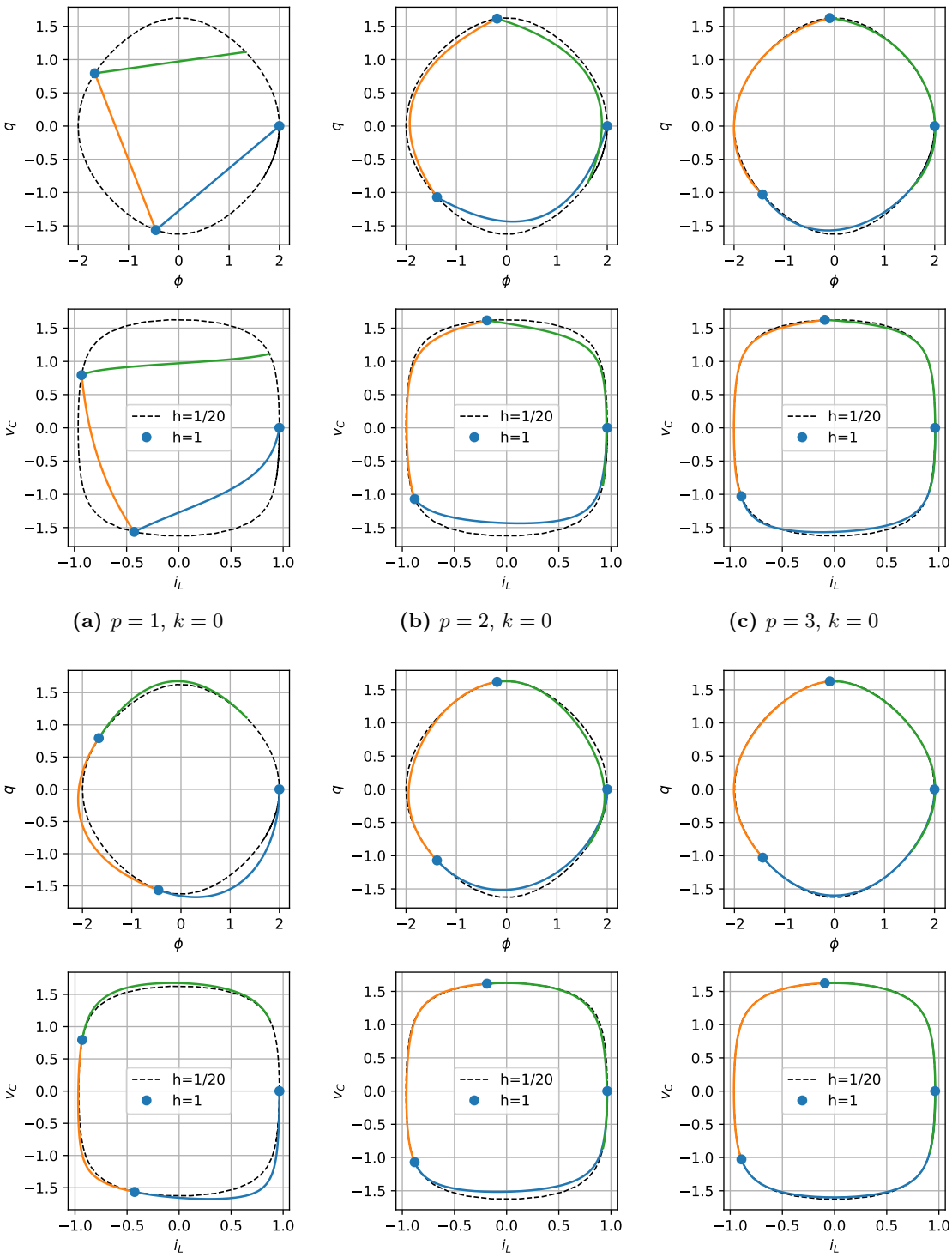
$$\mathcal{B}_\alpha^0(q) = -\nabla H_L(\phi_\alpha), \qquad \mathcal{B}_\alpha^1(q) = -\nabla^2 H_L(\phi_\alpha)\nabla H_C(q_\alpha), \qquad \text{etc}$$
$$\mathcal{B}_\alpha^0(\phi) = \nabla H_C(q_\alpha), \qquad \mathcal{B}_\alpha^1(\phi) = -\nabla^2 H_C(q_\alpha)\nabla H_L(\phi_\alpha), \qquad \text{etc}$$

  This regularisation process yields piecewise $\mathcal{C}^k$ solutions $q(t), \phi(t)$ thanks to the boundary functions $\left\{\psi_\alpha^m(\tau)\right\}$ defined in proposition 5.4 p.129.

Simulation results for different values of order p and regularity k are shown in figures 5.8-5.13.

15. In this example, we neglect hysteresis and use a generic tanh nonlinearity rather than a realistic one. We have supervised a work on a detailed inductor model for PHS (based on statistical physics) which includes hysteresis. This work, which is out of the scope of this thesis, has been published in [NMHR20].

16. For small values of ϕ, we have $\nabla H_L(\phi) = \phi/L + \mathcal{O}(\phi^3)$ so that the circuit reduces to a harmonic oscillator.

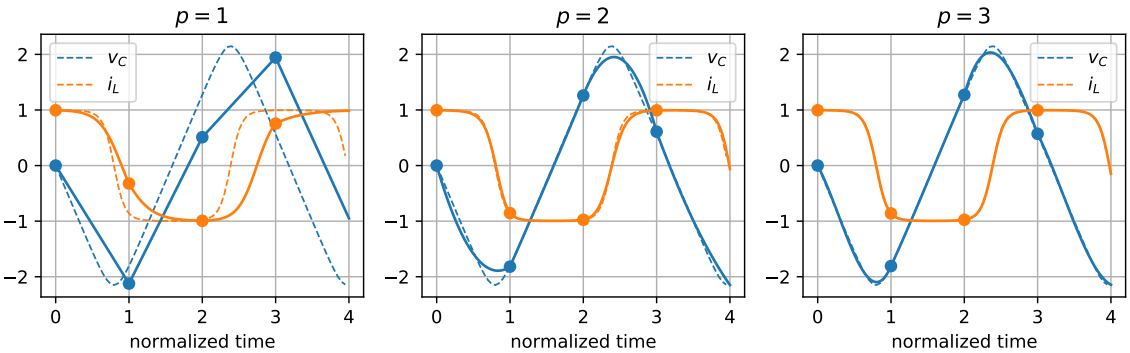


(a) $p = 1$, $k = 0$ **(b)** $p = 2$, $k = 0$ **(c)** $p = 3$, $k = 0$

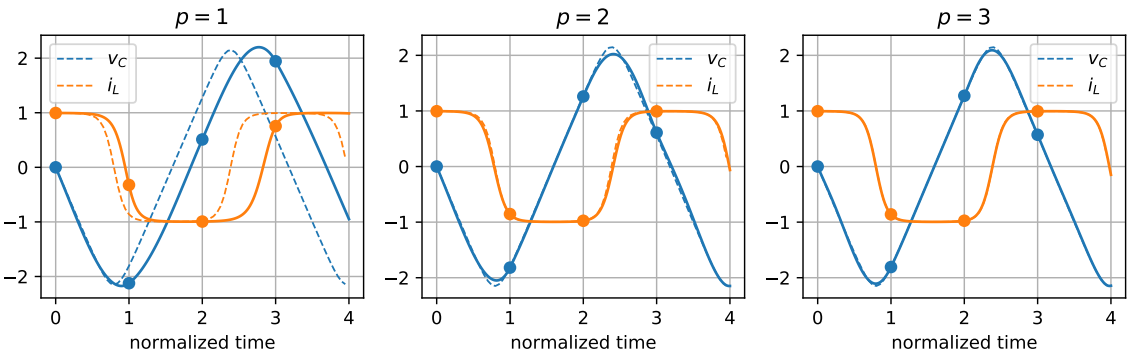
(d) $p = 1$, $k = 1$ **(e)** $p = 2$, $k = 1$ **(f)** $p = 3$, $k = 1$

Figure 5.8 – (Nonlinear LC) Orbits for projection order $p = 1, 2, 3$, and regularity order $k = 0, 1$, for a Nyquist pulsation $\omega = \pi$ (the actual pulsation is slower because of nonlinearities) and initial conditions $(q_0, \phi_0) = (0, 2)$. Plots are shown both in the phase space (ϕ, q) (first row), and in the flow/effort space (i_L, v_C) (second row).

Orbits and trajectories Orbits in the (ϕ, q) and (i_L, v_C) planes are shown on figure 5.8 and time trajectories are shown on figure 5.9. A pulsation close to the Nyquist frequency has been chosen in order to be able to show visual differences between different values of projection and regularity order p, k. On figure 5.8d ($p = 1$, $k = 1$), since the accuracy is only $\mathcal{O}(h^2)$ and we are close to the Nyquist frequency, we remark that the magnitude of derivatives is overestimated (overshoot). Despite this, orbits are much closer to the true manifold for regularity $k = 1$ (fig. 5.8d) than for $k = 0$ (fig. 5.8a). As the projection order p increases, orbits converge quickly to the true manifold but the derivatives remains discontinuous at the junctions. Increasing the regularity k improves the situation (the accuracy is now high enough to avoid derivative overestimation).



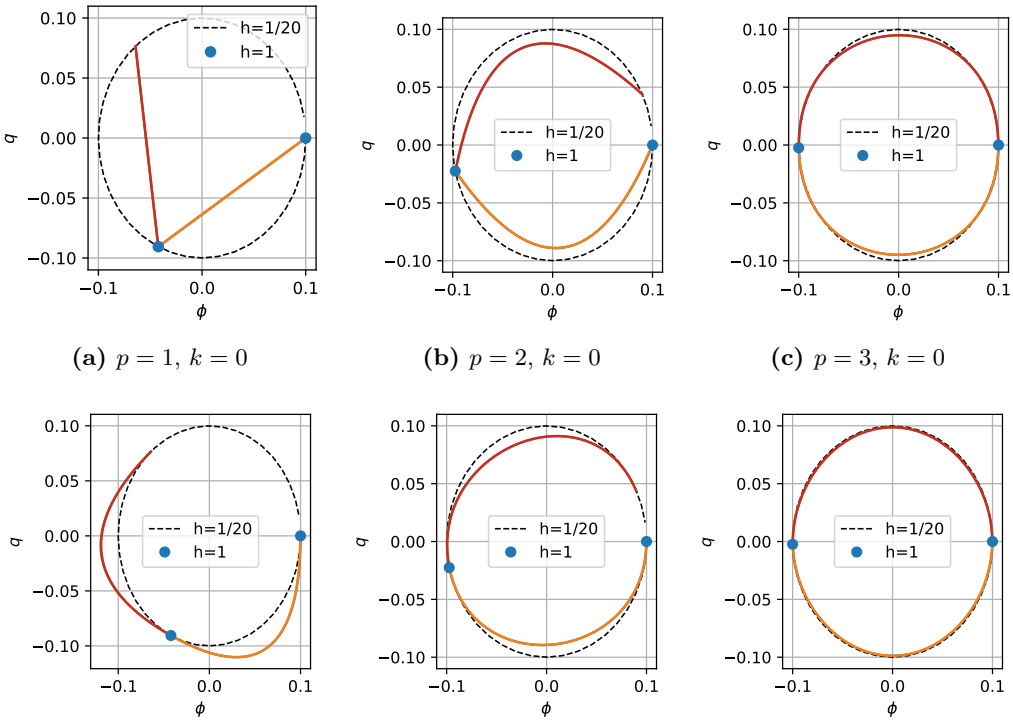
(a) regularity order $k = 0$



(b) regularity order $k = 1$

Figure 5.9 – (Nonlinear LC) Trajectories for projection order $p = 1, 2, 3$, regularity order $k = 0, 1$, pulsation $\omega = \pi$ and initial conditions $(q_0, \phi_0) = (0, 2)$. Oversampled trajectories by a factor of 20 are shown with dashed lines. Dots correspond to the boundaries of time frames.

Frequency warping and dispersion To emphasize the effect of projection order on frequency warping, it is shown on figure 5.10 that the frequency warping (dispersion) error diminishes greatly as p increases. In just two steps, the full circle is accurately reproduced. For $(p = 1, k = 1)$ (fig. 5.10d), we see that the accuracy is not high enough to simulate a pole at the Nyquist frequency: the magnitude of the vector field is over-estimated. Nevertheless, even in this extreme situation, the smooth solution ($k = 1$, fig. 5.10d) is still better than the affine approximation ($k = 0$, fig. 5.10a). For $p = 3$, the warping error becomes negligible. The effect of projection order on frequency and dissipation warping is further detailed in the appendix on figures D.2, D.3 p.298. A general formula to obtain the corresponding (A-stable) stability functions is proposed in section D.7 p.297.



(a) $p = 1$, $k = 0$ **(b)** $p = 2$, $k = 0$ **(c)** $p = 3$, $k = 0$

(d) $p = 1$, $k = 1$ **(e)** $p = 2$, $k = 1$ **(f)** $p = 3$, $k = 1$

Figure 5.10 – (Linear LC) Orbits for projection order $p = 1, 2, 3$, regularity order $k = 0, 1$, pulsation $\omega = \pi$ (Nyquist frequency) and initial conditions $(q_0, \phi_0) = (0, 0.1)$. Frequency warping can be observed by looking at blue dots (that should theoretically be aligned at angles 0 and π).

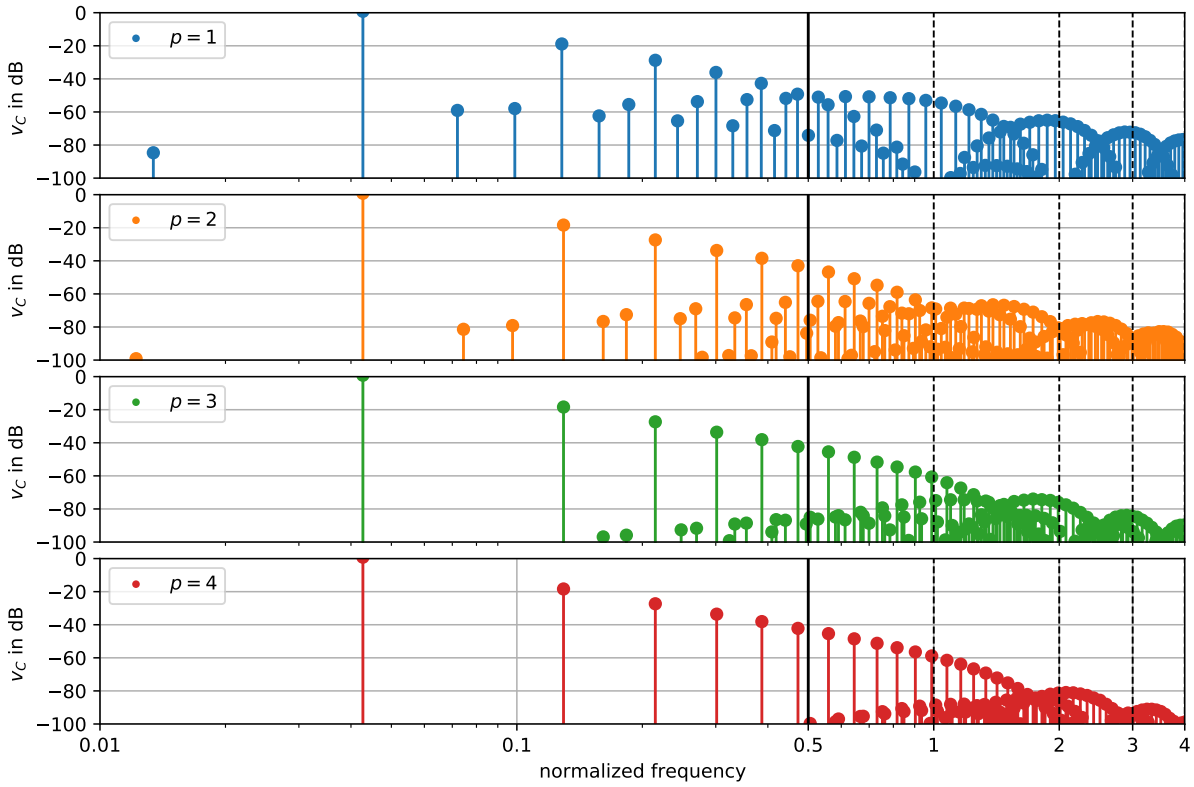
Aliasing To inspect aliasing, several oscillation cycles of $v_C(t)$ are simulated and examined in the Fourier domain (see figure 5.11). To exploit continuous-time trajectories, signals are (over)sampled over each time-step by a factor of 20, weighted by a Dolph–Chebyshev window (sidelobes rejection > 100 dB), and a Fast Fourrier Transform is performed. For reference, the Nyquist frequency of the time-stepping simulation scheme is shown in solid black and multiples of the sampling frequency are shown in dashed black lines.

We remark that, above the Nyquist frequency, the spectral content approaches the expected harmonic structure more and more closely, as the projection order p increases: this is due to the increasing bandwidth (w.r.t. p) in the sense of generalised sampling theory (see section 3.1 p.83). Accordingly, the aliasing decreases in the audio frequency range: the signal to (aliasing) noise ratio is above 100 dB for $p = 3$ in the frequency band below 20 kHz.

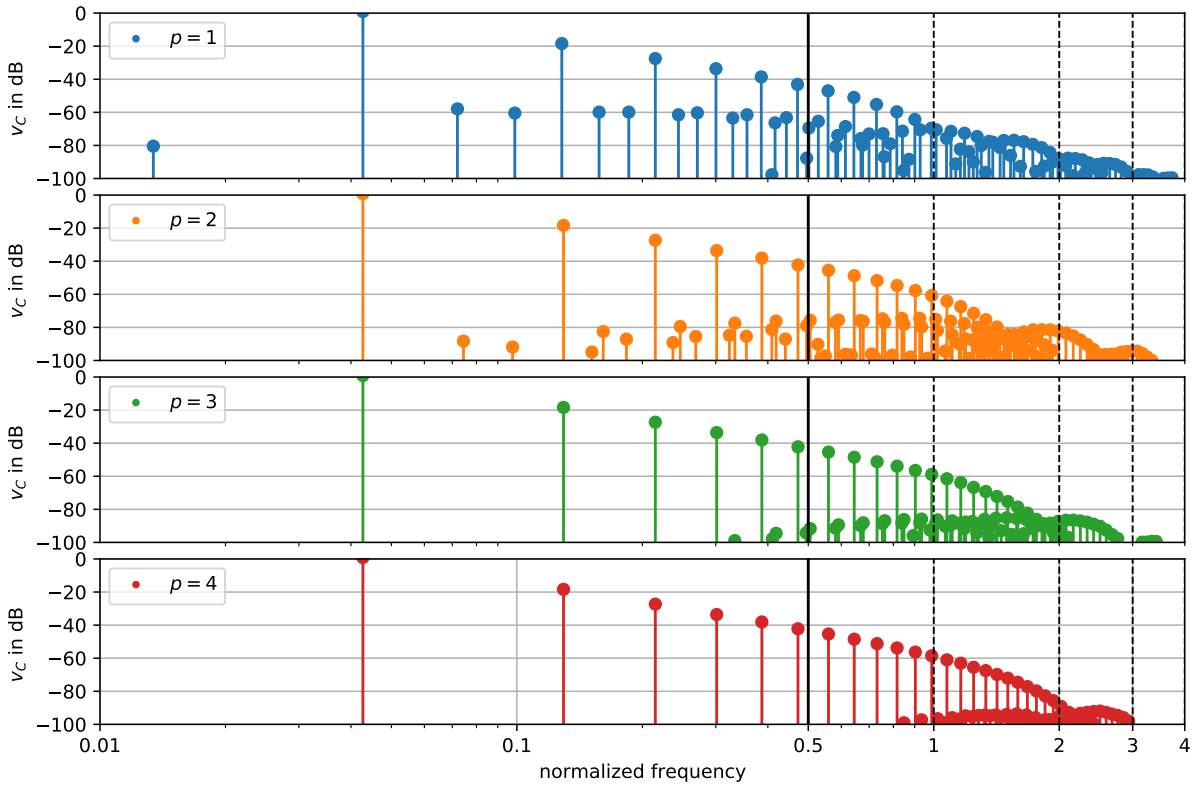
For $k = 0$, because of discontinuities, the high frequency spectrum has a slow spectral decay, but the magnitude of discontinuities diminishes when increasing accuracy. As expected, increasing the Sobolev regularity k exhibits a faster spectral decay. We remark that the signal to noise ratio and aliasing rejection are also improved in the frequency range around the Nyquist frequency, including in the frequency band *below the Nyquist frequency*.

However, as we have already warned before (see fig. 5.8), we experiment that, increasing the regularity k should be used with care (in regions where accuracy is high enough). Otherwise unwanted local frequency modulation can occur and create sub-harmonics in the pass-band.

A perspective that is left for further work would be to use backward error analysis theory [HLW06] to evaluate (multi-)derivatives of the modified vector field that are consistent with the frequency warping induced by projection operators.



(a) regularity $k = 0$, spectral decay increase with order, pass-band aliasing drops below -100dB for $p = 3$.



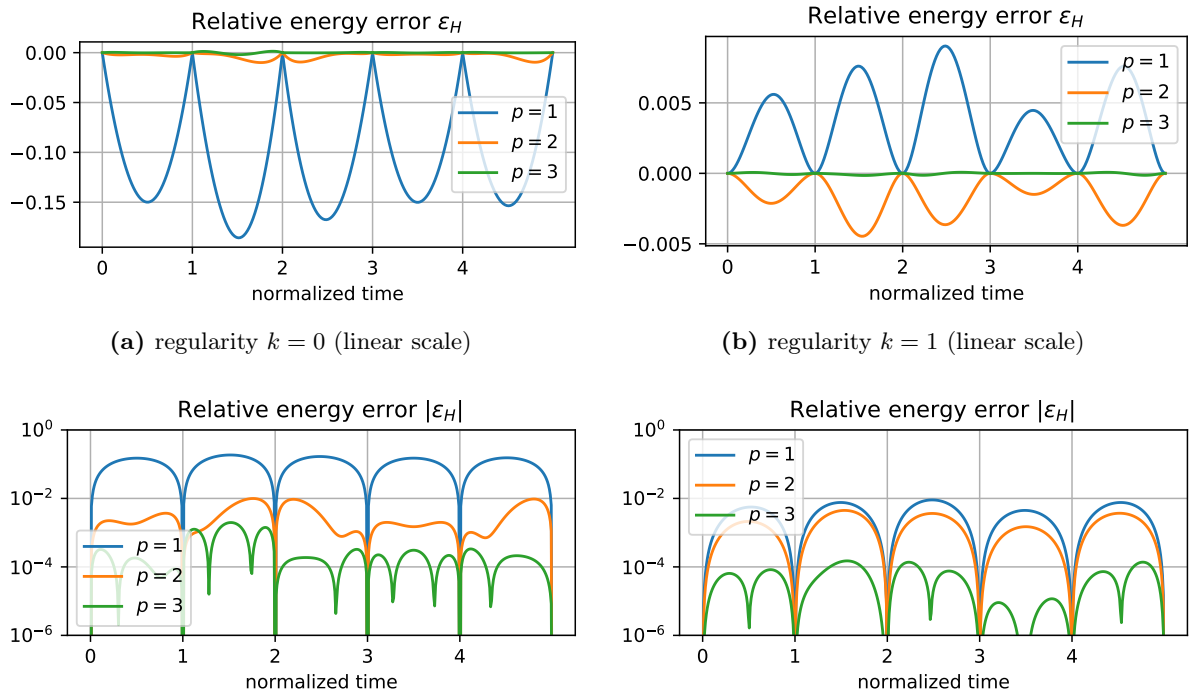
(b) regularity $k = 1$, faster spectral decay, better high frequency signal to noise ratio.

Figure 5.11 – (Nonlinear LC) Spectrum and aliasing of $v_C(t)$ according to projection order p and smoothness k. Note that state trajectories are $\mathcal{C}^k$ in the time domain. Spectral peaks are shown instead of the full spectrum to improve the visual contrast between signal harmonics and aliased partials.

Local and long-term energy error The local energy behaviour for $\omega = \pi/4$ is shown in figure 5.12. We remark as expected that the relative energy error

$$\epsilon_H(\tau) = \frac{H(q(\tau), \phi(\tau)) - H(q_0, \phi_0)}{H(q_0, \phi_0)},$$

vanishes on the time-stepping grid. Furthermore, its maximal also diminishes by an order of magnitude as the projection order p is increased. Finally, increasing the regularity order k also diminishes the local energy error (note the similarity with Peano kernels from figure 5.5 p.134). On figure 5.13, we show that energy conservation is satisfied on the time-stepping grid $\tau \in \mathbb{N}$ up to machine epsilon accuracy in double precision arithmetic.



(a) regularity $k = 0$ (linear scale)

(b) regularity $k = 1$ (linear scale)

(c) regularity $k = 0$ (log scale)

(d) regularity $k = 1$ (log scale)

Figure 5.12 – (Nonlinear LC) Continuous-time energy error $\epsilon_H(\tau)$ for $\tau \in \mathbb{R}$ according to projection order p and regularity order k for $\omega = \pi/2$ and $(q, \phi) = (0, 1)$.

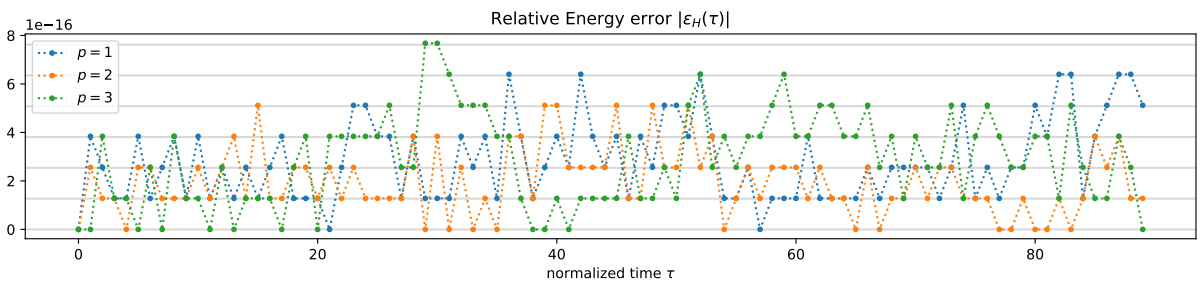
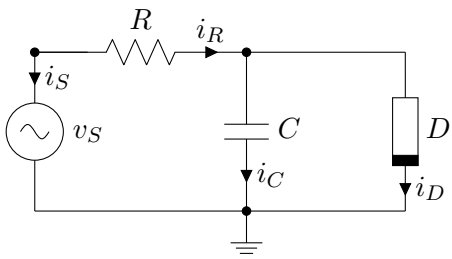


Figure 5.13 – (Nonlinear LC) Energy conservation on steps boundaries $\tau \in \mathbb{N}$ for $\omega = \pi/10$ and $(q, \phi) = (0, 1)$. Horizontal lines correspond to multiples of the machine epsilon.

5.5.2 Diode Clipper

We consider the diode clipper circuit and its semi-explicit PHS representation



$$\Longleftrightarrow \quad \begin{bmatrix} v_R \\ v_D \\ i_C \\ i_S \end{bmatrix} = \begin{bmatrix} & & -1 & 1 \\ & & 1 & 0 \\ 1 & -1 & & \\ -1 & 0 & & \end{bmatrix} \begin{bmatrix} i_R(v_R) \\ i_D(v_D) \\ v_C(Q^h_{q_0}(i_C)) \\ v_S \end{bmatrix}$$

where $i_R(v_R) = v_R/R$, $i_D(v_D) = 2I_S \sinh(v_D/V_T)$, $v_C(q) = q/C$, and v_S is a given input function. For the purpose of simulation, with $i_C = \dot{q}$, we can reduce it to the ODE

$$\dot{q} = -\frac{q}{RC} - i_D\left(\frac{q}{C}\right) + \frac{v_S}{R}, \qquad q = Q^h_{q_0}(i_C = \dot{q}) = q_0 + h\int_0^\tau i_C(s)\,\mathrm{d}s.$$

i) Let $\vec{\delta\mathbf{q}} := \left[\langle P_k \,|\, i_C\rangle\right]_{k=0}^{p-1}$, $\vec{\mathbf{v}}_S := \left[\langle P_k \,|\, v_S\rangle\right]_{k=0}^{p-1}$ be the Legendre coefficients of $i_C(\tau), v_S(\tau)$.

ii) Let $\mathbf{P} = \mathbf{I}_{(p-1)\times p}$ be the matrix representation of the projector and define the projected charge and diode current operators

$$\boldsymbol{Q}^h_{q_0}\left(\vec{\mathbf{i}}\right) := \left[\left\langle P_i \,\middle|\, q_0 + h\int_0^\tau \sum_{j=0}^{p-1} P_j(s)\vec{\mathbf{i}}_j\,\mathrm{d}s \right\rangle\right]_{i=0}^{p} = \begin{bmatrix} q_0 \\ \mathbf{0} \end{bmatrix} + h\mathbf{V}\vec{\mathbf{i}},$$

$$\boldsymbol{I}_D(\vec{\mathbf{v}}) := \left[\left\langle P_i \,\middle|\, i_D\left(\sum_{j=0}^{p} P_j(\tau)\vec{\mathbf{v}}_j\right)\right\rangle\right]_{i=0}^{p-1}$$

where $\mathbf{V}$ is the $p \times (p-1)$ operational matrix of the Volterra integration operator $\mathcal{V} = \int_0^\tau$. The projected ODE becomes the algebraic fixed point on $\vec{\delta\mathbf{q}}$

$$\vec{\delta\mathbf{q}} = -\mathbf{P}\frac{\vec{\mathbf{q}}}{RC} - \boldsymbol{I}_D\left(\frac{\vec{\mathbf{q}}}{C}\right) + \frac{\vec{\mathbf{v}}_S}{R} =: \boldsymbol{G}\left(\vec{\delta\mathbf{q}}\right), \qquad \text{where } \vec{\mathbf{q}} = \boldsymbol{Q}^h_{q_0}\left(\vec{\delta\mathbf{q}}\right).$$

We define the Newton function $\boldsymbol{F}\left(\vec{\delta\mathbf{q}}\right) := \vec{\delta\mathbf{q}} - \boldsymbol{G}\left(\vec{\delta\mathbf{q}}\right)$ and use the simplified Newton iteration to solve $\boldsymbol{F}\left(\vec{\delta\mathbf{q}}_\star\right) = \mathbf{0}$ given by

$$\vec{\delta\mathbf{q}}_{k+1} := \vec{\delta\mathbf{q}}_k + \Delta\vec{\delta\mathbf{q}}_k, \qquad \Delta\vec{\delta\mathbf{q}}_k := -(\boldsymbol{F}_0')^{-1}\left(\vec{\delta\mathbf{q}}_k - \boldsymbol{G}\left(\vec{\delta\mathbf{q}}_k\right)\right), \qquad \vec{\delta\mathbf{q}}_0 := \mathbf{0}.$$

Its Jacobian is tridiagonal positive definite (easy to invert) and equal to

$$\boldsymbol{F}_0' = \mathbf{I} + \alpha\mathbf{P}\mathbf{V} \succeq 0, \qquad \text{with} \qquad \alpha = \frac{h}{RC}\left(1 + R\frac{\partial i_D}{\partial v_D}\left(\frac{q_0}{C}\right)\right).$$

iii) For regulariy $k > 0$, we evaluate the boundary conditions at $\tau = \alpha \in \{0, 1\}$

$$\mathcal{B}^0_\alpha(\widetilde{i}_C) = \frac{1}{R}\left(v_S(\alpha) - \frac{q_\alpha}{C}\right) - i_D\left(\frac{q_\alpha}{C}\right),$$

$$\mathcal{B}^1_\alpha(\widetilde{i}_C) = \frac{1}{R}\left(\dot{v}_S(\alpha) - \frac{\mathcal{B}^0_\alpha(\widetilde{i}_C)}{C}\right) - i_D'\left(\frac{q_\alpha}{C}\right)\frac{\mathcal{B}^0_\alpha(\widetilde{i}_C)}{C}, \qquad \text{etc}$$

The regularized current $\widetilde{i}_C(\tau)$ is synthesized using the boundary functions $\left\{\psi^m_\alpha(\tau)\right\}$ defined in proposition 5.4 p.129. The voltage $\widetilde{v}_C(\tau)$ is then obtained from $\widetilde{i}_C(\tau)$ by integration.

Simulation results Simulation results for a ramp/step input and for different values of projection order p and regularity k are shown on figure 5.14. As expected we can observe diode clipping of the voltage about 0.7 V. Simulations differ mostly on how they behave when switching from the linear regime to the stiff clipping mode. For $p = 0$, we observe well-known Nyquist oscillations artefacts about the exact solution. These are due to the frequency warping of the method (stiff real poles are warped towards imaginary poles at the Nyquist frequency, see fig. D.2 p.298). Increasing the projection order p, we observe a significant reduction of this phenomenon thanks to higher order accuracy and bandwidth. Increasing the regularity order k yields smoother solutions, but for stiff poles (as we already noticed in fig. 5.10), we observe that additional smoothness also yields an amplification of artefacts. Increasing jointly p and k reduces both the amplitude of oscillations and their frequency. However small oscillations are still observable for $p = 2, k = 2$.

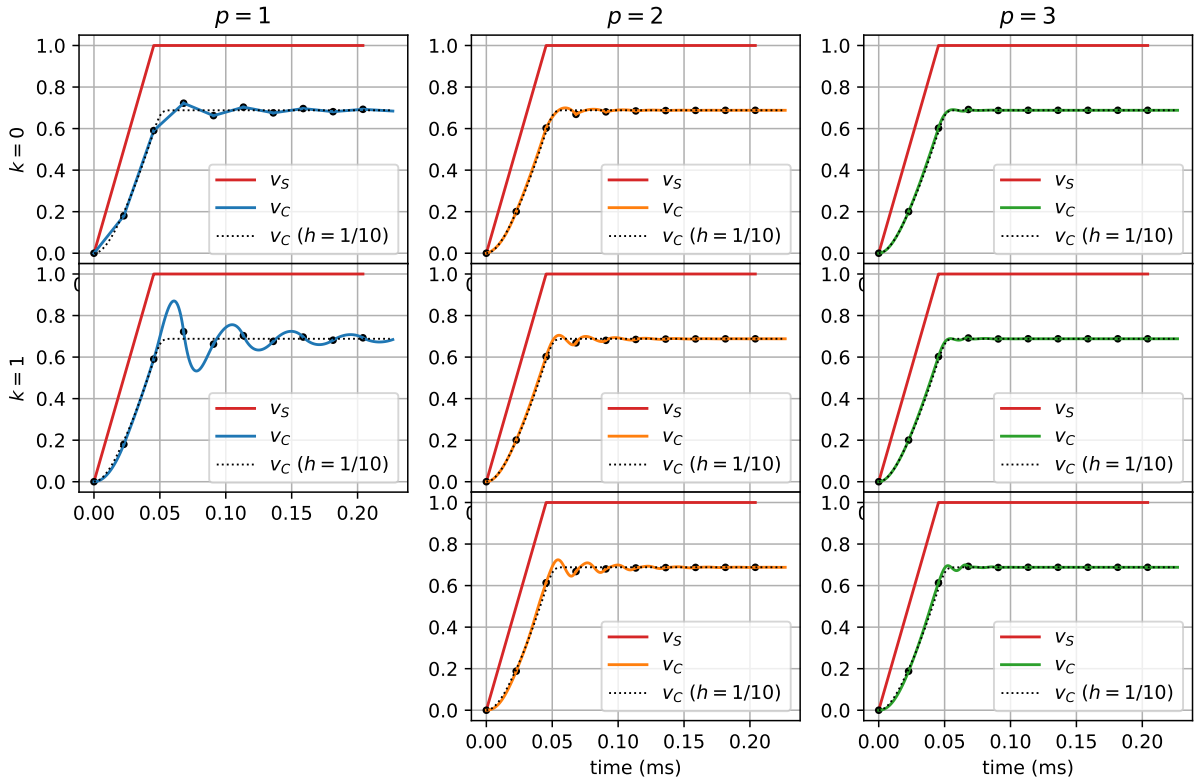
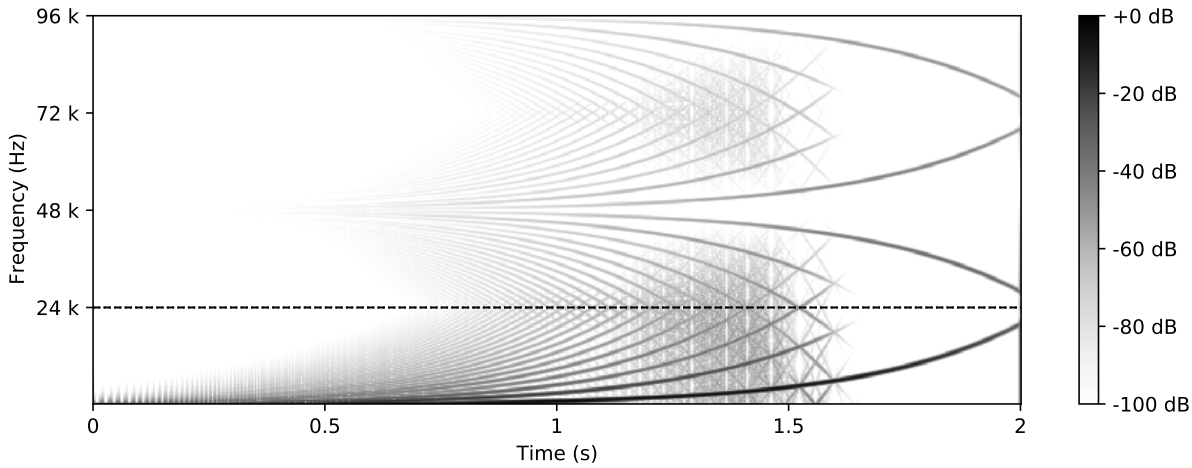
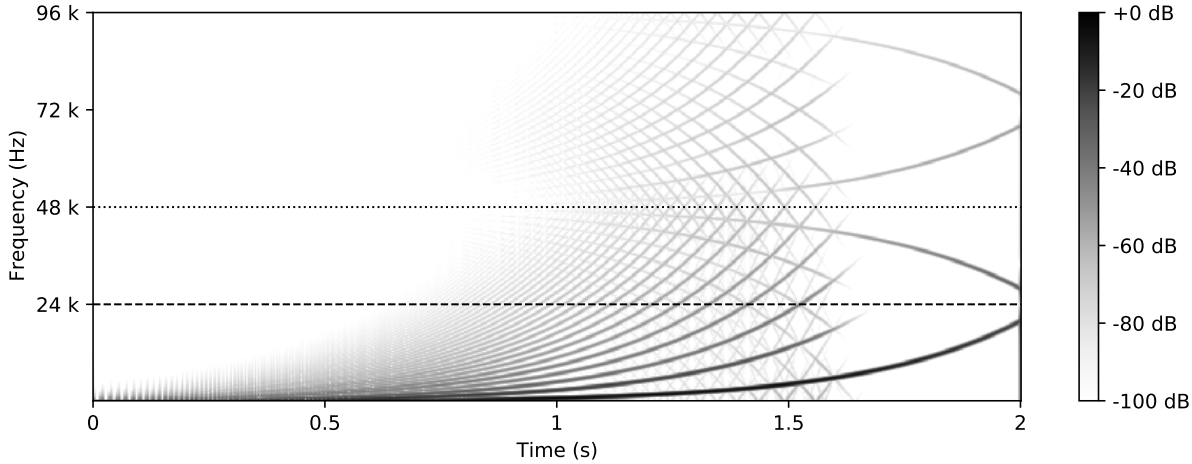


Figure 5.14 – (Diode clipper) Simulation for projection order $p = 1, 2, 3$, and regularity order $k = 0, 1, 2$ with $R = 1$ kΩ, $C = 20$ nF such that the diode clipper cutoff $f_c = 50$ kHz is set *above the sampling frequency* $f_s = 44.1$ kHz. We use $I_S = 1$ fA and $V_T = 26$ mV. The case $(k = 2, p = 1)$ is not shown because the accuracy order is not high enough to use second derivatives.

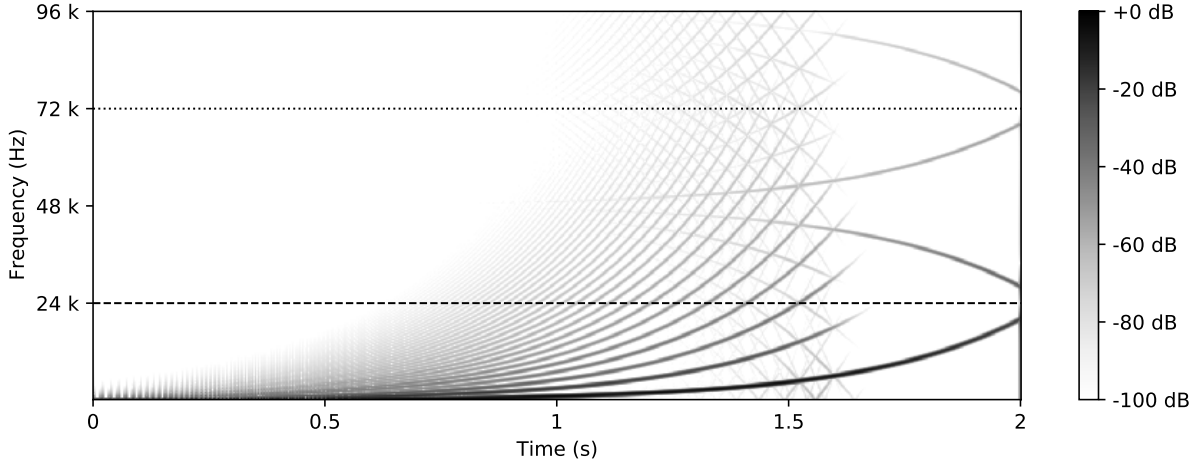
Sine sweep spectrograms and aliasing Spectrogram responses of the diode clipper to a sinusoidal sweep are also displayed on in 5.15 in linear scale and in figure 5.16 in log scale. The linear frequency scale is makes the visualisation easier to exhibit the generalized bandwidth and aliasing reduction of higher-order projection. The logarithmic frequency scale is closer to the human hearing resolution, the residual aliasing below 20kHz is easier to visualize with this scale. We see that with increasing order $p = 3$ the audible aliasing becomes barely noticeable, it only happens for input sinusoids above 5 kHz, and folded harmonics level stays below about 70/80dB. In comparison for low order $p = 1$, aliasing starts for sinusoids below 1kHz and its level is above −60dB.



(a) projection order $p = 1$

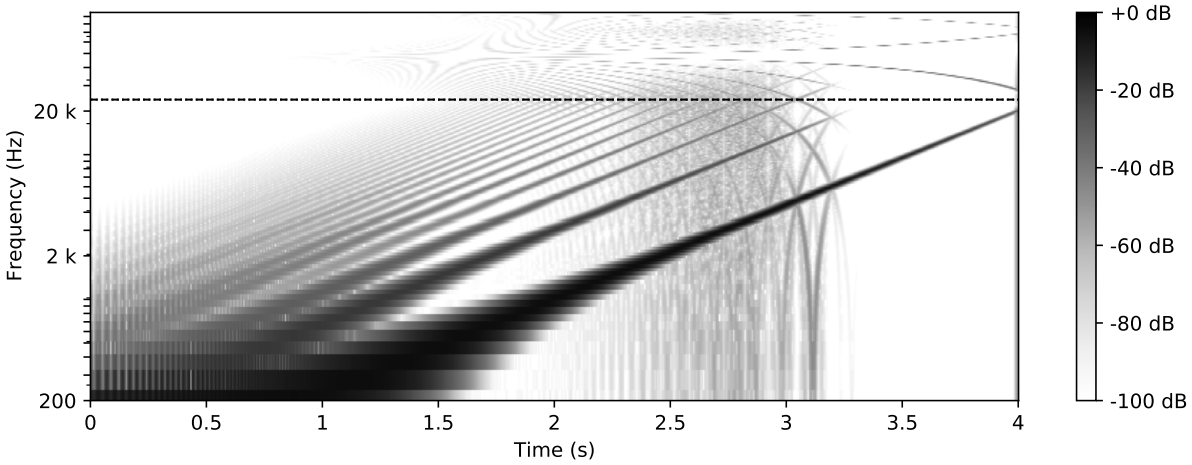


(b) projection order $p = 2$

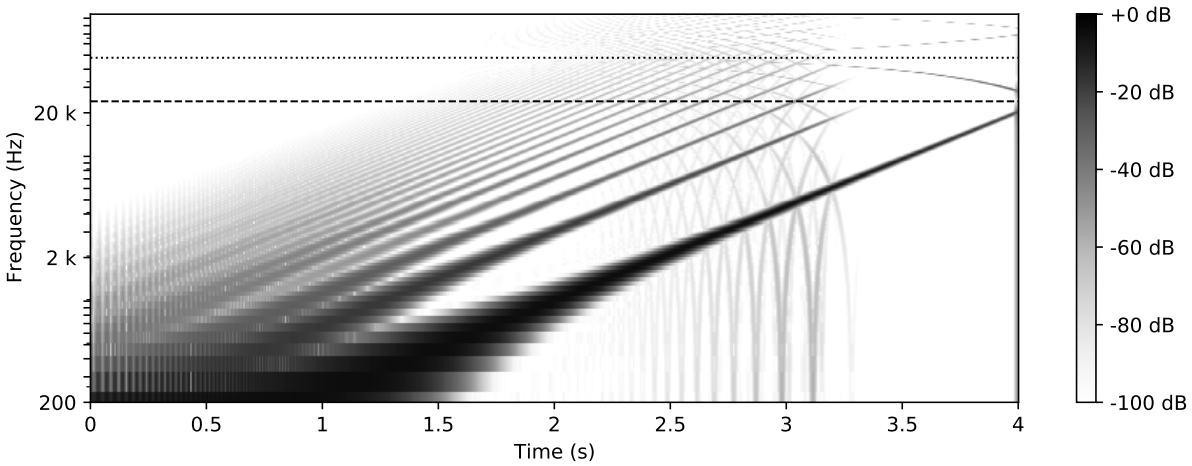


(c) projection order $p = 3$

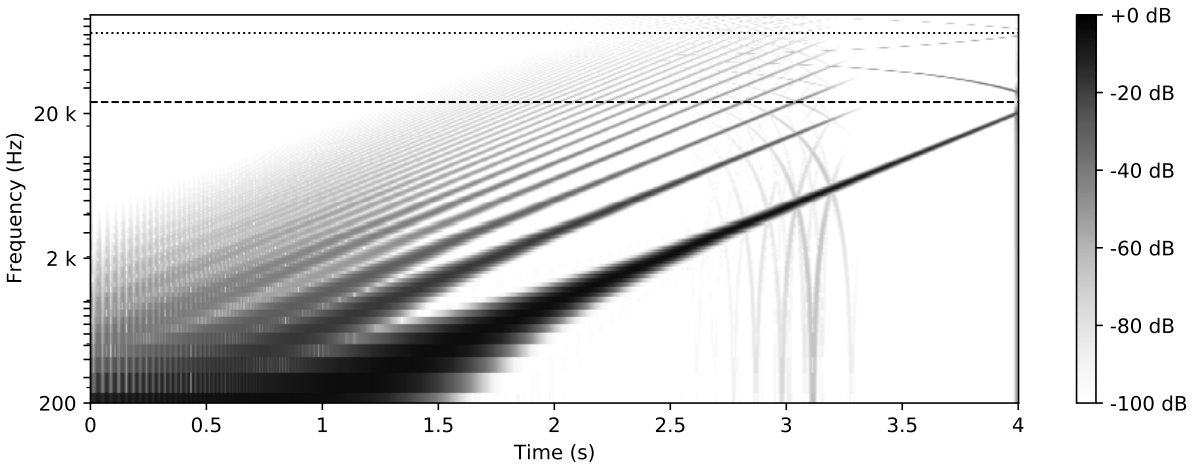
Figure 5.15 – (Diode clipper) Sinesweep spectrograms for $p = 1, 2, 3$, $k = 0$ with $R = 1$ kΩ, $C = 20$ nF such that the diode clipper cutoff is $f_c = 20$ kHz for a *fixed sampling frequency* $f_s = 48$ kHz. We use $I_S = 1$ fA and $V_T = 26$ mV and an input gain $g = 1.5$. The spectrum above the Nyquist frequency (24 kHz) is delimited by a dashed blacked line. The generalized bandwidth $f_p = pf_s/2$ is shown in dotted black. The non-bandlimited modelling power (and aliasing rejection) of high order projection clearly becomes more efficient as the projection order is increased.



(a) projection order $p = 1$



(b) projection order $p = 2$



(c) projection order $p = 3$

Figure 5.16 – (Diode clipper) Sinesweep spectrograms in logarithmic frequency scale (same simulation) to be compared with figure 5.15.

Discussion and perspectives for stiff dissipative systems

We reconsider the power balance functional ρ from chapter 4 in the case of an autonomous pH-ODE. In this chapter, using a self-adjoint scalar projector $\mathcal{P}$, we have by commmutation of $(\mathcal{P}, \mathbf{J} - \mathbf{R})$, self-adjointness of $\mathcal{P}$ and skew-symmetry of matrix $\mathbf{J}$

$$\rho(\boldsymbol{X}) = \left\langle \nabla H(\boldsymbol{X}) \,\middle|\, \boldsymbol{f}(\boldsymbol{X}) - \dot{\boldsymbol{X}} \right\rangle = \left\langle \nabla H(\boldsymbol{X}) \,\middle|\, (\mathbf{J} - \mathbf{R})(\mathcal{I} - \mathcal{P})\nabla H(\boldsymbol{X}) \right\rangle = -\left\| (\mathcal{I} - \mathcal{P})\nabla H(\boldsymbol{X}) \right\|_{\mathbf{R}}^2 .$$

This means that, after projection, conservative systems, are still unconditionally conservative and dissipative systems are still unconditionally dissipative. For conservative systems, the energy preservation is exact (since $\mathbf{R} = \mathbf{0}$). But for dissipative systems, comparing the functional projection approach in this chapter with the adaptive collocation strategy from chapter 4 p.107, the price to pay for unconditional passivity (and linear parametrization of the problem using projection coefficients) is an error on the dissipation rate which is in $\mathcal{O}\left(\left\| (\mathcal{I} - \mathcal{P})\nabla H(\boldsymbol{X}) \right\|_{\mathbf{R}}^2\right)$.

A perspective, for stiff dissipative systems (see oscillations in figure 5.14), is to combine the unconditional energy dissipation of RPM (see also [HL14]) with damping for infinitely damped poles (as in L-stable methods such as Radau IIa [HLW06]) while optimising the decay rate. A path towards this goal would be to combine a) the continuous-time functional projection in this chapter, b) the exact preservation (or minimisation) of the power-balance functional $\rho(\boldsymbol{X}) = 0$ introduced in (S)PAC methods.

Conclusion

In this chapter, we have demonstrated that representing flows and and efforts as functions of time in the Hilbert space L^2 (used as a pivot space) coupled with respectively skew-adjoint and self-adjoint approximations of PH structure matrices $\mathbf{J}$ and $\mathbf{R}$ (using projectors) is a key ingredient to yield energy-preserving and passivity-preserving methods for both pH-ODEs and pH-DAEs. Coupling this result with supplementary boundary conditions, we have proposed a class of methods called RPM(p, k) that satisfy properties $\mathbf{P1}, \mathbf{P2}, \mathbf{P3}$ (power-balance, accuracy, regularity) and whose principle is applicable to both pH-ODEs and pH-DAEs. A detailed analysis of RPM for ODE has been proposed where accuracy order, existence and uniqueness, local accuracy, Peano error kernels, etc have been studied. Works remains to be done in the case of DAE. First results show that the PH structure and its tree/cotree partitioning can be exploited advantageously. In particular, we were able to show that the Jacobian in Newton iteration is always invertible for convex Hamiltonians and incrementally passive dissipative component laws. The main advantages and drawbacks of the approach are listed below.

Advantages

- Unconditional energy preservation and passivity,
- Representation is linear in the parameters,
- Spectral projection converges exponentially fast for smooth functions,
- The method can be interpretated using the framework of CSRK methods,
- Order conditions directly stems from to the polynomial reproduction property of projectors,
- Orthonormal basis have optimal numerical conditioning and require less computations.

Drawbacks

- Projections integrals need to be computed exactly to have energy conservation,
- High orders require quadrature approximations (up to machine accuracy),
- Inexact dissipation rate and lack of damping for infinitely stiff systems.
- Regularity is a post-regularisation step rather than a built-in feature[17]: the increased regularity and local accuracy of projector $\mathcal{Q}$ does not improve the time-stepping accuracy.

> **Remark 5.3** (Discrete PHS)**.** Comparing with the discrete PHS definition proposed in [KL19], which is based on symplectic integration (such as Gauss-Legendre schemes), a main difference is that the functional projection approach in this chapter preserves the exact Hamiltonian (and passivity) while symplectic integrators preserve the symplectic structure (and possess a perturbed Hamiltonian).

17. The main reason is that orthogonal projection in H^k is not orthogonal in L^2. Since the power-balance is intimately linked to the L^2 inner product, we cannot choose a different inner product even if we look for regular solutions in H^k. However we can interpret L^2 solutions as weak solutions and H^k solutions as stronger solutions where the regularisation step is compatible with L^2 projection.

Chapter 6

Power-balanced Exponential Integrators

First Law of Numerical Analysis: Analytical and Numerical Difficulties Always Come Paired

J.W.Neuberger, "Sobolev Gradients and differential equations", [Neu09]

Contents

6.1 From functional Newton iteration to exponential integrators 160
6.2 Exponential Average Vector Field method 162
6.2.1 Notations and preliminary definitions 162
6.2.2 Energy preserving (resp. dissipating) Exponential AVF 163
6.2.3 LC example 165
6.2.4 Adding external ports 166
6.3 High-order energy-preserving exponential integrators 168

In the previous chapter, we have used functional orthogonal projection. It minimises the L^2-norm of the residual error between the exact and the projected vector field and preserves the power balance. In this chapter, we combine vector field projection with exponential integrators to obtain energy-preserving exponential integrators. A salient feature of exponential integrators is that they exactly integrates the (local) linear dynamics [1].

- In section 6.1 we motivate the choice of exponential integrators by showing that they naturally arise as optimal pre-conditioners in functional Newton iteration when minimizing the L^2-norm of the vector field residual error.

- In section 6.2, we propose an extended definition of the AVF discrete gradient and show how to combine it with exponential integrators to yield an energy (resp. dissipation) preserving numerical scheme.

- In section 6.3, we generalise this approach to power-balanced integrators with arbitrary high projection orders and basis functions.

1. This is a way to increase accuracy and to manage stiffness of the equations.

6.1 From functional Newton iteration to exponential integrators

In this section, instead of pre-specifying a finite-dimensional approximation space, we seek a solution of the ODE using infinite-dimensional Newton iteration in functional space. As a byproduct, exponential integrators naturally arise as pre-conditioners for Newton iteration [2].

Consider an autonomous ODE over a time interval Ω, governed by

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{f}(\boldsymbol{x}(t)), \quad \boldsymbol{x}(0) = \mathbf{x}_0 \in \mathbb{R}^n,$$

with $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^n$. We define the diffential operator $\mathcal{D} := \frac{\mathrm{d}}{\mathrm{d}t}$, and the residual vector field operator $\boldsymbol{E} : \mathcal{X} := H^1(\Omega, \mathbb{R}^n) \to L^2(\Omega, \mathbb{R}^n)$ by

$$\boldsymbol{E}(\boldsymbol{x}) := \mathcal{D}\boldsymbol{x} - \boldsymbol{f}(\boldsymbol{x}). \tag{6.1}$$

For an initial *trajectory* function $\boldsymbol{x}^0 \in \mathcal{X}$, we propose to formally solve the following minimisation problem using functional Newton iteration

$$\begin{aligned} \boldsymbol{x}^\star = \operatorname*{argmin}_{\boldsymbol{x}\in\mathcal{X}} \quad & \Phi(\boldsymbol{x}) = \tfrac{1}{2}\left\|\boldsymbol{E}(\boldsymbol{x})\right\|_{L^2}^2, \\ \text{s.t.} \quad & \boldsymbol{x}(0) = \mathbf{x}_0. \end{aligned} \tag{6.2}$$

Newton iteration consists in locally approximating the functional Φ about each functional iterate $\boldsymbol{x}^k$ by a *convex positive definite quadratic functional* $\tilde{\Phi}$ (detailed below) and solving the associated sequence of least-square problems. The Newton-Kantorovich theorem guarantees convergence with quadratic speed when the initial estimate is in the bassin of attraction of the solution (not detailed here). Using Frechet derivatives (see definition C.8 p.282), an extrema of the functional Φ corresponds to a zero of its first-order derivative

$$\Phi'(\boldsymbol{x})(\boldsymbol{u}) = \left\langle \boldsymbol{E}'_{\boldsymbol{x}}(\boldsymbol{u}), \boldsymbol{E}(\boldsymbol{x}) \right\rangle = \mathbf{0}, \tag{6.3}$$

where by definition of the Frechet derivative, $\boldsymbol{E}'_{\boldsymbol{x}}$ is the *linear operator* at $\boldsymbol{x}$ acting on $\boldsymbol{u}$ given by

$$\boldsymbol{E}'_{\boldsymbol{x}}(\boldsymbol{u}) = (\mathcal{D} - \mathbf{A}_{\boldsymbol{x}})\,\boldsymbol{u}, \qquad \text{where} \qquad \mathbf{A}_{\boldsymbol{x}} = \boldsymbol{f}'(\boldsymbol{x}). \tag{6.4}$$

To have a local minimum at each iteration, it is sufficient that the Hessian approximation $\Phi'' \approx \widetilde{\Phi}'' \succeq 0$ be a positive define bilinear form. For that purpose, we use the following *convex positive semi-definite approximation* [3] of the second Fréchet derivative

$$\widetilde{\Phi}''(\boldsymbol{x})(\boldsymbol{u}, \boldsymbol{v}) = \left\langle \boldsymbol{E}'_{\boldsymbol{x}}(\boldsymbol{u}), \boldsymbol{E}'_{\boldsymbol{x}}(\boldsymbol{v}) \right\rangle \geq 0. \tag{6.5}$$

Note that, $\widetilde{\Phi}''(\boldsymbol{x})$ being a positive bilinear form, it defines, for each function $\boldsymbol{x}$, a *Sobolev inner product*

$$\langle \boldsymbol{u}, \boldsymbol{v} \rangle_{\boldsymbol{E}'_{\boldsymbol{x}}} := \left\langle \boldsymbol{E}'_{\boldsymbol{x}}(\boldsymbol{u}), \boldsymbol{E}'_{\boldsymbol{x}}(\boldsymbol{v}) \right\rangle_{L^2}. \tag{6.6}$$

where the local metric is given by the linear self-adjoint differential operator

$$\mathcal{W} = (\boldsymbol{E}'_{\boldsymbol{x}})^*(\boldsymbol{E}'_{\boldsymbol{x}}) = (\mathcal{D} - \mathbf{A}_{\boldsymbol{x}})^*(\mathcal{D} - \mathbf{A}_{\boldsymbol{x}}). \tag{6.7}$$

2. Note that our goal is to guide the choice of optimal approximation space, not to actually implement Newton iteration in infinite-dimensional space. To focus on the idea, and not on functional details, the adjoint and inverse operators below are formal. Technically here, we assume that $\boldsymbol{f}$ is locally Lipschitz and that fixed-point Picard iteration converges to a unique solution, so that Newton iteration is only considered as a convergence acceleration tool. A similar derivation is available in [LC12].

3. Note that $\Phi''(\boldsymbol{x})(\boldsymbol{u}, \boldsymbol{v}) = \left\langle \boldsymbol{E}'_{\boldsymbol{x}}(\boldsymbol{u}), \boldsymbol{E}'_{\boldsymbol{x}}(\boldsymbol{v}) \right\rangle + \left\langle \boldsymbol{E}''_{\boldsymbol{x}}(\boldsymbol{u}, \boldsymbol{v}), \boldsymbol{E}(\boldsymbol{x}) \right\rangle$. We (classically) neglect the second term which is assumed to be small compared to the first term in a neighbourhood sufficiently close to a minimum.

An exponential integrator in disguise Starting from an initial functional estimate $\boldsymbol{x}^0 \in \mathcal{X}$, the approximate Newton step is formally given for all $k \geq 0$ by

$$\Delta \boldsymbol{x}^{k+1} = -\left[\widetilde{\Phi}''(\boldsymbol{x}^k)\right]^{-1} \Phi'(\boldsymbol{x}^k). \tag{6.8}$$

combined with a line-search iteration $\boldsymbol{x}^{k+1} = \boldsymbol{x}^k + \alpha \Delta \boldsymbol{x}^k$, $\alpha \in [0,1]$. Note that if $\boldsymbol{E}'_{\boldsymbol{x}}$ is invertible, we can simplify the pseudo-inverse in the Newton step as follows:

$$\Delta \boldsymbol{x} = -\left[\widetilde{\Phi}''(\boldsymbol{x})\right]^{-1} \Phi'(\boldsymbol{x}) = -\left[(\boldsymbol{E}'_{\boldsymbol{x}})^*(\boldsymbol{E}'_{\boldsymbol{x}})\right]^{-1} (\boldsymbol{E}_{\boldsymbol{x}})^* \boldsymbol{E}(\boldsymbol{x}) = -(\boldsymbol{E}'_{\boldsymbol{x}})^{-1} \boldsymbol{E}(\boldsymbol{x}) \tag{6.9}$$

where from (6.4), the inverse operator $(\boldsymbol{E}'_{\boldsymbol{x}})^{-1}$ is nothing but an exponential integrator

$$\left[(\boldsymbol{E}'_{\boldsymbol{x}})^{-1}\boldsymbol{u}\right](t) = \int_0^1 \exp\left(\int_s^t \mathbf{A}(\boldsymbol{x}(\xi))\,\mathrm{d}\xi\right) \Theta(t-s)\boldsymbol{u}(s)\,\mathrm{d}s. \tag{6.10}$$

It plays the role of a preconditioner applied to the residual $\boldsymbol{E}(\boldsymbol{x})$.

> **Remark 6.1.** The role of a Newton preconditioner is to enhance convergence and conditioning [Deu11] but it does not change the solution (of the fixed-point Picard iteration). Furthermore, since the exact operator can be difficult to approximate, we may instead use the following tractable approximation (simplified Newton iteration)
>
> $$\left[(\boldsymbol{E}'_{\boldsymbol{x}})^{-1}\boldsymbol{u}\right](t) \approx \int_0^1 \exp\left(\mathbf{A}_0(t-s)\right) \Theta(t-s)\boldsymbol{u}(s)\,\mathrm{d}s, \quad \text{where} \quad \mathbf{A}_0 = \boldsymbol{f}'(\mathbf{x}_0). \tag{6.11}$$

Functional Newton iteration automatically generates an exponential integrator $(\mathcal{D} - \mathbf{A}_{\boldsymbol{x}})^{-1}$ to precondition the residual $\boldsymbol{E}(\boldsymbol{x})$.

Sobolev Gradients Using the theory of Sobolev gradients [Neu09] and the Riesz representation theorem (see C.1), there exists respectively L^2 and Sobolev gradients $\nabla\Phi$ and $\nabla_S\Phi$ such that the Fréchet derivative can be represented either using the L^2 or the Sobolev inner product as

$$\Phi'(\boldsymbol{x})(\boldsymbol{u}) = \left\langle \nabla\Phi(\boldsymbol{x}), \boldsymbol{u}\right\rangle_{L^2} = \left\langle \nabla_S\Phi(\boldsymbol{x}), \boldsymbol{u}\right\rangle_{\boldsymbol{E}'_{\boldsymbol{x}}}, \tag{6.12}$$

where from (6.3) and (6.9) we find that

$$\nabla\Phi(\boldsymbol{x}) = (\boldsymbol{E}'_{\boldsymbol{x}})^*\boldsymbol{E}(\boldsymbol{x}), \qquad \nabla_S\Phi(x) = (\boldsymbol{E}'_{\boldsymbol{x}})^{-1}\boldsymbol{E}(\boldsymbol{x}). \tag{6.13}$$

Likewise there exists L^2 and Sobolev Hessians $\nabla^2\widetilde{\Phi}$ and $\nabla^2_S\widetilde{\Phi}$ such that $\nabla^2_S\widetilde{\Phi}$ is the identity.

$$\widetilde{\Phi}''(\boldsymbol{x})(\boldsymbol{u},\boldsymbol{v}) = \left\langle \boldsymbol{u} \,\middle|\, \nabla^2\widetilde{\Phi}(\boldsymbol{x}) \,\middle|\, \boldsymbol{v}\right\rangle_{L^2} = \langle \boldsymbol{u}, \boldsymbol{v}\rangle_{\boldsymbol{E}'_{\boldsymbol{x}}}.$$

Indeed, from (6.5), using the formal adjoint $(\boldsymbol{E}'_{\boldsymbol{x}})^*$, we find that $\nabla^2\widetilde{\Phi} = (\boldsymbol{E}'_{\boldsymbol{x}})^*\boldsymbol{E}'_{\boldsymbol{x}}$ (see (6.7)), and we can express the Fréchet derivatives as

$$\widetilde{\Phi}''(\boldsymbol{x})(\boldsymbol{u},\boldsymbol{v}) = \langle \boldsymbol{u}, \boldsymbol{v}\rangle_{\boldsymbol{E}'_{\boldsymbol{x}}} = \left\langle \boldsymbol{E}'_{\boldsymbol{x}}(\boldsymbol{u}), \boldsymbol{E}'_{\boldsymbol{x}}(\boldsymbol{v})\right\rangle = \left\langle \boldsymbol{u} \,\middle|\, (\boldsymbol{E}'_{\boldsymbol{x}})^*\boldsymbol{E}'_{\boldsymbol{x}} \,\middle|\, \boldsymbol{v}\right\rangle = \left\langle \boldsymbol{u} \,\middle|\, \nabla^2\Phi(\boldsymbol{x}) \,\middle|\, \boldsymbol{v}\right\rangle.$$

According to (6.9) and (6.13) we may conclude that

Functional Newton iteration is equivalent to steepest gradient descent in Sobolev space.

6.2 Exponential Average Vector Field method

6.2.1 Notations and preliminary definitions

Let $\langle \mathbf{u}, \mathbf{v} \rangle := \mathbf{u}^\mathsf{T}\mathbf{v}$, and $\|\mathbf{u}\| := \sqrt{\langle \mathbf{u}, \mathbf{u} \rangle}$ denote the euclidean inner product and norm in $\mathbb{R}^n$. For an invertible symmetric positive definite matrix $\mathbb{R}^{n\times n} \ni \mathbf{Q} = \mathbf{Q}^\mathsf{T} \succ 0$, we define the associated inner product and norm by $\langle \mathbf{u}, \mathbf{v} \rangle_\mathbf{Q} := \mathbf{u}^\mathsf{T}\mathbf{Q}\mathbf{v}$, and $\|\mathbf{u}\|_\mathbf{Q} := \langle \mathbf{u}, \mathbf{u} \rangle_\mathbf{Q}^{1/2}$.

Definition 6.1. Let $H : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function on $\mathbb{R}^n$. Using the Riesz representation theorem, we define the euclidean gradient ∇H and the $\mathbf{Q}$-gradient $\nabla_\mathbf{Q} H$ as the unique elements satisfying

$$H'(\mathbf{x})(\cdot) = \left\langle \cdot, \nabla H(\mathbf{x}) \right\rangle_{\mathbb{R}^n} = \left\langle \cdot, \nabla_\mathbf{Q} H(\mathbf{x}) \right\rangle_\mathbf{Q}, \quad \forall \mathbf{x} \in \mathbb{R}^n. \tag{6.14}$$

where $H'(\mathbf{x})(\cdot)$ denotes the Frechet derivative of H at $\mathbf{x}$. It follows that $\nabla_\mathbf{Q} H(\cdot) = \mathbf{Q}^{-1}\nabla H(\cdot)$.

Lemma 6.1. *Let $\mathbf{A} = (\mathbf{J} - \mathbf{R})\mathbf{Q}$ with $\mathbb{R}^{n\times n}$ matrices $\mathbf{J} = -\mathbf{J}^\mathsf{T}$, $\mathbf{R} = \mathbf{R}^\mathsf{T} \succeq 0$ and $\mathbf{Q} = \mathbf{Q}^\mathsf{T} \succ 0$. Then the semigroup $e^{t\mathbf{A}}$ is norm-preserving (resp. non expansive) in the $\mathbf{Q}$-norm, i.e. it satisfies for all $\mathbf{u} \in \mathbb{R}^n$, for all $t \geq 0$*

$$\left\| e^{t\mathbf{A}}\mathbf{u} \right\|_\mathbf{Q} = \|\mathbf{u}\|_\mathbf{Q} \quad \textit{if} \quad \mathbf{R} = \mathbf{0}, \qquad \textit{otherwise} \qquad \left\| e^{t\mathbf{A}}\mathbf{u} \right\|_\mathbf{Q} \leq \|\mathbf{u}\|_\mathbf{Q} \quad \textit{if} \quad \mathbf{R} \succeq \mathbf{0}. \tag{6.15}$$

Proof. For a pH-ODE $\dot{\boldsymbol{x}} = (\mathbf{J} - \mathbf{R})\mathbf{Q}\boldsymbol{x}$, $\boldsymbol{x}(0) = \mathbf{x}_0$, the Hamiltonian $H(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_\mathbf{Q}^2$ is preserved (resp. dissipated) along the solution $\boldsymbol{x}(t) = e^{t\mathbf{A}}\mathbf{x}_0$ (see equation (1.49) p.33). □

Definition 6.2. Let $\Omega = (0, 1)$, We define the orthogonal averaging projector $\mathcal{P} : L^2(\Omega, \mathbb{R}^n) \to L^2(\Omega, \mathbb{R}^n)$, and the associated Sobolev projector $\mathcal{P}_S : H^1(\Omega, \mathbb{R}^n) \to H^1(\Omega, \mathbb{R}^n)$ by

$$(\mathcal{P}\boldsymbol{u})(\tau) := \int_0^1 \boldsymbol{u}(\sigma)\,\mathrm{d}\sigma, \qquad (\mathcal{P}_S\boldsymbol{u})(\tau) := \boldsymbol{u}(0) + \int_0^\tau (\mathcal{P}\dot{\boldsymbol{u}})(\sigma)\,\mathrm{d}\sigma. \tag{6.16}$$

In particular, they satisfy the commutation identity $\frac{\mathrm{d}}{\mathrm{d}\tau}(\mathcal{P}_S\boldsymbol{u}) = \mathcal{P}(\frac{\mathrm{d}}{\mathrm{d}\tau}\boldsymbol{u}) = \boldsymbol{u}_1 - \boldsymbol{u}_0$.

Using these operators we give an extended functional definition of the average discrete gradient

Thm-definition 6.1 (Generalized Average Discrete ($\mathbf{Q}$)-Gradient)**.** Let $V \in \mathcal{C}^1(\mathbb{R}^n, \mathbb{R})$. For all $\boldsymbol{x} \in H^1(\Omega, \mathbb{R}^n)$, we define the *generalized average discrete gradient* (GADG)

$$\overline{\overline{\nabla}} V(\boldsymbol{x}) := (\mathcal{P} \circ \nabla V \circ \mathcal{P}_S)(\boldsymbol{x}). \tag{6.17}$$

and the discrete $\mathbf{Q}$-gradient $\overline{\overline{\nabla}}_\mathbf{Q} V(\boldsymbol{x}) := \mathbf{Q}^{-1}\overline{\overline{\nabla}} V(\boldsymbol{x})$ satisfying the discrete gradient identity

$$V(\boldsymbol{x}_1) - V(\boldsymbol{x}_0) = \left\langle \overline{\overline{\nabla}} V(\boldsymbol{x}), \boldsymbol{x}_1 - \boldsymbol{x}_0 \right\rangle_{\mathbb{R}^n} = \left\langle \overline{\overline{\nabla}}_\mathbf{Q} V(\boldsymbol{x}), \boldsymbol{x}_1 - \boldsymbol{x}_0 \right\rangle_\mathbf{Q}. \tag{6.18}$$

Proof. $V(\boldsymbol{x}_1) - V(\boldsymbol{x}_0) \overset{a}{=} \left\langle \nabla V(\mathcal{P}_S\boldsymbol{x}) \,\middle|\, \frac{\mathrm{d}}{\mathrm{d}\tau}(\mathcal{P}_S\boldsymbol{x}) \right\rangle_{L^2} \overset{b}{=} \left\langle \nabla V(\mathcal{P}_S\boldsymbol{x}) \,\middle|\, \mathcal{P}\dot{\boldsymbol{x}} \right\rangle_{L^2} \overset{c}{=} \left\langle \overline{\overline{\nabla}} V(\boldsymbol{x}), \boldsymbol{x}_1 - \boldsymbol{x}_0 \right\rangle_{\mathbb{R}^n}$. using (a) the gradient theorem, (b) $\frac{\mathrm{d}}{\mathrm{d}\tau}(\mathcal{P}_S\boldsymbol{x}) = \mathcal{P}(\frac{\mathrm{d}}{\mathrm{d}\tau}\boldsymbol{x})$, (c) $\mathcal{P}^2 = \mathcal{P} = \mathcal{P}^*$ and $\mathcal{P}\dot{\boldsymbol{x}} = \boldsymbol{x}_1 - \boldsymbol{x}_0$. □

6.2.2 Energy preserving (resp. dissipating) Exponential AVF

We consider (for each time frame) the semi-linear splitting of an autonomous pH-ODE

$$\dot{\boldsymbol{x}} = (\mathbf{J} - \mathbf{R})(\mathbf{Q}\boldsymbol{x} + \nabla V(\mathbf{x})), \qquad \boldsymbol{x}(t_0) = \mathbf{x}_0, \tag{6.19}$$

with matrices $\mathbf{J} = -\mathbf{J}^\top$, $\mathbf{R} = \mathbf{R}^\top \succeq 0$, and $\mathbf{Q} = \mathbf{Q}^\top \succ 0$. The Hamiltonian is decomposed as

$$H(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_{\mathbf{Q}}^2 + V(\mathbf{x}). \tag{6.20}$$

A typical choice for $H \in \mathcal{C}^2$ is to use the Hessian $\mathbf{Q} = \nabla^2 H(\bar{\mathbf{x}})$ about an expansion point $\bar{\mathbf{x}}$ and define the potential V as the difference $V(\mathbf{x}) = H(\mathbf{x}) - \frac{1}{2}\|\mathbf{x}\|_{\mathbf{Q}}^2$.

Introducing the linear operator $\boldsymbol{\mathcal{L}} = \frac{\mathrm{d}}{\mathrm{d}\tau} - \mathbf{A}$, with matrix $\mathbf{A} = h(\mathbf{J} - \mathbf{R})\mathbf{Q}$, we rewrite (6.19) as the normalized-time initial value problem.

$$\boldsymbol{\mathcal{L}}\,\boldsymbol{x} = \mathbf{A}\nabla_{\mathbf{Q}} V(\boldsymbol{x}), \quad \tau \in (0,1) \qquad \boldsymbol{x}(0) = \mathbf{x}_0, \tag{6.21}$$

where $\nabla_{\mathbf{Q}} V = \mathbf{Q}^{-1}\nabla V$ denotes the $\mathbf{Q}$-gradient [4] of V and $t = t_0 + h\tau$, for $\tau \in [0,1]$.

Theorem 6.1. *If $\boldsymbol{X}(\tau)$ is the solution of the projected Initial Value Problem (6.21) using the generalized average discrete gradient (6.17)*

$$\boldsymbol{\mathcal{L}}\boldsymbol{X} = \mathbf{A}\overline{\overline{\nabla}}_{\mathbf{Q}} V(\boldsymbol{X}), \qquad \boldsymbol{X}(0) := \mathbf{x}_0. \tag{6.22}$$

Then the time stepping $\Phi : \mathbf{x}_0 \mapsto \mathbf{x}_1 = \boldsymbol{X}(1)$ is energy (resp. passivity) preserving i.e.

$$H(\mathbf{x}_1) - H(\mathbf{x}_0) \quad = 0 \quad \textit{if} \quad \mathbf{R} = \mathbf{0} \quad \textit{otherwise} \quad \leq 0 \quad \textit{if} \quad \mathbf{R} \succeq \mathbf{0}. \tag{6.23}$$

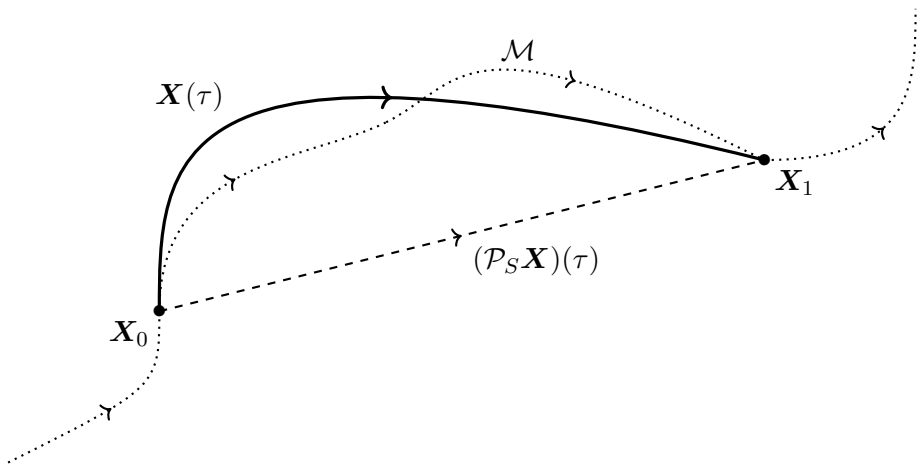


Figure 6.1 – (Exponential AVF) Schematic description of the method. The linear part $\dot{\boldsymbol{x}} = \mathbf{Q}\boldsymbol{x}$ of the ODE is exactly integrated by the exponential integrator. The nonlinear part $\overline{\overline{\nabla}} V(\boldsymbol{X}) = \mathcal{P}\nabla V(\mathcal{P}_S\boldsymbol{X})$ is averaged along the trajectory $\mathcal{P}_S\boldsymbol{X}$ where by construction $\boldsymbol{X}$ and $\mathcal{P}_S\boldsymbol{X}$ share the same endpoints on the manifold $\mathcal{M} = \{\mathbf{x} \in \mathbb{R}^n \mid H(\mathbf{x}) = H(\mathbf{x}_0)\}$ and thus the same average slope.

4. The reason for using the $\mathbf{Q}$-gradient and the $\mathbf{Q}$-norm will become apparent in the proof of theorem 6.1.

Proof. We prove the result in three steps:

i) Using the Green kernel $e^{(\tau-\sigma)\mathbf{A}}\Theta(\tau-\sigma)$ of operator $\boldsymbol{\mathcal{L}}$, we have the trajectory

$$\boldsymbol{X}(\tau) = e^{\tau\mathbf{A}}\mathbf{x}_0 + \int_0^\tau e^{(\tau-\sigma)\mathbf{A}}\mathbf{A}\,\mathrm{d}\sigma\overline{\overline{\nabla}}_{\mathbf{Q}}V(\boldsymbol{X}) = e^{\tau\mathbf{A}}\mathbf{x}_0 + \left(e^{\tau\mathbf{A}} - \mathbf{I}\right)\overline{\overline{\nabla}}_{\mathbf{Q}}V(\boldsymbol{X}).$$

ii) It follows after evaluating $\mathbf{x}_1 = \boldsymbol{X}(1)$ and factorisation that we have the identities

$$\mathbf{x}_1 - \mathbf{x}_0 = \left(e^{\mathbf{A}} - \mathbf{I}\right)\left(\mathbf{x}_0 + \overline{\overline{\nabla}}_{\mathbf{Q}}V(\boldsymbol{X})\right), \tag{6.24a}$$

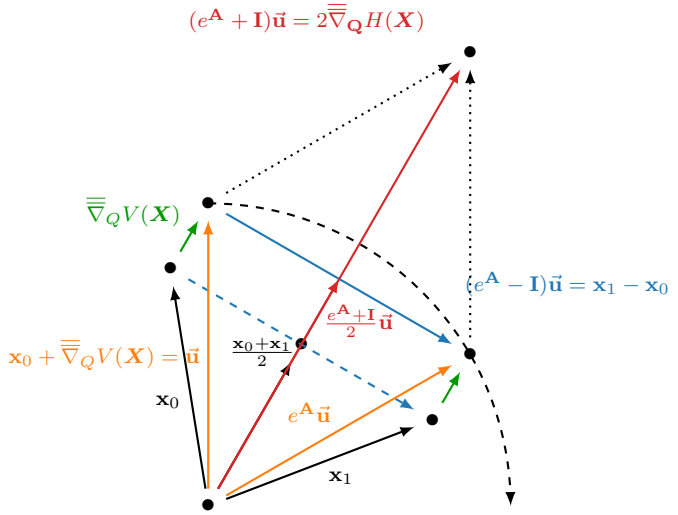
$$\overline{\overline{\nabla}}_{\mathbf{Q}}H(\boldsymbol{X}) = \frac{\mathbf{x}_1 + \mathbf{x}_0}{2} + \overline{\overline{\nabla}}_{\mathbf{Q}}V(\boldsymbol{X}) = \left(e^{\mathbf{A}} + \mathbf{I}\right)\left(\mathbf{x}_0 + \overline{\overline{\nabla}}_{\mathbf{Q}}V(\boldsymbol{X})\right)\frac{1}{2}. \tag{6.24b}$$

iii) Energy-conservation (resp. dissipation) follows from the sequence of relations

$$\begin{aligned} H(\mathbf{x}_1) - H(\mathbf{x}_0) &\overset{a}{=} \left\langle \overline{\overline{\nabla}}_{\mathbf{Q}}H(\boldsymbol{X}), \mathbf{x}_1 - \mathbf{x}_0 \right\rangle_{\mathbf{Q}} \\ &\overset{b}{=} \frac{1}{2}\left\langle \left(e^{\mathbf{A}} + \mathbf{I}\right)\left(\mathbf{x}_0 + \overline{\overline{\nabla}}_{\mathbf{Q}}V(\boldsymbol{X})\right), \left(e^{\mathbf{A}} - \mathbf{I}\right)\left(\mathbf{x}_0 + \overline{\overline{\nabla}}_{\mathbf{Q}}V(\boldsymbol{X})\right)\right\rangle_{\mathbf{Q}} \\ &\overset{c}{=} \frac{1}{2}\left\| e^{\mathbf{A}}\left(\mathbf{x}_0 + \overline{\overline{\nabla}}_{\mathbf{Q}}V(\boldsymbol{X})\right)\right\|_{\mathbf{Q}}^2 - \frac{1}{2}\left\|\mathbf{x}_0 + \overline{\overline{\nabla}}_{\mathbf{Q}}V(\boldsymbol{X})\right\|_{\mathbf{Q}}^2 \qquad (6.24c) \\ &\overset{d}{=} 0 \text{ if } \mathbf{R} = \mathbf{0} \quad \text{and} \quad \overset{d}{\leq} 0 \text{ if } \mathbf{R} \succeq \mathbf{0}. \qquad (6.24d) \end{aligned}$$

using (a) the discrete $\mathbf{Q}$-gradient identity (6.18), (b) identities from step ii), (c) expansion of the inner product (with vanishing cross terms), (d) lemma 6.1 with $\mathbf{u} = \mathbf{x}_0 + \overline{\overline{\nabla}}_{\mathbf{Q}}V(\boldsymbol{X})$.

A geometric interpretation of the proof is shown on figure 6.2. □



(a) Geometry

1) Form $\mathbf{u} = \mathbf{x}_0 + \overline{\overline{\nabla}}_{\mathbf{Q}}V(\boldsymbol{X})$ and its rotation $e^{\mathbf{A}}\mathbf{u}$ in orange, form the differences $\mathbf{e}^{\mathbf{A}}\mathbf{u} - \mathbf{u}$ and $\mathbf{x}_1 - \mathbf{x}_0$ (in blue and dashed blue). We get (6.24a) .
2) Form the sum $\mathbf{u} + \mathbf{e}^{\mathbf{A}}\mathbf{u}$ and its half, form the barycenter $(\mathbf{x}_0 + \mathbf{x}_1)/2$ (red) and add $\overline{\overline{\nabla}}_{\mathbf{Q}}V(\boldsymbol{X})$ (green), we get (6.24b)
3) Note the orthogonality in (6.24d) $\mathbf{x}_1 - \mathbf{x}_0 \perp_{\mathbf{Q}} \overline{\overline{\nabla}}_{\mathbf{Q}}H(\boldsymbol{X})$

(b) Explanations

Figure 6.2 – (Exponential AVF) Visual illustration of geometric objects in the proof of theorem 6.1. For simplicity, we consider the conservative case $\mathbf{R} = 0$, $\mathbf{Q} = \mathbf{I}$. Energy conservation stems from the orthogonality $\mathbf{x}_1 - \mathbf{x}_0 \perp_{\mathbf{Q}} \overline{\overline{\nabla}}_{\mathbf{Q}}H(\boldsymbol{X})$.

6.2.3 LC example

In order to perform a comparison betwen the AVF method (i.e. projection order $p = 1$, regularity $k = 0$) and the Exponential AVF method, we reconsider the nonlinear LC example of subsection 5.5.1.

$$\begin{bmatrix} \dot{q} \\ \dot{\phi} \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \nabla H_C(q) \\ \nabla H_L(\phi) \end{bmatrix}, \qquad \mathbf{Q} = \nabla^2 H(q_0, \phi_0) = \begin{bmatrix} \frac{1}{C} & 0 \\ 0 & \frac{1}{L_0}, \end{bmatrix}$$

where the Hessian $\mathbf{Q}$ is governed by the local inductance $L_0 = L/(1 - \tanh^2(\phi_0/LI_S))$. We decompose the Hamiltonian as $H(q, \phi) = \frac{q^2}{2C} + \frac{\phi^2}{2L_0} + V(\phi)$ where

$$V(\phi) = LI_S^2 \ln\cosh\left(\frac{\phi}{LI_S}\right) - \frac{\phi^2}{2L_0} = LI_S^2 \ln\cosh\left(\frac{\phi_0}{LI_S}\right) + \mathcal{O}\left(\left(\frac{\phi - \phi_0}{LI_S}\right)^4\right). \tag{6.25}$$

Using (6.24a), we solve the fixed point equation on $(\delta q, \delta\phi)$

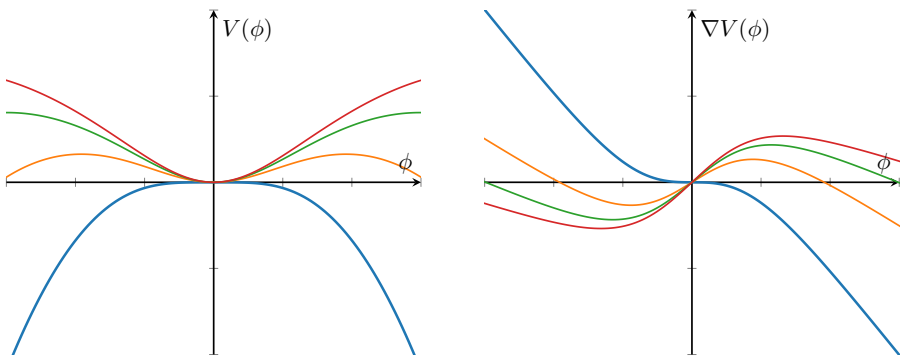
$$\begin{bmatrix} \delta q \\ \delta\phi \end{bmatrix} = (e^{\mathbf{A}} - \mathbf{I}) \left(\begin{bmatrix} q_0 \\ \phi_0 \end{bmatrix} + \begin{bmatrix} 0 \\ L_0 \overline{\overline{\nabla}} V(\phi) \end{bmatrix} \right), \qquad \mathbf{A} = h \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{C} & 0 \\ 0 & \frac{1}{L_0} \end{bmatrix}, \tag{6.26}$$

using the closed-form formula of the AVF discrete gradient (see Equation 5.41)

$$\overline{\overline{\nabla}} V(\phi) = \begin{cases} \dfrac{V(\phi_0 + \delta\phi) - V(\phi_0)}{\delta\phi} & \delta\phi \neq 0, \\ \nabla V(\phi_0) & \delta\phi = 0. \end{cases}$$

See [MVL78, CI01, MVL03] to compute the matrix exponential $e^{\mathbf{A}}$ (simulation results use scipy's [VGO+20] function expm which is based on the scaling and squaring method from [AMH10]).

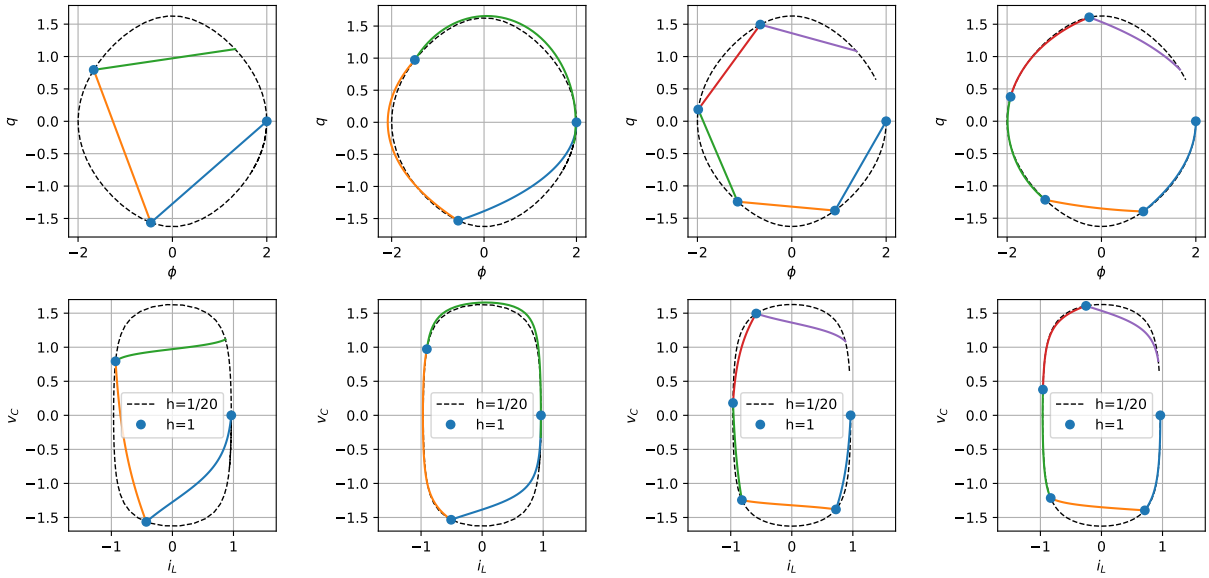
The potential V and its gradient ∇V are shown on Figure 6.3.



(a) $V(\phi)$ **(b)** $\nabla V(\phi)$

Figure 6.3 – (EAVF) Potential function V and its gradient ∇V for $L_0 = L, 2L, 3L, 4L$. Note that although V is not a positive function, the Hamiltonian H remains positive: the quadratic part of H is handled by matrix $\mathbf{Q}$.

Simulations results comparing the AVF method with the Exponential AVF (EAVF) method are shown on Figure 6.4. As expected, the exponential AVF trajectories are closer to the true solution (in dashed black), and exhibit very good accuracy when the nonlinearities changes slowly over the time step. Because of exact integration of the linear dynamic, we note that frequency warping is also improved in the EAVF compared to the AVF method (compare figures 6.4a and 6.4b).



(a) AVF $\omega = \pi$ **(b)** EAVF $\omega = \pi$ **(c)** AVF $\omega = \pi/2$ **(d)** EAVF $\omega = \pi/2$

Figure 6.4 – (Exponential AVF) Comparison of EAVF and AVF methods on Nonlinear LC. The pulsation is set to $\omega \in \{\pi, \pi/2\}$.

6.2.4 Adding external ports

We generalize the exponential AVF method to input-state-output pH-ODEs (definition 1.22). For that purpose, we remark that compared to projection methods of chapter 5 , the crucial element of the proof of theorem 6.1 relies on making a distinction between the exponential trajectory[5] $\boldsymbol{X}$ and its affine Sobolev projection $\boldsymbol{X}_S = \mathcal{P}_S \boldsymbol{X}$ (sharing the same endpoints) such that, thanks to path independence of the gradient theorem, a the proof of equation (6.18) relies on the following identity (see Figure 6.1)

$$H(\boldsymbol{X}_1) - H(\boldsymbol{X}_0) = \left\langle \nabla H(\boldsymbol{X}) \,\middle|\, \dot{\boldsymbol{X}} \right\rangle_{L^2} = \left\langle \nabla H(\boldsymbol{X}_S) \,\middle|\, \dot{\boldsymbol{X}}_S \right\rangle_{L^2} = \left\langle \overline{\overline{\nabla}} H(\boldsymbol{X}) \,\middle|\, \boldsymbol{X}_1 - \boldsymbol{X}_0 \right\rangle_{L^2}.$$

Exponential AVF for input-state-output pH-ODEs

Consider the pH-ODE

$$\begin{cases} \dot{\boldsymbol{x}} = (\mathbf{J} - \mathbf{R})\nabla H(\boldsymbol{x}) + \mathbf{G}\boldsymbol{u}, \\ \boldsymbol{y} = -\mathbf{G}^\mathsf{T} \nabla H(\boldsymbol{x}). \end{cases}, \quad \boldsymbol{x}(t_0) = \mathbf{x}_0, \quad \text{where} \quad \nabla H(\boldsymbol{x}) = \mathbf{Q}\boldsymbol{x} + \nabla V(\boldsymbol{x}). \tag{6.27}$$

Note that, compared to the autonomous case, special care has to be paid for the treatment of inputs and outputs: in the following, we use AVF projection of the input term $\bar{\boldsymbol{u}} = \mathcal{P}\boldsymbol{u}$; dually, we have to use a dual output $\boldsymbol{y} = -\mathbf{G}^\mathsf{T}\overline{\overline{\nabla}} H(\boldsymbol{X})$ to ensure that we still have a passive power-balance (see vanishing cross terms in the proof of theorem 6.2 below). Otherwise the method follows the same construction as in the autonomous case.

5. which brings accuracy by exact integration of the linear part of the vector field

Method 6.1 (Exponential AVF for pH-ODE)**.** Denote $\mathcal{P} = \int_0^1$ the AVF projector and denote $\boldsymbol{X}(\tau)$ an approximation of $\boldsymbol{x}(t_0 + h\tau)$ solution of the system

$$\frac{1}{h}\dot{\boldsymbol{X}}(\tau) = (\mathbf{J} - \mathbf{R})(\mathbf{Q}\boldsymbol{X} + \overline{\overline{\nabla}}V(\boldsymbol{X})) + \mathbf{G}\bar{\boldsymbol{u}}, \qquad \boldsymbol{X}(0) = \mathbf{x}_0, \tag{6.28a}$$

$$\boldsymbol{y} = -\mathbf{G}^{\mathsf{T}}\overline{\overline{\nabla}}H(\boldsymbol{X}), \tag{6.28b}$$

where $\bar{\boldsymbol{u}} = \mathcal{P}\boldsymbol{u}$. The associated time-stepping method is $\Phi_h : \mathbf{x}_0 \mapsto \mathbf{x}_1 = \boldsymbol{X}(1)$.

Let $\mathbf{A} = h(\mathbf{J} - \mathbf{R})\mathbf{Q}$, the exponential trajectory should be a solution of the fixed-point

$$\boldsymbol{X}(\tau) = e^{\tau\mathbf{A}}\mathbf{x}_0 + \int_0^\tau e^{(\tau-\sigma)\mathbf{A}} h \left((\mathbf{J} - \mathbf{R})\overline{\overline{\nabla}}V(\boldsymbol{X}) + \mathbf{G}\bar{\boldsymbol{u}}\right) \mathrm{d}\sigma.$$

Theorem 6.2. *If system (6.27) is discretized using the exponential AVF method (6.28a)-(6.28b). Then, it satisfies the passive average power balance*

$$\frac{H(\mathbf{x}_1) - H(\mathbf{x}_0)}{h} + \langle \boldsymbol{u} \,|\, \boldsymbol{y} \rangle \leq 0.$$

Proof. Take the inner product of (6.28a) with $\overline{\overline{\nabla}}H(\boldsymbol{X}_S)$ on the left, of (6.28b) with $\boldsymbol{u}$ and sum the results to get

$$\begin{aligned}
&\frac{1}{h}\left\langle \overline{\overline{\nabla}}H(\boldsymbol{X}) \,\middle|\, \dot{\boldsymbol{X}} \right\rangle + \langle \boldsymbol{u} \,|\, \boldsymbol{y} \rangle = \left\langle \overline{\overline{\nabla}}H(\boldsymbol{X}) \,\middle|\, (\mathbf{J} - \mathbf{R})(\mathbf{Q}\boldsymbol{X} + \overline{\overline{\nabla}}V(\boldsymbol{X})) + \mathbf{G}\bar{\boldsymbol{u}} \right\rangle - \left\langle \boldsymbol{u} \,\middle|\, \mathbf{G}^{\mathsf{T}}\overline{\overline{\nabla}}H(\boldsymbol{X}) \right\rangle \\
&\overset{a}{\iff} \frac{1}{h}\left\langle \overline{\overline{\nabla}}H(\boldsymbol{X}) \,\middle|\, \mathbf{x}_1 - \mathbf{x}_0 \right\rangle + \langle \boldsymbol{u} \,|\, \boldsymbol{y} \rangle = \frac{1}{2}\left\| e^{\mathbf{A}}\left(\mathbf{x}_0 + \overline{\overline{\nabla}}_{\mathbf{Q}}V(\boldsymbol{X})\right)\right\|_{\mathbf{Q}}^2 - \frac{1}{2}\left\|\mathbf{x}_0 + \overline{\overline{\nabla}}_{\mathbf{Q}}V(\boldsymbol{X})\right\|_{\mathbf{Q}}^2. \\
&\overset{b}{\implies} \frac{H(\mathbf{x}_1) - H(\mathbf{x}_0)}{h} + \langle \boldsymbol{u} \,|\, \boldsymbol{y} \rangle \leq 0 \quad \text{if} \quad \mathbf{R} \succeq \mathbf{0}.
\end{aligned}$$

The following identities were used to obtain the result:

a) By construction (see def.6.2), we have $\mathcal{P}\dot{\boldsymbol{X}} = \dot{\boldsymbol{X}}_S = \mathbf{x}_1 - \mathbf{x}_0$, $\mathcal{P}^2 = \mathcal{P}$ and $\mathcal{P}^* = \mathcal{P}$ so that

$$\begin{aligned}
\left\langle \overline{\overline{\nabla}}H(\boldsymbol{X}) \,\middle|\, \dot{\boldsymbol{X}} \right\rangle = \left\langle \mathcal{P}\nabla H(\boldsymbol{X}_S) \,\middle|\, \dot{\boldsymbol{X}} \right\rangle = \left\langle \mathcal{P}^2\nabla H(\boldsymbol{X}_S) \,\middle|\, \dot{\boldsymbol{X}} \right\rangle &= \left\langle \mathcal{P}\nabla H(\boldsymbol{X}_S) \,\middle|\, \mathcal{P}\dot{\boldsymbol{X}} \right\rangle \\
&= \left\langle \overline{\overline{\nabla}}H(\boldsymbol{X}) \,\middle|\, \mathbf{x}_1 - \mathbf{x}_0 \right\rangle.
\end{aligned}$$

Furthermore we use (6.24c) in the proof of theorem 6.1, the main difference compared to the proof of theorem 6.1 is the presence of input-output cross-terms. They vanish thanks to $\overline{\overline{\nabla}}H(\boldsymbol{X}) = \mathcal{P}\nabla H(\boldsymbol{X}_S)$, $\bar{\boldsymbol{u}} = \mathcal{P}\boldsymbol{u}$ and the self-adjoint property of projector $\mathcal{P}$ (as above).

$$\left\langle \overline{\overline{\nabla}}H(\boldsymbol{X}) \,\middle|\, \mathbf{G}\bar{\boldsymbol{u}} \right\rangle - \left\langle \boldsymbol{u} \,\middle|\, \mathbf{G}^{\mathsf{T}}\overline{\overline{\nabla}}H(\boldsymbol{X}) \right\rangle = \left\langle \overline{\overline{\nabla}}H(\boldsymbol{X}) \,\middle|\, \mathbf{G}\mathcal{P}\boldsymbol{u} \right\rangle - \left\langle \mathbf{G}\mathcal{P}\boldsymbol{u} \,\middle|\, \overline{\overline{\nabla}}H(\boldsymbol{X}) \right\rangle = 0.$$

b) we use (6.18) for the left hand side and we use (6.24d) for the right hand side.

□

6.3 High-order energy-preserving exponential integrators

In this section, we propose an extension of the results from section 6.2 to arbitrary projection orders. The price we pay with this approach, is that the linear dynamic is no longer integrated exactly: a perturbation term is introduced by the projector to satisfy the power balance.

For simplicity, we consider the autonomous Hamiltonian IVP

$$\dot{\boldsymbol{x}} = \mathbf{J}\nabla H(\boldsymbol{x}), \qquad \boldsymbol{x}(t_0) = \mathbf{x}_0,$$

with matrix $\mathbf{J}$ skew symmetric. Choose a matrix $\mathbf{A}$ (usually $\mathbf{A} = \mathbf{J}\nabla^2 H(\mathbf{x}_0)$), decompose the vector field into a linear part and a *deflated vector field* as

$$\dot{\boldsymbol{x}} = \mathbf{A}\boldsymbol{x} + \left(\mathbf{J}\nabla H(\boldsymbol{x}) - \mathbf{A}\boldsymbol{x}\right),$$

and introduce the differential operator $\mathcal{L}\boldsymbol{x} = \dot{\boldsymbol{x}} - \mathbf{A}\boldsymbol{x}$ to define the equivalent IVP

$$\mathcal{L}\boldsymbol{x} = \left(\mathbf{J}\nabla H(\boldsymbol{x}) - \mathbf{A}\boldsymbol{x}\right), \qquad \boldsymbol{x}(t_0) = \mathbf{x}_0. \tag{6.29}$$

We define the following discretization scheme.

Definition 6.3 (Exponential Projection Method (EPM))**.** Let $\Omega = (t_0, t_0 + h)$. Let $\mathcal{P}$ be a projector in $L^2(\Omega)$ reproducing constant functions and satisfying $\mathcal{P}\mathbf{J} = \mathbf{J}\mathcal{P}^*$. Denote $\boldsymbol{X}, \boldsymbol{X}_S \in H^1(\Omega)$ the approximations of the IVP (6.29), that solve the implicit equations

$$\mathcal{L}\boldsymbol{X} = \mathcal{P}\left(\mathbf{J}\nabla H(\boldsymbol{X}_S) - \mathbf{A}\boldsymbol{X}\right) \quad \text{in } \Omega, \quad \text{and} \quad \boldsymbol{X}(t_0) = \mathbf{x}_0, \tag{6.30a}$$

$$\text{where} \quad \dot{\boldsymbol{X}}_S := \mathcal{P}\dot{\boldsymbol{X}}, \quad \text{in } \Omega, \quad \text{and} \quad \boldsymbol{X}_S(t_0) = \boldsymbol{X}(t_0). \tag{6.30b}$$

We call $\boldsymbol{X}$ the *exponential trajectory*, $\boldsymbol{X}_S$ its *Sobolev projection* [a] and $\Phi_h : \mathbf{x}_0 \mapsto \mathbf{x}_1 = \boldsymbol{X}(t_0+h)$ the *time-stepping function* of the *exponential projection method* (EPM).

a. See definition 6.2.

Then, the following results holds.

Proposition 6.1 (Energy preservation)**.** *EPMs are energy-preserving.*

Proof. Rewrite equation (6.30a) to express the derivative $\dot{\boldsymbol{X}}$

$$\dot{\boldsymbol{X}} - \mathbf{A}\boldsymbol{X} = \mathcal{P}\left(\mathbf{J}\nabla H(\boldsymbol{X}_S) - \mathbf{A}\boldsymbol{X}\right) \quad \Longleftrightarrow \quad \dot{\boldsymbol{X}} = \mathcal{P}\mathbf{J}\nabla H(\boldsymbol{X}_S) + (\mathcal{I} - \mathcal{P})\mathbf{A}\boldsymbol{X}. \tag{6.31}$$

From proposition 6.3 below, we have $\mathbf{x}_1 = \boldsymbol{X}(t_0 + h) = \boldsymbol{X}_S(t_0 + h)$. Express the power-balance

$$\begin{aligned} H(\mathbf{x}_1) - H(\mathbf{x}_0) &\overset{a}{=} \left\langle \nabla H(\boldsymbol{X}_S) \,\middle|\, \dot{\boldsymbol{X}}_S \right\rangle \overset{b}{=} \left\langle \nabla H(\boldsymbol{X}_S) \,\middle|\, \mathcal{P}\dot{\boldsymbol{X}} \right\rangle \\ &\overset{c}{=} \left\langle \nabla H(\boldsymbol{X}_S) \,\middle|\, \mathcal{P}^2\mathbf{J}\nabla H(\boldsymbol{X}_S) + \mathcal{P}(\mathcal{I} - \mathcal{P})\mathbf{A}\boldsymbol{X} \right\rangle \\ &\overset{d}{=} \left\langle \nabla H(\boldsymbol{X}_S) \,\middle|\, \mathcal{P}\mathbf{J}\mathcal{P}^* | \nabla H(\boldsymbol{X}_S) \right\rangle \overset{e}{=} 0. \end{aligned}$$

This result stems from (a) the gradient theorem, (b) equation(6.30b), (c) equation (6.31), (d) identities $\mathcal{P}^2\mathbf{J} = \mathcal{P}\mathbf{J}\mathcal{P}^*$ and $\mathcal{P}(\mathcal{I} - \mathcal{P}) = 0$, (e) skew-adjointness of $\mathcal{P}\mathbf{J}\mathcal{P}^*$. □

Proposition 6.2 (passivity preservation)**.** *EPMs are passivity-preserving.*

Proof. Replacing the skew symmetric matrix $\mathbf{J}$ by $\mathbf{J}-\mathbf{R}$ with $\mathbf{R}=\mathbf{R}^\mathsf{T}\succeq 0$ in the proof of proposition 6.1 yields $H(\mathbf{x}_1)-H(\mathbf{x}_0)=-\left\langle \nabla H(\boldsymbol{X}_S)\,\middle|\,\mathcal{P}\mathbf{R}\mathcal{P}^*|\nabla H(\boldsymbol{X}_S)\right\rangle \leq 0$. □

Proposition 6.3. *The exponential trajectory $\boldsymbol{X}$ and its Sobolev projection $\boldsymbol{X}_S$ in definition 6.3 share the same endpoint $\mathbf{x}_1=\boldsymbol{X}(t_0+h)=\boldsymbol{X}_S(t_0+h)$.*

Proof. By definition 6.3, $\mathbf{x}_1=:\boldsymbol{X}(t_0+h)$. Let $\mathcal{P}_0=\frac{1}{h}\int_\Omega$ denote the averaging projector from $L^2(\Omega)$ to the space of constant functions. Since $\mathcal{P}$ reproduces constants, we have $\mathcal{P}_0\mathcal{P}=\mathcal{P}_0$. Then,

$$\boldsymbol{X}(t_0+h)=\mathbf{x}_0+\int_\Omega \dot{\boldsymbol{X}}(t)\,\mathrm{d}t=\mathbf{x}_0+h\mathcal{P}_0\dot{\boldsymbol{X}}$$

and

$$\boldsymbol{X}_S(t_0+h)=\mathbf{x}_0+\int_\Omega \mathcal{P}\dot{\boldsymbol{X}}(t)\,\mathrm{d}t=\mathbf{x}_0+h\mathcal{P}_0\mathcal{P}\dot{\boldsymbol{X}}=\mathbf{x}_0+h\mathcal{P}_0\boldsymbol{X}.$$

It follows that $\mathbf{x}_1=:\boldsymbol{X}(t_0+h)=\boldsymbol{X}_S(t_0+h)$. □

Remarks We make the following observations regarding EPMs

a) As in chapter 5, we only require the projector to reproduce constants and satisfy the commutation condition $\mathcal{P}\mathbf{J}=\mathbf{J}\mathcal{P}^*$.

b) A sufficient condition is fulfilled when $\mathcal{P}$ is scalar (commuting with matrices) and self-adjoint ($\mathcal{P}=\mathcal{P}^*$) projector. But using adjoint pairs of non-scalar projectors is an interesting option for partitionable equations that gives more freedom over the choice of projection space(s).

c) As in the EAVF method, using the Sobolev projected trajectory $\boldsymbol{X}_S$ to evaluate the nonlinearity is a key aspect of the method [6]. Without this double projection, we would have

$$\dot{\boldsymbol{X}}=\mathbf{J}\nabla H(\boldsymbol{X})+(\mathcal{I}-\mathcal{P})\mathbf{A}\boldsymbol{X}$$

and evaluating the power balance would result (in general) in the non-vanishing term

$$\left\langle \nabla H(\boldsymbol{X})\,\middle|\,\dot{\boldsymbol{X}}\right\rangle=\left\langle \nabla H(\boldsymbol{X})\,\middle|\,\mathcal{P}\mathbf{J}\nabla H(\boldsymbol{X})+(\mathcal{I}-\mathcal{P})\mathbf{A}\boldsymbol{X}\right\rangle=\left\langle \nabla H(\boldsymbol{X})\,\middle|\,(\mathcal{I}-\mathcal{P})\mathbf{A}\boldsymbol{X}\right\rangle\neq 0.$$

d) A drawback of the proposed approach, compared to the EAVF method, is that the linear dynamic is no longer integrated exactly: the projection induces the perturbation term $(\mathcal{I}-\mathcal{P})\mathbf{A}\boldsymbol{X}$ in equation 6.31). The proof is simple (and closer to the approach of chapter 5), but we lose in linear accuracy.

e) Adding H^k regularity has been left for future research.

Note that when we approximate the linear vector field differently from its non-linear part, i.e. if we dissociate the flow space from the projection space (using exponential integration), we have to be more careful (than we had to in chapter 5) to ensure that energy is preserved.

6. A perspective of the proposed approach is to look more closely at the properties of the equivalence class of trajectories that share the same projected vector field: i.e. $\mathcal{P}\dot{\boldsymbol{X}}=\mathcal{P}\dot{\boldsymbol{X}}_S$.

Conclusion

We found after bibliographical research that energy-preserving exponential AVF methods had already been proposed in [SL19] (but restricted to commuting matrices $\mathbf{J}, \mathbf{Q}$) and also in [WW18]. For this reason, we chose not to publish our derivation of the exponential AVF method.

However, since the genesis of the method and the structure of the proof are different, we hope that our presentation, specially in the context of pH-ODE and continuous-time projection, brings a complementary viewpoint which paves the way towards different approximation strategies.

Using the tools and methodology from chapter 5, we were able to generalize this result to higher projection orders for an arbitrary choice of basis. We call this approach (energy/passivity preserving) Exponential Projection Methods (EPM). The proof is simpler than the proof of the EAVF method: it avoids (sometimes tedious) manipulation of convolutions and identities involving matrix exponentials, however, the price to pay is that the linear dynamics is no longer integrated exactly (it is perturbed by a projection term).

The results in this chapter have been obtained late in the redaction of this manuscript. For this reason, analysis of order conditions, existence and uniqueness of solutions and detailed simulations are not included and are left for future research. To this end, a theory of (stiff) order conditions for exponential integrators, using exponential B-series, can be found in the reference [LO13] see also [BOS05, But10]. Finding an alternative strategy to generalise the approach to higher projection orders while exactly integrating the linear dynamic is also left for further research.

Part III

Applications

Chapter 7

Passive Operational Amplifier models

Contents

7.1 A minimal passive model of the operational amplifier 175
7.1.1 Introduction 175
7.1.2 Operational Amplifier Model 176
7.1.3 Case study 179
7.1.4 Sallen-Key analog lowpass filter 183
7.2 A passive fully differential amplifier model with infinite gain 190
7.2.1 Ideal Fully Differential Amplifier (FDA) model 190
7.2.2 Continuous parametrisation 192
7.2.3 Explicit formulation using common and differential ports 194
7.3 Towards a grey-box passive model of the OPA 195

Introduction

The Operational Amplifier is widely used in analog audio circuits. This chapter is concerned with its passive power-balanced modelling as a PHS, which, to our knowledge, has not yet been explored. Our motivation arises by examining the two following questions:

- Do not we learn (in high school) that an operational amplifier is an *active device*?
- Why should we consider a pH model rather than the state-of-the-art [1] ?

First, the OPA component does not create energy by itself: it is passive without a power supply. Thus, our first motivation is to model the passive component separately from the power supply, introducing *explicit power supply ports* [2].

Second, to understand the interest of such a modelling, we perform a simple passivity test. Consider the circuit in figure 7.1, involving a resistor, an OPA and a capacitor. This capacitor replaces the traditional power supply of the OPA: this circuit is thus fully passive. Indeed, the capacitor is initialised with zero charge, all ports except the positive supply and the output are grounded [3]. Then, according to charge conservation, the sum of all currents should be zero, so

1. such as the macro models in SPICE-like simulation software that have been used for decades by engineers.
2. Note that qualifying the OPA as an "active device" and hiding the power supply ports is common practice. But this is a huge source of confusion for many students. Examination of the second question will show that this confusion is not limited to vocabulary but also affects modelling.
3. The output resistor is meant as a short circuit, but LTSPICE solvers requires a non zero resistance.

that no current can flow in the capacitor and its output voltage should stay equal to zero.

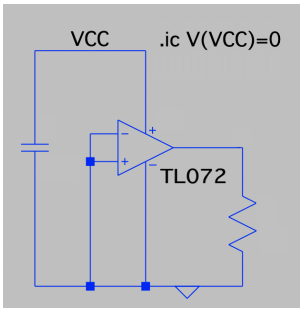
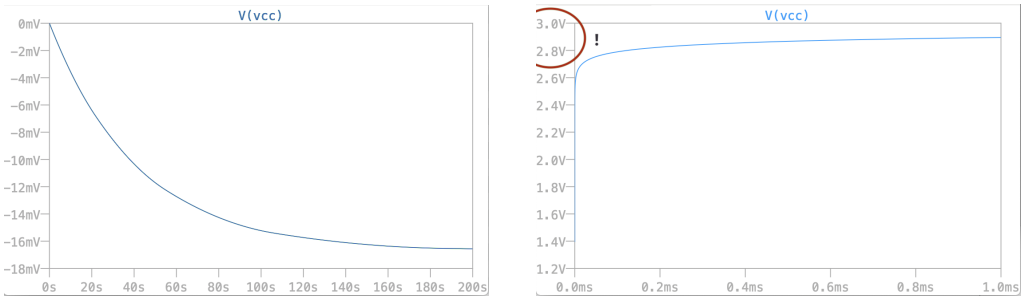


Figure 7.1 – (Passivity test) operational amplifier circuit.

Simulations in LTSPICE (see figure 7.2) yield ill results that do not pass the passivity test for various OPAs models used in audio amplifiers (TL1366 and TL072). Indeed, both conservation of charge and passivity are violated since the OPA charges the capacitor to a significant non-zero value (dependent on the OPA macro model). The reason lies in the common practice of using controlled current and voltage sources in behavioural macro-modelling [4] of components.



(a) LT1366 **(b)** TL072

Figure 7.2 – (Passivity test) Simulation result in LTSPICE for two different OPA macro models. The OPA is charging the capacitor, violating both passivity and conservation of charge.

Our passivity test may seem far-fetched for real-life applications as OPA, transistor and tube amplifiers are usually designed and biased to avoid non-ideal behaviour. But musicians are known for pushing devices outside of their intended use (e.g. overdrive). It is not unusual for guitarists and effect pedal designers to use what is called *voltage sag* for creative purposes [5].

All these practical elements strongly motivate our strategy to build passive OPA models, including in overdriven and under-powered configuration. Section 7.1 presents a first idealized (conservative, memoryless, saturating) model with an illustrative application (this section repeats the original content published in [MH19]). Section 7.2 considers a limit-case: a fully-differential amplifier with infinite gain. Section 7.3 paves the way towards a grey-box model incorporating non-ideal behaviours (limited bandwidth, and slew-rate, dissipation...)

4. As a counter example, the Ebers–Moll transistor model is often depicted using diodes and voltage-controlled current sources to describe PN coupling. Despite this, we proved in example 1.10 p.32 that this model is passive. However establishing such proofs can be difficult and has to be performed for each component. By contrast, the pH modelling strategy is to exclusively rely on provably passive formulations.

5. The power supply voltage is voluntarily (and even dynamically) lowered to push a circuit outside of its ideal operating point, resulting in all kinds of unexpected behaviours (dead zone, self-oscilations, etc).

7.1 A minimal passive model of the operational amplifier

This section repeats the original content published in [MH19].

Abstract

This papers stems from the fact that, whereas there are passive models of transistors and tubes, a minimal passive model of the operational amplifier does not seem to exist. A new behavioural model is presented that is memoryless, fully described by its interaction ports, with a minimal number of equations, for which a passive power balance can be defined. The proposed model handles saturation, asymmetric power supply, and can be used with non-ideal voltage references. To illustrate the model in audio applications, the non-inverting voltage amplifier and a saturating Sallen-Key lowpass filter are considered.

7.1.1 Introduction

Operational Amplifier (OPA) models can be roughly categorized into a) Controlled Source (CS) models, b) white box macro models and c) Nullor models .

In CS models (see [CDK87]), the power supplies are lumped within the OPA and controlled sources can provide an infinite amount of power. It has the advantage of being simple and hides most of the internal complexity. This is the method of choice used by students to study the functional behaviour of OPA circuits. The main drawback comes from the absence of external supply ports. This results in non passive models, and forbids simulations with non-ideal voltage sources (e.g. in low-budget guitar stomboxes).

White box macro models (see references [BPCS74] [CB01] [AB90]) use dozens of transistors to accurately reproduce the inner structure and non-ideal characteristics of particular devices. While this is appropriate for offline simulation and circuit design, the main drawback of this approach comes from the high number of (implicit) nonlinear equations which makes it often unsuitable for real-time simulation.

Nullors (see references [Car64] [Tel66] [OU80] [Mar65]), are singular two-port elements where the input flow and effort variables are both zero: $e_1 = f_1 = 0$, while the output flow and effort variables e_2, f_2 are unconstrained. One drawback is the lack of flow / effort duality. In addition, similar to CS, Nullors have no explicit power supply ports and thus are not passive devices, inheriting the same drawbacks mentioned above.

For audio applications, dedicated Wave Digital Filters (WDF) models of the OPA for specific circuit topologies have been proposed in [PdPV12], more recently, using Modified Nodal Analysis to WDF adaptors, both Nullor and CS general purpose models of the OPA and OTA have been proposed in [WDR+16] [BW17] and Sallen-key filters have been modelled with WDF in [VBS17].

We propose a passive, quasi-ideal, black-box, behavioural model of the OPA, simple enough for realtime simulation, with explicit power supply and modelling nonlinear saturation. In particular, a by-product of this research is to have a model compatible with the port-Hamiltonian formalism [VdS06].

The paper is structured as follows. First a general purpose passive model of the OPA is proposed in subsection 7.1.2, then it is illustrated by treating the non-inverting voltage amplifier circuit in subsection 7.1.3, finally a detailed study and simulation of a saturating Sallen-Key lowpass filter is presented in subsection 7.1.4.

7.1.2 Operational Amplifier Model

The objective of this paper is to find the simplest class of Operational Amplifier models satisfying the following requirements:

a) **Memoryless**: infinite bandwidth, infinite slew rate,

b) **Passivity**: the power dissipated by the OPA is non-negative (i.e. hidden sources of energy are forbidden),

c) **Quasi-ideal**: infinite input impedance, zero output impedance, infinite common-mode rejection ratio,

d) **Finite output voltage range** and saturation: explicit non-constant power-supply ports,

e) **Minimal**: behavioural model with a minimum number of equations (i.e. not a white box model containing dozen of transistors).

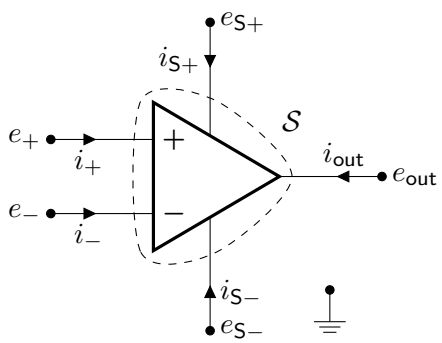


Figure 7.3 – Circuit diagram of an Operational Amplifier (OPA) with currents drawn in receiver convention. The gaussian surface $\mathcal{S}$ enclosing the component is shown in dashed line.

Notations

The OPA shown on figure 7.3 is modelled as a 5-port device with node voltages being measured relatively to the ground, node currents directed toward the element using the receiver convention and pins labelled $P = \{+, -, \mathsf{S}+, \mathsf{S}-, \mathsf{out}\}$. In this paper, we assume that the ports of the OPA can be partitioned into a voltage-driven set T, and a current-controlled co-set $\overline{T}$

$$T := \{+, -, \mathsf{S}+, \mathsf{S}-\}, \qquad \overline{T} := \{\mathsf{out}\}, \qquad T \cup \overline{T} = P. \tag{7.1}$$

The respective inputs and outputs are collected into the vectors

$$\mathbf{u} := [\mathbf{e}_T, \mathbf{i}_{\overline{T}}]^\mathsf{T} = [e_+, e_-, e_{\mathsf{S}+}, e_{\mathsf{S}-}, i_{\mathsf{out}}]^\mathsf{T}, \tag{7.2a}$$

$$\mathbf{y} := [\mathbf{i}_T, \mathbf{e}_{\overline{T}}]^\mathsf{T} = [i_+, i_-, i_{\mathsf{S}+}, i_{\mathsf{S}-}, e_{\mathsf{out}}]^\mathsf{T}, \tag{7.2b}$$

Finally, the common supply, the differential supply and the differential input voltages are respectively defined by

$$V_{\mathsf{cm}} = \frac{e_{\mathsf{S}+} + e_{\mathsf{S}-}}{2}, \qquad V_{\mathsf{dm}} = \frac{e_{\mathsf{S}+} - e_{\mathsf{S}-}}{2}, \qquad \epsilon = e_+ - e_-. \tag{7.3}$$

Constitutive equations

Since there are 5 ports with dual flow and efforts variables, 5 independent equations are required to specify the device:

1-2) **Non-energetic input ports**: the current entering the pins $\{+,-\}$ is zero (infinite input impedance)

$$i_+ = i_- = 0, \tag{7.4}$$

3) **Conservation of charge**: Kirchoff Current Law applied over the gaussian surface [6] $\mathcal{S}$ enclosing the AOP implies that the sum of all currents is zero

$$\sum_{\ell \in \mathcal{P}} i_\ell = 0, \tag{7.5}$$

4) **Passivity**: the power absorbed by the OPA is greater or equal to zero

$$P_{\mathsf{diss}} = \mathbf{y}^\mathsf{T}\mathbf{u} = \sum_{\ell \in P} e_\ell \cdot i_\ell \geq 0, \tag{7.6}$$

5) **Differential gain and saturation**: the tensions are tied by a continuous relation

$$e_{\mathsf{out}} = f(e_+, e_-, e_{\mathsf{S}+}, e_{\mathsf{S}-}), \text{ with } \begin{cases} \dfrac{\partial f}{\partial \epsilon} \geq 0, & \text{monotonicity} \\ \max\left(\dfrac{\partial f}{\partial \epsilon}\right) = K, & \text{differential gain} \\ \max(f) = e_{\mathsf{S}+},\ \epsilon \to +\infty & \text{positive saturation} \\ \min(f) = e_{\mathsf{S}-},\ \epsilon \to -\infty & \text{negative saturation} \end{cases}. \tag{7.7}$$

This gives 4 equalities and 1 inequality

$$i_+ = 0 \tag{7.8a}$$
$$i_- = 0 \tag{7.8b}$$
$$i_{\mathsf{S}+} + i_{\mathsf{S}-} + i_{\mathsf{out}} = 0 \tag{7.8c}$$
$$P_{\mathsf{diss}} = i_{\mathsf{S}+} \cdot e_{\mathsf{S}+} + i_{\mathsf{S}-} \cdot e_{\mathsf{S}-} + i_{\mathsf{out}} \cdot e_{\mathsf{out}} \geq 0 \tag{7.8d}$$
$$f(e_{\mathsf{S}+}, e_{\mathsf{S}-}, e_+, e_-) - e_{\mathsf{out}} = 0 \tag{7.8e}$$

Since there is an inequality and the relation f is not specified yet, there is an infinite class of models satisfying these equations. A particular instance is chosen as follows.

Toward a unique model

Substituting (7.3) into the passivity equation (7.8d), using the conservation of charge (7.8c) and simplifying by i_{out} gives the constraint [7]

$$V_{\mathsf{cm}} + V_{\mathsf{dm}}\left(\frac{i_{\mathsf{S}+} - i_{\mathsf{S}-}}{i_{\mathsf{S}+} + i_{\mathsf{S}-}}\right) = e_{\mathsf{out}} - \frac{P_{\mathsf{diss}}}{i_{\mathsf{out}}}, \quad (i_{\mathsf{out}} \neq 0) \tag{7.9}$$

which imposes a lot of structure on the form of the output function. In order to specify a unique model, the following choices are made.

6. The Gaussian surface $\mathcal{S}$ is shown on figure 7.3. For more details see [CDK87].
7. see appendix D.9.1 for a detailed proof.

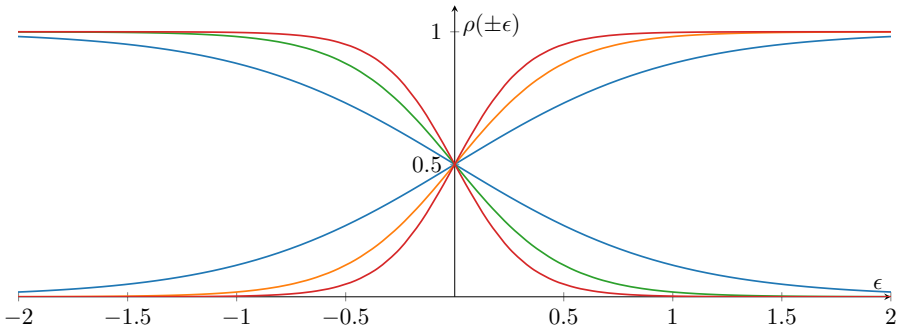


Figure 7.4 – The adimensioned modulation factor $\rho(\pm\epsilon)$, for $K/V_{\mathsf{dm}} = 1, 2, 3$

Push–Pull current splitting First, motivated by the typical structure of an OPA, composed of a differential pair of transistors, gain stages and a push-pull output (see [SS98] p.707), the adimensioned modulation factor [8]

$$\rho(\epsilon) := -\frac{i_{\mathsf{S}+}}{i_{\mathsf{out}}} = \frac{\exp(x)}{\exp(x) + \exp(-x)}, \quad x = \frac{K\epsilon}{V_{\mathsf{dm}}}, \tag{7.10}$$

is introduced and shown on figure 7.4. According to the conservation of charge (7.8c), this leads to the symmetrical current splitting

$$i_{\mathsf{S}+} = -\rho(\epsilon) i_{\mathsf{out}}, \qquad i_{\mathsf{S}-} = -\rho(-\epsilon) i_{\mathsf{out}}. \tag{7.11}$$

The conservative OPA choice Second, among all passive OPA models, the conservative ones are chosen, neglecting internal dissipation:

$$P_{\mathsf{diss}} = 0. \tag{7.12}$$

The power supply ports provide the amount of power necessary to balance the power consumed at the output port. This is an instance of a nonlinear nonenergic n-port [WC77].

Final model Substituting (7.11) and (7.12) into (7.9) uniquely defines the output function (a similar result was also derived in [Mac12a])

$$e_{\mathsf{out}} = V_{\mathsf{cm}} + V_{\mathsf{dm}} \tanh\left(\frac{K\epsilon}{V_{\mathsf{dm}}}\right). \tag{7.13}$$

Expressed as a function of $e_{\mathsf{S}+}, e_{\mathsf{S}-}$ this gives

$$e_{\mathsf{out}} = \rho(+\epsilon) e_{\mathsf{S}+} \; + \; \rho(-\epsilon) e_{\mathsf{S}-}. \tag{7.14}$$

Finally gathering equations (7.4) (7.11) (7.14) in matrix form reveals the modulated hybrid Dirac structure [9] of the conservative OPA model given by the skew-symmetric matrix $\mathbf{J}(\mathbf{u})$:

8. This choices is reminiscent of a BJT push-pull. Different choices for the function ρ can be made to adapt to other transistors types, for examples MOSFETs as long as it defines a complimentary splitting function compatible with charge conservation (7.8c) (i.e. $\rho(\epsilon) + \rho(-\epsilon) = 1$) and saturation constraints (7.7).

9. Please refer to the references [Cou90] [VdS17] [VdS06] for more details on Dirac structures and to [CDK87] for hybrid parameters.

$$\underbrace{\begin{bmatrix} i_+ \\ i_- \\ i_{\mathsf{S}+} \\ i_{\mathsf{S}-} \\ e_{\mathsf{out}} \end{bmatrix}}_{\mathbf{y}} = \underbrace{\left[\begin{array}{cc|cc|c} \cdot & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & 0 \\ \hline \cdot & \cdot & \cdot & \cdot & -\rho(+\epsilon) \\ \cdot & \cdot & \cdot & \cdot & -\rho(-\epsilon) \\ \hline 0 & 0 & \rho(\epsilon) & \rho(-\epsilon) & \cdot \end{array}\right]}_{\mathbf{J}(\mathbf{u})} \underbrace{\begin{bmatrix} e_+ \\ e_- \\ e_{\mathsf{S}+} \\ e_{\mathsf{S}-} \\ i_{\mathsf{out}} \end{bmatrix}}_{\mathbf{u}}. \tag{7.15}$$

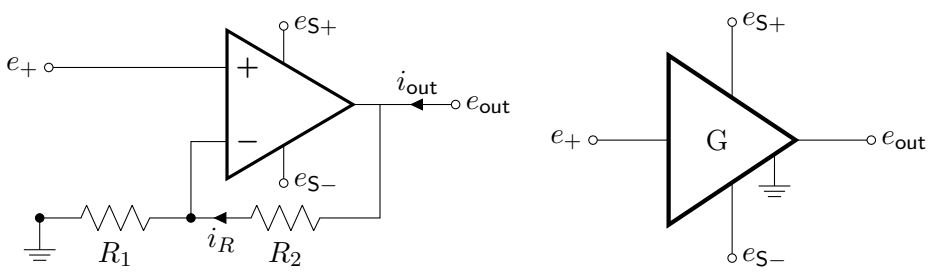
The singularity of the structure matrix $\mathbf{J}$ encodes the conservation of the so-called Casimir invariants $i_+ = i_- = 0$, in addition to the conservative power-balance

$$P_{\mathsf{diss}} = \mathbf{u}^\mathsf{T}\mathbf{y} = \mathbf{u}^\mathsf{T}\mathbf{J}(\mathbf{u})\mathbf{u} = 0, \quad (\text{because } \mathbf{J} = -\mathbf{J}^\mathsf{T}). \tag{7.16}$$

7.1.3 Case study

To study the behaviour of the proposed model in practical applications, the case of the voltage amplifier is examined. Then as a pedagogical example, the voltage amplifier is driven by a sinusoidal voltage source and asymmetrically powered by a single capacitor to simulate a discharging battery. The voltage amplifier will be used as a building block of the Sallen-Key lowpass filter shown in subsection 7.1.4.

The non-inverting voltage amplifier



(a) non inverting amplifier **(b)** equivalent component

Figure 7.5 – Non-inverting voltage amplifier circuit with explicit alimentation ports.

A non-inverting voltage amplifier (figure 7.5) is achieved by feeding back the output e_{out} to the negative input e_- through a voltage divider

$$\epsilon = e_+ - \frac{e_{\mathsf{out}}}{G}, \qquad G = \frac{R_1 + R_2}{R_1} = 1 + \frac{R_2}{R_1}. \tag{7.17}$$

The instantaneous feedback makes the circuit act as a proportional corrector with high proportional gain K in order to satisfy the constraint $e_{\mathsf{out}} \approx Ge_+$ within the range $e_{\mathsf{out}} \in [e_{\mathsf{S}+}, e_{\mathsf{S}-}]$.

The voltage divider induces an internal current $i_R = e_{\mathsf{out}}/R$, where $R = R_1 + R_2$, and the current splitting (7.11) becomes

$$i_{\mathsf{S}+} = -\rho(\epsilon)(i_{\mathsf{out}} - i_R), \qquad i_{\mathsf{S}-} = -\rho(-\epsilon)(i_{\mathsf{out}} - i_R). \tag{7.18}$$

This results in the following law for the voltage amplifier

$$\begin{bmatrix} i_+ \\ i_{\mathsf{S}+} \\ i_{\mathsf{S}-} \\ e_{\mathsf{out}} \end{bmatrix} = \left[\begin{array}{c|cc|c} \cdot & \cdot & \cdot & \cdot \\ \hline \cdot & g_+(\epsilon) & g_\pm(\epsilon) & -\rho(\epsilon) \\ \cdot & g_\pm(\epsilon) & g_-(\epsilon) & -\rho(-\epsilon) \\ \hline \cdot & \rho(\epsilon) & \rho(-\epsilon) & \cdot \end{array}\right] \begin{bmatrix} e_+ \\ e_{\mathsf{S}+} \\ e_{\mathsf{S}-} \\ i_{\mathsf{out}} \end{bmatrix}. \tag{7.19}$$

with conductances

$$g_+(\epsilon) = \frac{\rho(\epsilon)^2}{R}, \qquad g_-(\epsilon) = \frac{\rho(-\epsilon)^2}{R}, \qquad g_\pm(\epsilon) = \frac{\rho(\epsilon)\rho(-\epsilon)}{R}. \tag{7.20}$$

In the following, it is assumed that $R \to \infty$ such that internal losses are negligible. In particular, this is the case of the classical voltage follower circuit for which $R_2 = 0$, and $R_1 = \infty$.

Implicit constraint The relation (7.19) is still implicitly defined since ϵ depends on both input and output variables e_+ and e_{out}. To avoid apparent difficulties with discontinuous functions, consider the curve

$$\mathcal{F} = \left\{ (u, y) \in \mathbb{R}^2 \quad | \quad F(u, y) = 0 \right\}, \tag{7.21}$$

specified by the function

$$F(u, y) = V_{\mathsf{cm}} + V_{\mathsf{dm}} \tanh\left(\frac{K}{V_{\mathsf{dm}}} \left(u - \frac{y}{G} \right) \right) - y, \tag{7.22}$$

and given e_+, look for e_{out} such that $(e_+, e_{\mathsf{out}}) \in \mathcal{F}$.

Since the output function is monotonous with respect to ϵ and bounded in $[e_{\mathsf{S}-}, e_{\mathsf{S}+}]$, a unique solution exists within that range. A global method such as the bisection method is guaranteed to find it, whereas, since K is typically about 10^6, it is very difficult to use either fixed-point or derivative-based methods because of bad numerical conditioning. Numerical simulations are shown on figure 7.6.

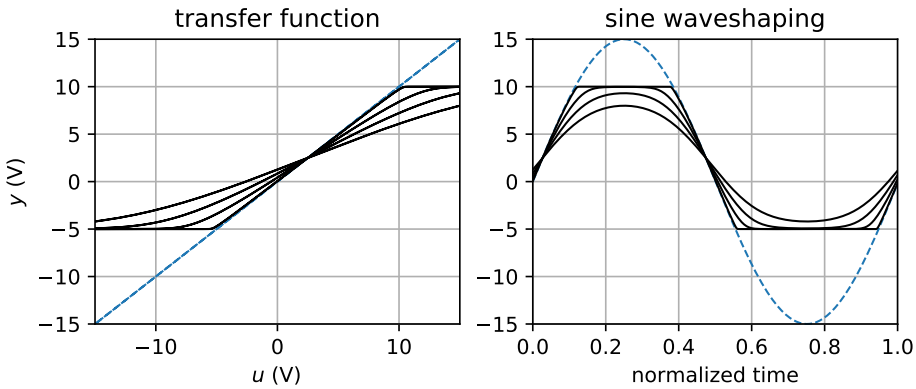


Figure 7.6 – Transfer function of the voltage amplifier for $G = 1$, $K \in \{1, 2, 5, 50\}$, $e_{\mathsf{S}+} = 10\text{V}$, $e_{\mathsf{S}-} = -5\text{V}$. Smaller values than the typical OPA gain $K \approx 10^6$ are used for visualisation purposes.

Explicit representation Taking the limit when $K \to \infty$ gives an explicit representation of $\mathcal{F}$ as the piecewise continuous curve

$$\mathcal{F}_\infty = \lim_{K \to \infty} \mathcal{F} : \begin{cases} y = e_{\mathsf{S}+}, & Gu > y \\ y = e_{\mathsf{S}-}, & Gu < y \\ y \in [e_{\mathsf{S}-}, e_{\mathsf{S}+}], & y = Gu \end{cases}. \tag{7.23}$$

One can see on figure 7.6 that convergence to $\mathcal{F}_\infty$ is very fast even for moderate values of K. This justifies the use of this limit process in following developments.

For $(e_+, e_{\mathsf{out}}) \in \mathcal{F}_\infty$ this gives the explicit form

$$e_{\mathsf{out}} = V_{\mathsf{cm}} + V_{\mathsf{dm}} \operatorname{sat}\left(\frac{Ge_+ - V_{\mathsf{cm}}}{V_{\mathsf{dm}}}\right), \qquad \text{where} \quad \operatorname{sat}(x) = \min(\max(x, -1), 1). \tag{7.24}$$

Alternatively one can represent this function as

$$e_{\mathsf{out}} \;=\; \mu_+(e_+, V_{\mathsf{cm}}, V_{\mathsf{dm}}) \cdot e_{\mathsf{S}+} \quad + \quad \mu_-(e_+, V_{\mathsf{cm}}, V_{\mathsf{dm}}) \cdot e_{\mathsf{S}-} \tag{7.25}$$

where the implicit modulation factor $\rho(\pm\epsilon)$ in (7.19) has been replaced by the explicit one

$$\mu_\pm(e_+, V_{\mathsf{cm}}, V_{\mathsf{dm}}) = \frac{1 \pm \operatorname{sat}(x)}{2}, \qquad x = \frac{Ge_+ - V_{\mathsf{cm}}}{V_{\mathsf{dm}}}. \tag{7.26}$$

A single-rail voltage follower powered by a capacitor

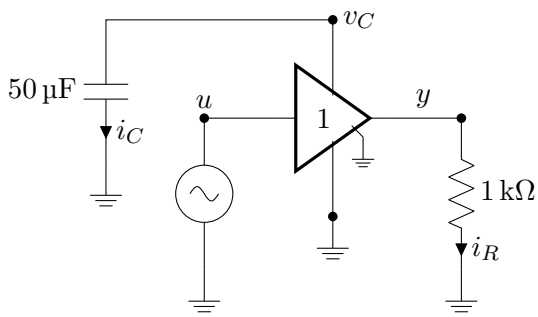


Figure 7.7 – A single-rail voltage amplifier powered by a capacitor.

To illustrate one of the practical interest of having explicit power supply ports, the voltage amplifier is used with the negative supply port grounded, and the positive supply port powered by a capacitor to simulate a discharging battery (figure 7.7).

Using (7.15) with $V_{\mathsf{cm}} = V_{\mathsf{dm}} = q/(2C)$, and $i_{\mathsf{out}} = -y/R$, yields the algebro-differential equations

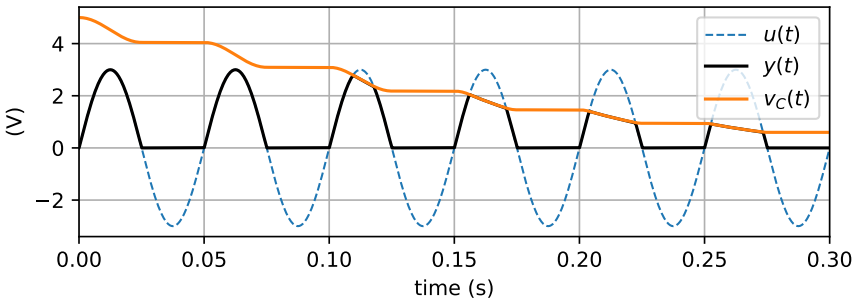
$$\begin{cases} \dot{q} = -\eta(u, q)\dfrac{y}{R}, \\ y = \eta(u, q)\dfrac{q}{C} \end{cases}, \qquad \eta(u, q) = \mu_+\left(u, \frac{q}{2C}, \frac{q}{2C}\right). \tag{7.27}$$

The energy stored in the capacitor is $H(q) = q^2/2C$. Then its differential equation is governed by the monotonous discharge

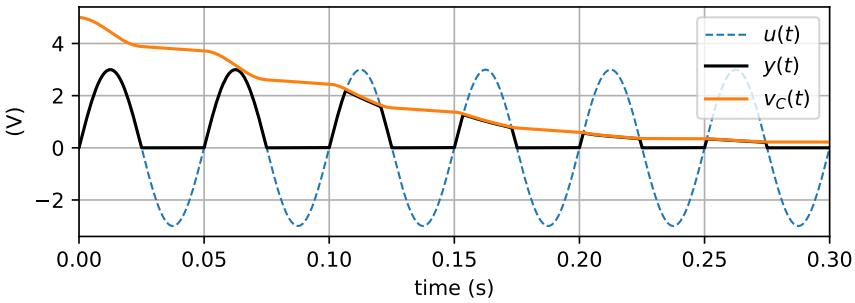
$$\frac{\mathrm{d}}{\mathrm{d}t}H(q) = \frac{\partial H}{\partial q}\frac{\mathrm{d}q}{\mathrm{d}t} = -\frac{q}{C}\eta(q, u)\frac{y}{R} = -\frac{y^2}{R}. \tag{7.28}$$

The circuit acts as a half-wave rectifier with a positive clipping threshold governed by the discharge of the capacitor as shown on figure 7.8.

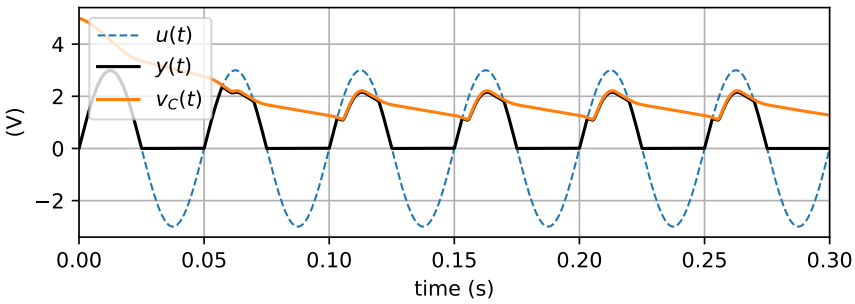
Comparison between models As expected, with the proposed model (fig.7.8 (a)), the capacitor does not discharge during negative saturation (energy-preservation), and has a monotonous discharge otherwise. Comparison with LTspice's universal model (fig.7.9) shows that the two simulations are very close. With the LT1366 (fig.7.8 (b)), the discharge is monotonous and qualitatively similar, but decays faster due to internal dissipation. Finally the LTC6241 (fig.7.8 (c)) exhibits unexpected behaviour: it starts charging back the capacitor once the capacitors drops below a threshold (probably linked non-ideal rail-rail behaviour).



(a) Simulation of the single-rail voltage follower driven by a sinusoid and powered by a capacitor



(b) Same simulation using the LT1366



(c) Same simulation using LTC6241. It is obvious that the LTC6241 is not passive. The capacitor is being charged by the OPA! Instead of discharging monotonously.

Figure 7.8 – Time domain simulation of the capacitor-powered single rail voltage amplifier with $v_C(0) = 5\text{V}$ and $|u| = 3V$.

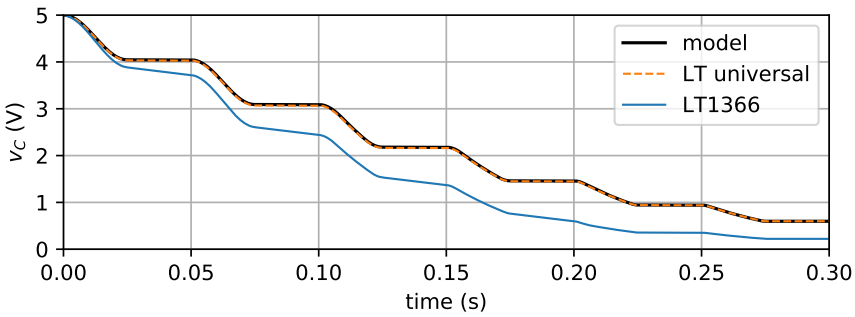


Figure 7.9 – Comparison of discharge rate with LTspice's Universal OPA level.2 and the LT1366 opamp [Dev19].

7.1.4 Sallen-Key analog lowpass filter

The class of Sallen-Key Filters (SKF), introduced in [SK55], is perhaps one of the most common analog filter design topology. It is used for the realization of analog biquadratic filters, for example in parametric equalisers. It is also the basis of the multimode Steiner filter [Ste74], the Korg MS-20 [Sti06] and the Buchla Lowpass-Gate [Pd13].

A Sallen-Key lowpass filter schematic is shown on figure 7.11a. The linear regime and its control parameters are studied in 7.1.4, the circuit is then converted into equations in 7.1.4. Discretization is performed using the Average Vector Field method in 7.1.4, finally simulation results are shown in 7.1.4.

Linear behaviour and control parameters

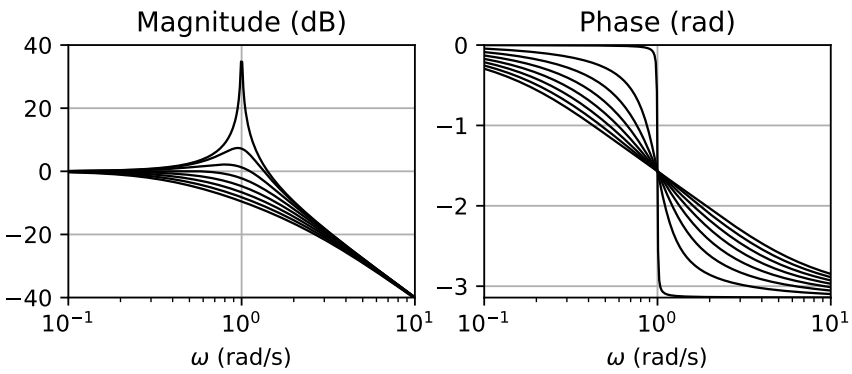


Figure 7.10 – Bode plot of the Sallen-Key filter for $\omega = 1$, $G \in [0, 3]$.

It is recalled that the Laplace transfer function (shown on figure 7.10) of a second order resonant lowpass filters with pulsation ω and quality factor Q is

$$H_{\mathsf{LP}}(s) = \frac{1}{1 + \frac{1}{Q}\left(\frac{s}{\omega}\right) + \left(\frac{s}{\omega}\right)^2}, \tag{7.29}$$

In the linear regime, the Laplace transfer function of the lowpass Sallen-Key filter is

$$H_{\mathsf{SK}}(s) = \mathcal{L}\left\{\frac{y_{\mathsf{SK}}}{v_{\mathsf{IN}}}\right\} = \frac{1}{1 + a_1 s + a_2 s^2}, \tag{7.30}$$

where

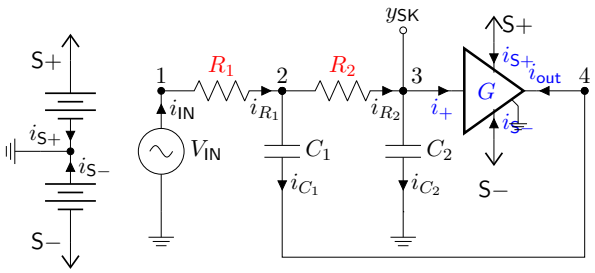
$$a_1 = \left((1-G)R_1C_1 + (R_1+R_2)C_2\right), \tag{7.31a}$$
$$a_2 = C_1C_2R_1R_2. \tag{7.31b}$$

Since there are only two target controls (ω, Q), for 5 design parameters (R_1, R_2, C_1, C_2, G), there are many possible design decisions that are often decided according to electronic constraints.

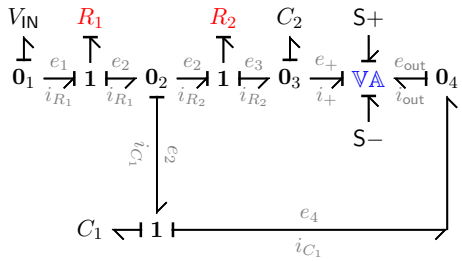
In this paper, the Steiner filter parametrization is used with $R_1 = R_2 = R$, and $C_1 = C_2 = C$ because of its simplicity. The transfer function (7.30) simplifies to

$$H_{\mathsf{SK}}(s) = \frac{1}{1 + (3-G)\left(\frac{s}{\omega}\right) + \left(\frac{s}{\omega}\right)^2}, \tag{7.32}$$

with $\omega = 1/(RC)$, and $Q = 1/(3-G)$. In simulations, capacitances are both set to $C = 4.7\text{nF}$ and the resistors are adjusted to achieve the target cutoff frequencies.



(a) Circuit



(b) Bondgraph

$$
\begin{array}{c}
 \\ i_{IN} \\ i_{C_1} \\ i_{C_2} \\ i_{S+} \\ i_{S-} \\ v_{R_1} \\ v_{R_2} \\ e_+ \\ e_{S+} \\ e_{S-} \\ i_{out}
\end{array}
\begin{array}{c}
\begin{array}{ccccccccccc}
V_{IN} & v_{C_1} & v_{C_2} & S+ & S- & i_{R_1} & i_{R_2} & i_+ & i_{S+} & i_{S-} & e_{out}
\end{array} \\
\begin{pmatrix}
 & & & & & -1 & . & . & . & . & \\
 & & & & & 1 & -1 & . & . & . & \\
 & & & & & . & 1 & -1 & . & . & \\
 & & & & & . & . & . & -1 & . & \\
 & & & & & . & . & . & . & -1 & \\
1 & -1 & . & . & . & & & & & & -1 \\
. & 1 & -1 & . & . & & & & & & 1 \\
. & . & 1 & . & . & & & & & & . \\
. & . & . & 1 & . & & & & & & . \\
. & . & . & . & 1 & & & & & & . \\
 & & & & & 1 & -1 & . & . & . & .
\end{pmatrix}
\end{array}
$$

(c) Skew-symmetric Dirac structure (KCL+KVL)

$$
\begin{pmatrix} i_{C_1} \\ i_{C_2} \\ \hline i_{IN} \\ i_{out} \end{pmatrix}
= \underbrace{\frac{1}{R}\left(\begin{array}{cc|cc} -2 & 1 & 1 & -2 \\ 1 & -1 & 0 & 1 \\ \hline 1 & 0 & -1 & 1 \\ -2 & 1 & 1 & -2 \end{array}\right)}_{Y}
\begin{pmatrix} v_{C_1} \\ v_{C_2} \\ \hline v_{IN} \\ e_{out} \end{pmatrix}
$$

(d) Reduced model (admittance form)

Figure 7.11 – a) The original Sallen-Key lowpass filter circuit, b) its corresponding bondgraph (see references [Pay61] [Bre86] [Bro99b]) with computational causality assignment. c) the skew-symmetric Dirac structure representing Kirchoff conservation laws. d) the reduced dynamical model.

Modelling

To model the Sallen-Key filter, the following systematic approach is used: (See also chapters 1 and 2)

- **Bondgraph**: The circuit on figure 7.11a is first converted to an equivalent bondgraph 7.11b using the rules in [Bre86]. A bond between two ports $\mathbf{A} \longrightarrow \mathbf{B}$ stands for a pair of dual port-variables (e, f). The half-arrow indicates the power sign convention $P = ef \geq 0$. **0** denotes a parallel junction where all bonds share the same voltage, and **1** denotes a serial junction where all bonds share the same current.

- **Causality assignment**: to convert an acausal bidirectional bondgraph to a causal, computable, block-diagram, one needs to partition the flows and efforts into inputs and outputs. The convention uses a vertical stroke $\mathbf{A} \longmapsto \mathbf{B}$ next to ports that are effort-controlled. *Computational causalities* can be assigned graphically by propagating the following rules: voltage sources and capacitors have an effort-out causality, **0** junctions can only have one input effort, while the dual **1** junctions can only have one output effort.

- **Dirac Structure**: given the causality assignment, shown on 7.11b, into inputs and outputs, it is now straightforward to fill the Dirac Structure matrix 7.11c by inspecting circuit 7.11a and expressing Kirchoff's current and voltage laws.

- **Reduced model**: one can reduce the model by solving trivial equalities like $e_+ = v_{C_2}$, $e_{S+} = V_+$, $e_{S-} = V_-$, treating $V_\pm$ as constants and replacing the linear resistive currents (i_{R_1}, i_{R_2}) by their constitutive laws. This results in the reduced admittance model shown on figure 7.11d.

Nonlinear feedback To separate the linear and nonlinear feedback, one can write

$$\hat{e}_{\mathsf{out}}(v) = Gv - \nabla N(v) \tag{7.33}$$

where the nonlinear law is

$$\nabla N(v) := Gv - \hat{e}_{\mathsf{out}}(v) = \min(0, Gv - e_{\mathsf{S}-}) + \max(0, Gv - e_{\mathsf{S}+}). \tag{7.34}$$

and its algebraic potential (figure 7.12) is given by the line integral

$$N(v) := \int_0^v \nabla N(s) \cdot \mathrm{d}s = \frac{\min(0, Gv - e_{\mathsf{S}-})^2}{2G} + \frac{\max(0, Gv - e_{\mathsf{S}+})^2}{2G}. \tag{7.35}$$

This potential will used by the Average Vector Field discretization (an instance of Anti-Derivative Anti-Aliasing).

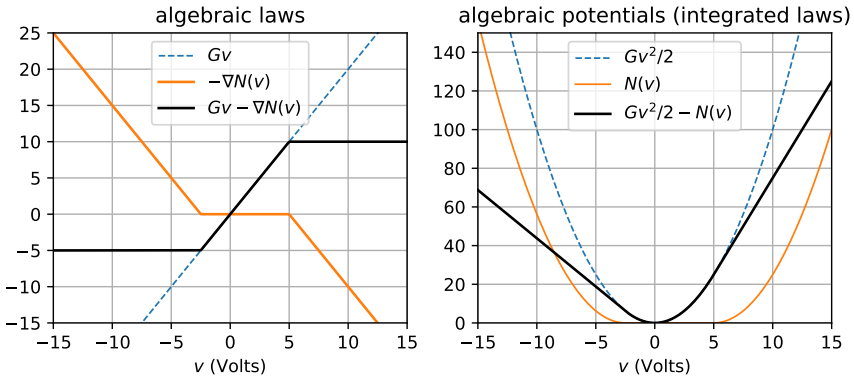


Figure 7.12 – Algebraic feedback laws and their potentials shown for $G = 2$, $e_{\mathsf{S}+} = 10\mathrm{V}$, $e_{\mathsf{S}-} = -5\mathrm{V}$.

State-space model

Finally replacing the flow and effort variables by their constitutive laws, and only considering the input-state-output, one gets

$$\begin{cases} \dot{\mathbf{x}} = \omega \left[\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} - \mathbf{F}\nabla N(\mathbf{C}\mathbf{x})\right] \\ \mathbf{y} = \mathbf{C}\mathbf{x} \end{cases}, \tag{7.36}$$

where $\mathbf{u} = v_{\mathsf{IN}}$, $\mathbf{y} = y_{\mathsf{SK}}$, $\mathbf{x} = [v_{C_1}, v_{C_2}]^{\mathsf{T}}$, $\omega = 1/(RC)$ and

$$\mathbf{A} = \begin{bmatrix} -2 & 1-2G \\ 1 & -1+G \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \qquad \mathbf{C} = \begin{bmatrix} 0 & 1 \end{bmatrix}, \qquad \mathbf{F} = \begin{bmatrix} -2 \\ 1 \end{bmatrix}. \tag{7.37}$$

Using the co-energy variables v_{C_1}, v_{C_2} instead of the energy variables q_{C_1}, q_{C_2} is justified here by the fact that the capacitors are linear and time-invariant, i.e. the co-energy $H^*(v) = Cv^2/2$ equals the energy $H(q) = q^2/(2C)$ for the linear law $v = q/C$.

Discretization using the AVF method

The Average Vector Field (AVF) method is used to discretize (7.36) because of its structure-preserving properties: it preserves the energy (resp. dissipativity) of conservative (resp. dissipative) systems (see [CGM+12]). One can also refer to [Hél11] where it has been shown that the bilinear transform doesn't always guarantee the dissipativity of nonlinear filters (whether time-varying or not). Furthemore, the interrest of generalizing the Average Discrete Gradient to algebraic potentials has been shown in [MH18]. As an important side-effect, the AVF method can also be interpreted as a first-order instance of anti-derivative antialiasing [BEPV17].

The Average Vector Field method Let $\Omega = [t_0, t_0 + h]$ be a time-step, $\boldsymbol{x} \in \mathbb{P}^1(\Omega \to \mathbb{R}^n)$ a locally affine trajectory parametrized by the normalized variable $\tau \in [0, 1]$

$$\boldsymbol{x}(t_0 + h\tau) = \mathbf{x}_0 + \tau(\mathbf{x}_1 - \mathbf{x}_0). \tag{7.38}$$

Introduce the averaging projector $\mathcal{A}$, defined for all functions $f : \mathbb{R}^n \to \mathbb{R}^n$ or operators $f : \mathcal{H} \to \mathcal{H}$, where $\mathcal{H}$ is a functional space from $\Omega \to \mathbb{R}^n$, by

$$(\mathcal{A} f)(\boldsymbol{x}) := \int_0^1 f(\boldsymbol{x}(t_0 + h\tau))\, \mathrm{d}\tau. \tag{7.39}$$

For the time derivative and identity operators, one gets first order finite difference and average

$$\overline{\dot{\boldsymbol{x}}} := \left(\mathcal{A}\,\frac{\mathrm{d}}{\mathrm{d}t}\right)\boldsymbol{x} = \frac{\mathbf{x}_1 - \mathbf{x}_0}{h}, \qquad \bar{\boldsymbol{x}} := (\mathcal{A}\mathcal{I})\boldsymbol{x} = \frac{\mathbf{x}_0 + \mathbf{x}_1}{2}. \tag{7.40}$$

For ∇N, using the gradient theorem, this gives the average discrete gradient

$$\overline{\nabla} N(v_0, v_1) := (\mathcal{A}\nabla N)(v_0 + \tau(v_1 - v_0)) = \begin{cases} \dfrac{N(v_1) - N(v_0)}{v_1 - v_0} & v_0 \neq v_1 \\ \nabla N(v_0) & v_0 = v_1 \end{cases}. \tag{7.41}$$

Computing its derivative with respect to v_1 leads to the discrete pseudo-Hessian

$$\frac{\partial \overline{\nabla} N}{\partial v_1}(v_0, v_1) = \begin{cases} \dfrac{\nabla N(v_1) - \overline{\nabla} N(v_0, v_1)}{v_1 - v_0} & v_0 \neq v_1 \\ \dfrac{1}{2}\nabla^2 N(v_0) & v_0 = v_1 \end{cases}. \tag{7.42}$$

One can refer to [MH18], where the discrete gradient's derivative is also used for numerical simulation. Note that the average discrete gradient of the nonlinearity $\overline{\nabla} N$ is continuously derivable for $v_0 \neq v_1$, while ∇N is not.

Averaged state space system Applying the averaging projector $\mathcal{A}$ to (7.36), leads to the structure-preserving discrete algebraic system

$$\begin{cases} \overline{\dot{\boldsymbol{x}}} = \omega\left[\mathbf{A}\bar{\boldsymbol{x}} + \mathbf{B}\bar{\boldsymbol{u}} - \mathbf{F}\,\overline{\nabla} N(\mathbf{C}\mathbf{x}_0, \mathbf{C}\mathbf{x}_1)\right] \\ \bar{\mathbf{y}} = \mathbf{C}\bar{\boldsymbol{x}} \end{cases}. \tag{7.43}$$

Solving the linear part for $\mathbf{x}_1$ gives the discrete state-space update

$$\mathbf{x}_1 = \mathbf{A}_d\mathbf{x}_0 + \mathbf{B}_d\bar{\boldsymbol{u}} - \mathbf{F}_d\,\overline{\nabla} N(\mathbf{C}\mathbf{x}_0, \mathbf{C}\mathbf{x}_1), \tag{7.44}$$

with the normalised pulsation $\omega_d = h\omega$ and

$$\mathbf{A}_d = \mathbf{D}^{-1}\left(\mathbf{I} + \frac{\omega_d}{2}\mathbf{A}\right), \quad \mathbf{B}_d = \mathbf{D}^{-1}(\omega_d\mathbf{B}), \quad \mathbf{D} = \left(\mathbf{I} - \frac{\omega_d}{2}\mathbf{A}\right), \quad \mathbf{F}_d = \mathbf{D}^{-1}(\omega_d\mathbf{F}). \tag{7.45}$$

Simulation

Simulation results [10] are shown on figures 7.13 and 7.14 and exhibit a very close match with offline simulations performed in LTspice. To solve (7.44), one can either use the simple fixed-point iteration, or Newton's method.

Fixed-point iteration A simple numerical scheme is to look for the fixed-point $\mathbf{x}_1 = \phi(\mathbf{x}_1)$ of the pre-conditioned fixed-point function

$$\phi(\mathbf{x}_1) := \mathbf{A}_d\mathbf{x}_0 + \mathbf{B}_d\bar{\mathbf{u}} - \mathbf{F}_d\,\overline{\nabla}N(\mathbf{C}\mathbf{x}_0, \mathbf{C}\mathbf{x}_1), \tag{7.46}$$

with the fixed-point iteration

$$\mathbf{x}_1^{k+1} = \phi\left(\mathbf{x}_1^k\right), \quad \mathbf{x}_1^0 = \mathbf{x}_0. \tag{7.47}$$

A sufficient convergence condition is detailed in appendix D.9.2.

In practice, thanks to the non linear feedback splitting in (7.33), when the OPA is in the linear regime, $\nabla N = 0$. Then the iteration reduces to an explicit one-step trapezoidal integrator and converges in only one iteration.

Newton iteration To accelerate convergence, one can use Newton's method [Deu11] as follows: define the auxiliary function

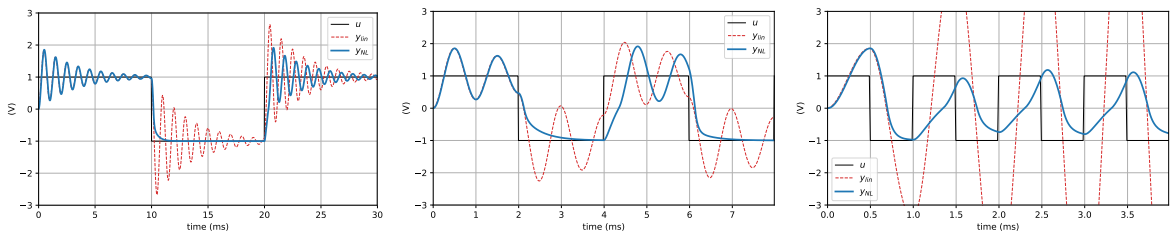
$$\varphi(\mathbf{x}_1) = \mathbf{x}_1 - \phi(\mathbf{x}_1), \tag{7.48}$$

and look for the root $\mathbf{x}_1^*$ such that $\varphi(\mathbf{x}_1^\star) = 0$ with the Newton iteration

$$\mathbf{x}_1^{k+1} = \mathbf{x}_1^k - \left(\varphi'(\mathbf{x}_1^k)\right)^{-1}\varphi(\mathbf{x}_1^k), \quad \mathbf{x}_1^0 = \mathbf{x}_0. \tag{7.49}$$

where the Jacobian of φ is given by

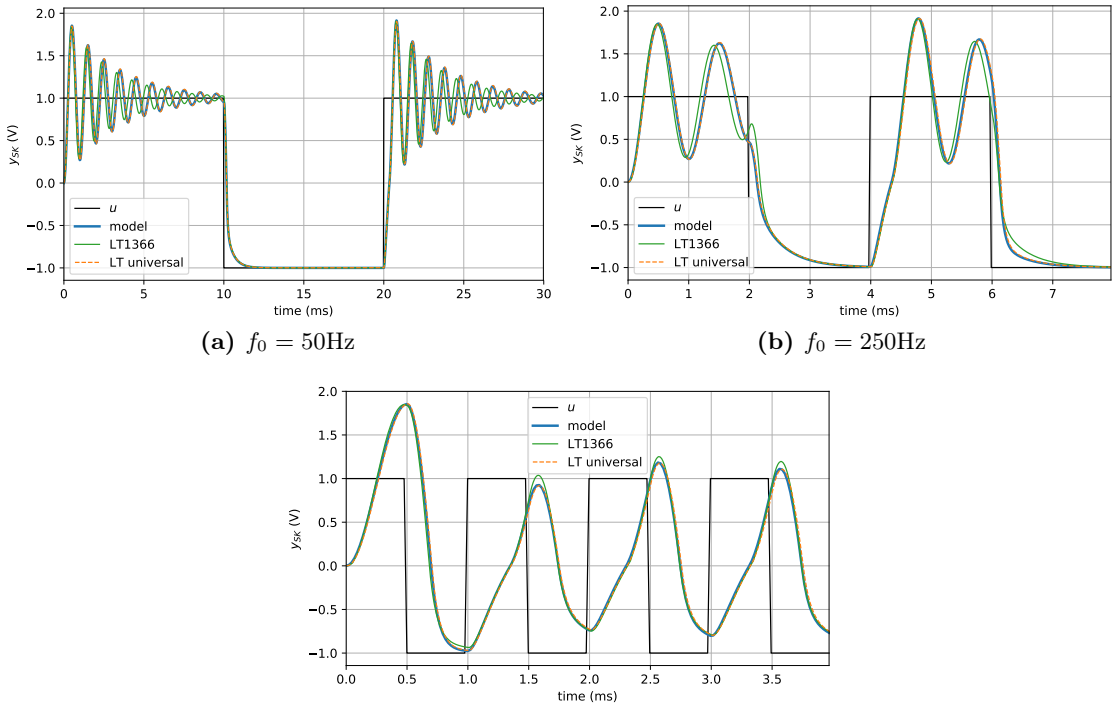
$$\varphi'(\mathbf{x}_1) = \mathbf{I} + \mathbf{F}_d\mathbf{C}\,\frac{\partial\overline{\nabla}N}{\partial v_1}\,(\mathbf{C}\mathbf{x}_0, \mathbf{C}\mathbf{x}_1). \tag{7.50}$$



(a) $f_0 = 50$Hz **(b)** $f_0 = 250$Hz **(c)** $f_0 = 1$kHz

Figure 7.13 – SKF filter response to a square wave input with sampling frequency $f_s = 44.1$kHz, $C = 4.7$nF, cutoff $f_c = 1$kHz ($R = 33.8k\Omega$), $Q = 10$, asymmetric saturation $V_+ = 15$V, $V_- = 0$V and different fundamental frequencies. The non linear SKF response is shown in solid blue, with the linear SKF response in dashed red for reference.

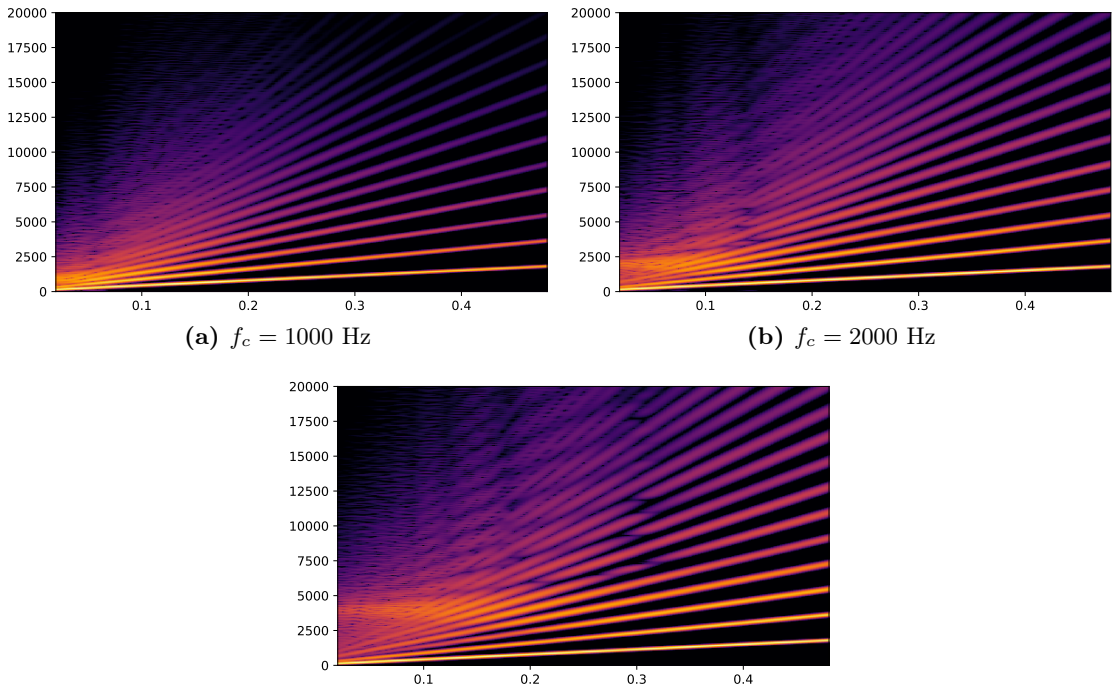
10. Sound examples and LTspice files are available at the accompanying website: https://github.com/remymuller/dafx19-opa.



(a) $f_0 = 50$Hz **(b)** $f_0 = 250$Hz

(c) $f_0 = 1$kHz

Figure 7.14 – Comparison between the proposed model, LTspice's universal OPA level.2 and the LT1366 opamp. The proposed model output is almost indistinguishable from LTspice's universal model, whereas the tuning of the LT1366 is slightly different because of dissipation.



(a) $f_c = 1000$ Hz **(b)** $f_c = 2000$ Hz

(c) $f_c = 4000$ kHz

Figure 7.15 – Spectrogram responses to a sine sweep for $f_c \in \{1000, 2000, 4000\}$ Hz. Intermodulation between the input and the resonance is noticeable.

Conclusions and perspectives

In this paper, a static, passive, black-box model of the operational amplifier with explicit power supply has been examined. It is suitable for the modelling of audio circuits and simple enough for real-time simulation. Furthermore the explicit modelling of external power supply ports allows the use of non-ideal voltage sources.

The choice has been made to ignore internal dissipation to keep the model minimal. However, non-ideal characteristics such as input and output impedance or power supply voltage drop can be achieved by modular composition of the model with other circuit elements. This will be the topic of further research.

The non inverting amplifier is also derived as a dedicated building block. Numerical simulations justify the use of an infinite OPA gain to get an explicit formulation. Having a pre-solved amplifier model also greatly simplifies its use in electronic circuits, avoiding numerical stiffness and high index DAE.

Finally, the amplifier is used for audio simulations to model a saturating Sallen-Key lowpass filter of second order. A reduced state-space model is derived from the circuit schematic, and a struc-ture-preserving discretization is performed using the average vector field method. A comparison with LTspice shows that our results are very close to those of more complex macro models.

The perspectives of this study are a) modelling other non-ideal OPA characteristics such as finite slew-rate and bandwidth, current and voltage offsets, non-zero common-mode input gain... b) studying the behaviour of the model in other typical circuits (oscillator, rectifier, comparator) and c) experimental comparison with specific devices such as the common μA741, or TL072 audio OPAs which are not rail-to-rail opamps.

7.2 A passive fully differential amplifier model with infinite gain

This section deals with the pH modelling of fully-differential operational amplifiers having symmetric power supply, infinite gain, and differential input and output. This component is common in textbooks, but usually, the power supply port is not represented (and passivity not addressed). Moreover, the linear or saturation behaviours are usually modelled separately, on a case-by-case basis. The model proposed below solves this problem. To this end, the model of section 7.1 is extended to the case of a differential output and simplified to the degenerated case of an infinite differential gain (and symmetric power-supply).

This limit case yields a multi-valued relation (see subsection 7.2.1 and appendix A p.271) that requires special care for numerical simulation. In this thesis, we do not consider solvers based on non-smooth dynamics and differential inclusions (see [AB08]). Instead, in subsection 7.2.2, we propose an alternative strategy based on *implicit continuous parametrisation* of the idealised amplifier relation (see definition 1.21 p.28). This follows the approach that we proposed in [MH20] and exploits the fact that the nonlinear law is in fact geometrically C^0-continuous.

7.2.1 Ideal Fully Differential Amplifier (FDA) model

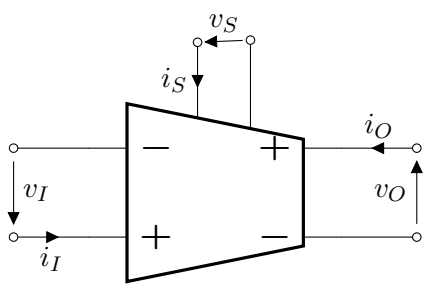


Figure 7.16 – (FDA) Ideal non-energetic Fully Differential Amplifier 3-port.

In this section, compared to section 7.1, we assume the following additional hypothesis:

- the supply voltages are symmetric $v_{\mathsf{S}+} = -v_{\mathsf{S}-} = v_S$,
- the output port is no longer referenced to the ground,
- we consider the limit case of the amplification gain $K \to \infty$,

Moreover, using the common-differential variable change introduced in section 2.5 p.73, because of symmetries (e.g. $e_S^\Sigma = e_S^+ + e_S^- = 0$ on fig. 7.3 p.176), the common-mode input and common-mode power supply have no influence on the model behaviour. We can reduce the FDA to a 3-port. We label ports $\{I, S, O\}$ for Input, Supply, Output, satisfying the set *relations* (see appx A p.271)

$$i_I \in \{0\} \qquad \text{(infinite input impedance)} \tag{7.51a}$$

$$v_O \in v_S \operatorname{sign}(v_I) \qquad \text{(saturating fully differential amplifier)} \tag{7.51b}$$

$$v_I i_I + v_S i_S + v_O i_O \in \{0\} \qquad \text{(conservative power balance)} \tag{7.51c}$$

Rewriting (7.51a)-(7.51c) yields $i_S \in -\operatorname{sign}(v_I) i_O$, which we summarize by the vector relation

$$\begin{bmatrix} i_I \\ v_O \\ i_S \end{bmatrix} \in \begin{bmatrix} . & . & . \\ . & . & \operatorname{sign}(v_I) \\ . & -\operatorname{sign}(v_I) & . \end{bmatrix} \begin{bmatrix} v_I \\ i_O \\ v_S \end{bmatrix}, \quad \text{where} \quad \operatorname{sign}(x) := \begin{cases} \{-1\} & x \in (-\infty, 0), \\ (-1, 1) & x \in \{0\}, \\ \{1\} & x \in (0, +\infty). \end{cases} \tag{7.52}$$

When solving circuits with (7.52), we may distinguish two situations:

- Either $v_I \neq 0$, the amplifier is in *saturation mode* (black curve in figure 7.17). Then v_O is *single-valued* and equal to either v_S or $-v_S$. This corresponds to the situation where the amplifier is used as a comparator to implement flip-flops, Schmidt triggers, etc.

- Or $v_I = 0$, the amplifier is in the vertical branch of the sign relation (red curve). This corresponds to infinite amplification. We call it the singular *nullor mode* (see [Car64, Mar65, Tel66, OU80]). This situation is very common. It is used to implement voltage buffers, virtual grounds, active filters, etc. Although v_O (and i_S) appear as multi-valued functions of v_I, in practice, a unique operating point is imposed by the external circuit.

The next sub section proposes a single-valued parametric representation to overcome the apparent difficulty of dealing with this multi-valued property.

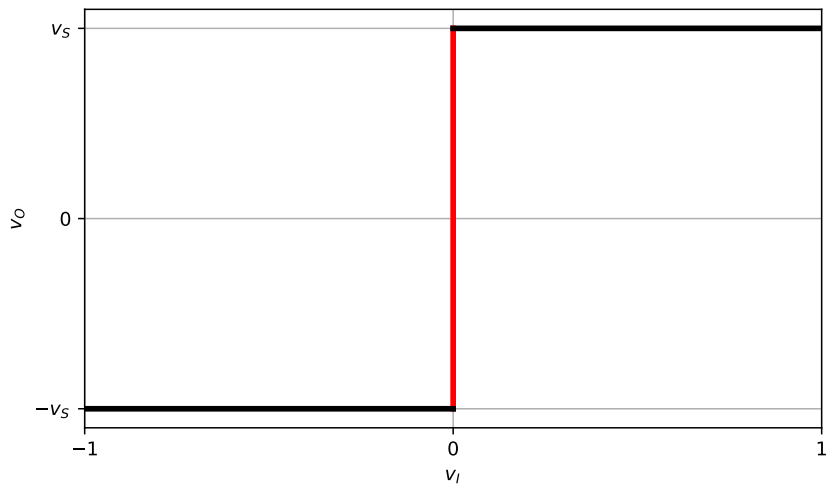


Figure 7.17 – (FDA) Ideal law in the (v_I, v_0)-plane expressed as a multi-valued function.

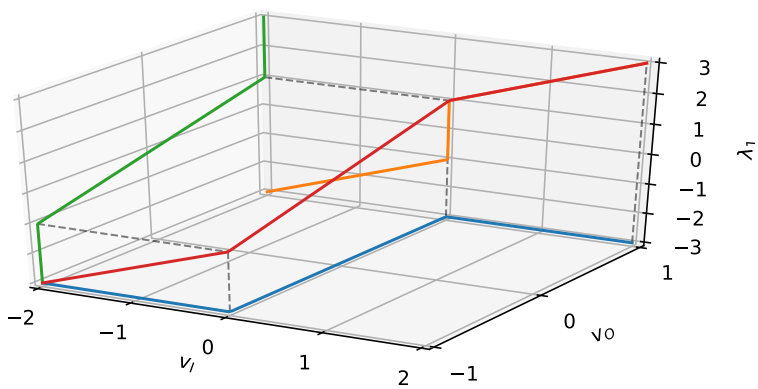


Figure 7.18 – (FDA) Ideal law in (v_I, v_0, λ_1) coordinates. The law is represented by an implicit $\mathcal{C}^0$-*continuous* map $\boldsymbol{\lambda} \mapsto (i_I, i_O, i_S, v_I, v_O, v_S)$ parametrised by $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3)$.

7.2.2 Continuous parametrisation

The (non-energetic [11]) relation (7.52) between $(v_I, i_O, v_S) \in \mathbb{R}^3$ and $(i_I, v_O, i_S) \in \mathbb{R}^3$ is multi-valued and may seem difficult to simulate. But this equation hides that the FDA admits a continuous geometrical description . The underlying continuous 3D manifold in this $\mathbb{R}^3 \times \mathbb{R}^3$-space can be described by the following parametric description (recall def. 1.21 p.28).

Introduce parameters $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3) \in \Lambda = \mathbb{R}^3$ to relate the currents $\mathbf{i} = (i_I, i_O, i_S) \in \mathbb{R}^3$ and voltages $\mathbf{v} = (v_I, v_O, v_S) \in \mathbb{R}^3$ of the FDA according to the *single-valued* relation

$$\mathcal{R}_{FDA} = \left\{ (\mathbf{i}, \mathbf{v}) \in \mathbb{R}^3 \times \mathbb{R}^3 \mid \begin{bmatrix} i_I \\ i_O \\ i_S \end{bmatrix} = \lambda_2 \begin{bmatrix} 0 \\ 1 \\ -\mu(\lambda_1) \end{bmatrix}, \begin{bmatrix} v_I \\ v_O \\ v_S \end{bmatrix} = \lambda_3 \begin{bmatrix} \mu^*(\lambda_1) \\ \mu(\lambda_1) \\ 1 \end{bmatrix}, \forall \boldsymbol{\lambda} \in \Lambda \right\}. \quad (7.53)$$

where the complementary modulation functions [12] μ, μ^* are defined by

$$\mu(x) := \begin{cases} -1 & x \leq -1 \\ x & x \in (-1, 1) \\ 1 & x \geq 1 \end{cases}, \qquad \mu^*(x) := x - \mu(x) = \begin{cases} x+1 & x \leq -1 \\ 0 & x \in (-1, 1) \\ x-1 & x \geq 1 \end{cases}, \quad (7.54)$$

and for which equations (7.51a)-(7.51c) are satisfied: this is obvious for (7.51a), straightforward for (7.51b) (compare also the (v_I, v_O)-planes of figures 7.17 and 7.18), and the (non-energetic) power balance (7.51c) is pointwise satisfied since

$$v_I \cdot i_I + v_O \cdot i_O + v_S \cdot i_S = \mu^*(\lambda_1)\lambda_3 \cdot 0 + \lambda_2\lambda_3\mu(\lambda_1) - \mu(\lambda_1)\lambda_2\lambda_3 = 0.$$

Description (7.53) (see fig. 7.20) shows that λ_2 and λ_3 are respectively controlled by i_O and v_S ($i_O = \lambda_2$ and $v_S = \lambda_3$). Because of the dual complementary functions μ, μ^* (see figure 7.19), parameter λ_1 is alternatively controlled by v_I in saturation mode and v_O in Nullor mode (but it still corresponds to a single one-dimensional constraint). This description can be reformulated as the single-valued relation (to be compared to the multi-valued one (7.52))

$$\begin{bmatrix} i_I \\ v_O \\ i_S \end{bmatrix} = \begin{bmatrix} . & . & . \\ . & . & \mu(\lambda_1) \\ . & -\mu(\lambda_1) & . \end{bmatrix} \begin{bmatrix} v_I = \mu^*(\lambda_1)\lambda_3 \\ i_O = \lambda_2 \\ v_S = \lambda_3 \end{bmatrix}. \quad (7.55)$$

An important property of (7.55) is that, contrary to (7.52), it is now explicit that for all λ_1 (for both linear and saturation modes) there exists a unique pair (v_O, i_S) and not a multi-valued set

Discussion: Nullors and computational causality To simplify circuit design and analysis, a common practice in electronic engineering is to use OPA in nullor mode, that is, to impose the double constraint $i_I = 0$, $v_I = 0$ (while i_O and v_O are unconstrained). But, as mentioned by Breedveld [Bre85, V.4], it is physically impossible to impose or control both effort and flow of one port. So, is the nullor mode paradoxical? How shall we interpret its double contraint $i_I = 0$, $v_I = 0$? To reconcile both viewpoints, thanks to (7.53), one can remark that the current constraint $i_I = 0$ is inherent to the device (it must be considered as an *output* of the FDA since it cannot be controlled whatever the mode). Conversely, v_I is an input of the device determining

11. See (7.51c) and references [WC77], [Bre85, VII.4] for the theory of nonlinear non-energetic n-ports.
12. Note the complementarity $\mu' + (\mu^*)' = 1$ and the Legendre transform duality $\int_0^x \mu \, \mathrm{d}x + \int_0^x \mu^* \, \mathrm{d}x = \frac{1}{2}x^2$.

its modes (through λ_1). The case $v_I = 0$ is *a consequence of the circuit operating point*. It holds only if v_O can be maintained in $(-v_S, v_S)$ out of the saturation mode. Indeed, as soon as v_O saturates, v_I is no longer zero. In practice, the Nullor mode region can be extended at will by increasing the supply voltage v_S. A clear analysis of causality arises by reformulating the FDA according to input-output common-differential ports introduced in subsection 7.2.3 p.194.

Figure 7.19 – (FDA) Dual functions μ, μ^* (left) and their derivative (right) used to implicitly parametrise the FDA relation (7.53). Note that similar functions have already been used (without being formalised) in figure 7.12 for the OPA.

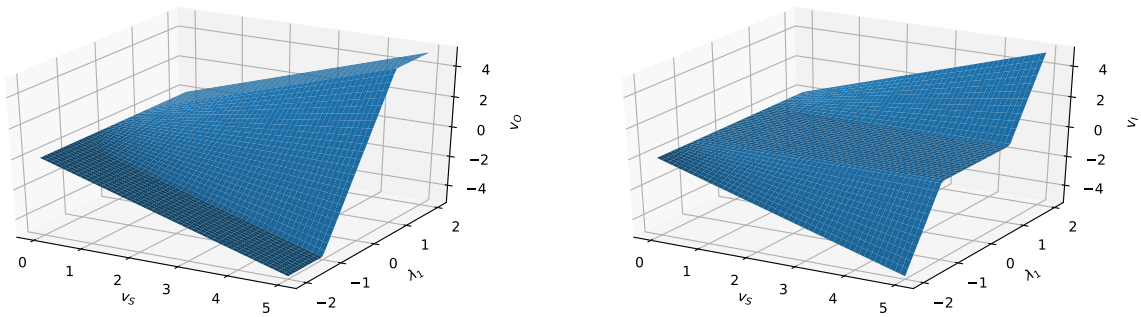


Figure 7.20 – (FDA) Ideal laws in the (v_S, λ_1, v_O)-space (left) and (v_S, λ_1, v_I)-space (right). Note that, according to (7.53), these laws are independent of the output current i_O and corresponds to a continuous function $(v_S, \lambda_1) \mapsto (v_O, v_I)$ and remind that $v_S = \lambda_3$.

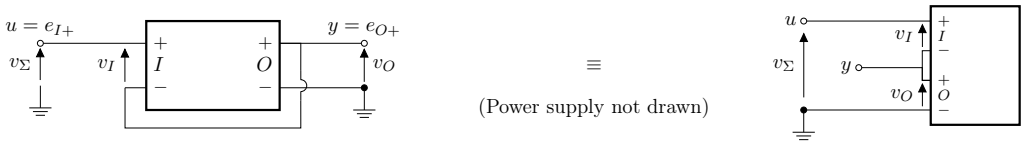


Figure 7.21 – (FDA) Voltage buffer. This examples shows a physical interpretation for the input-output common mode voltage ($v_\Sigma = v_O + v_I$) which is equal to the buffer input u.

7.2.3 Explicit formulation using common and differential ports

The understanding of causality is greatly simplified by switching to the unconventional [13] common and differential ports $\{\Sigma, \Delta\}$ built from input and output ports $\{I, O\}$. Indeed we show that parameter λ_1 can be *explicitly* controlled from the sum of input and output voltages [14] (see fig.7.21). Using theorem 2.5 (p.73), we perform the *power-preserving* port variables change $\{I, O\} \mapsto \{\Sigma, \Delta\}$ *between input and output* variables. We introduce the quantities

$$v_\Sigma := v_O + v_I, \qquad i_\Sigma := \frac{1}{2}\left(i_O + i_I\right), \tag{7.56a}$$

$$v_\Delta := v_O - v_I \qquad i_\Delta := \frac{1}{2}\left(i_O - i_I\right). \tag{7.56b}$$

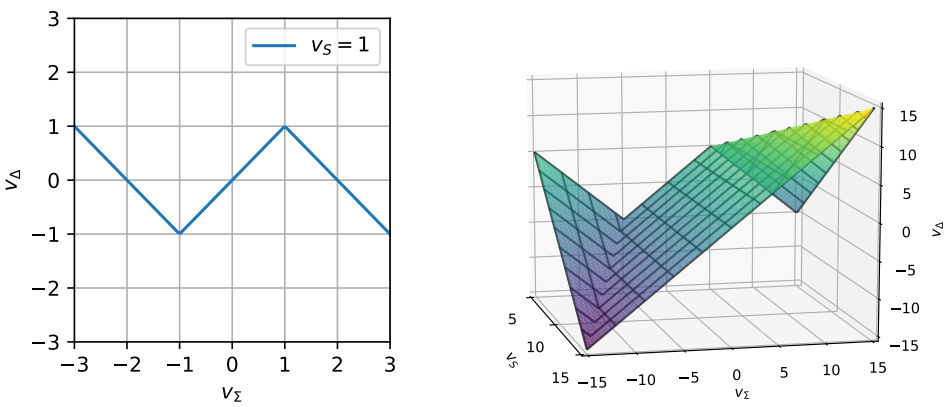
to form the alternative $\mathbb{R}^3 \times \mathbb{R}^3$ system of coordinates given by currents $\tilde{\mathbf{i}} = (i_\Sigma, i_\Delta, i_S) \in \mathbb{R}^3$, and voltages $\tilde{\mathbf{v}} = (v_\Sigma, v_\Delta, v_S) \in \mathbb{R}^3$. Substituting (7.56a) (7.56b) into equation (7.53) yields

$$v_\Sigma = \lambda_1 v_S, \qquad v_\Delta = (2\mu(\lambda_1) - \lambda_1) v_S, \qquad i_\Sigma = i_O/2, \qquad i_\Delta = i_O/2.$$

This shows that we can control parameter λ_1 (in (7.53)) from the (input-output) common mode voltage v_Σ and the power supply voltage v_S, while the map $i_\Delta \mapsto i_\Sigma$ is just the identity. We consider the differential mode v_Δ as an *output* and the common mode v_Σ as an *input* (see fig.7.21). By consequence the relation in eq. (7.53) can be written as the explicit skew-symmetric map

$$\mathcal{R}_{FDA} = \left\{ (\tilde{\mathbf{i}}, \tilde{\mathbf{v}}) \in \mathbb{R}^3 \times \mathbb{R}^3 \;\middle|\; \begin{bmatrix} i_S \\ i_\Sigma \\ v_\Delta \end{bmatrix} = \begin{bmatrix} . & . & -2\mu(\lambda_1) \\ . & . & 1 \\ 2\mu(\lambda_1) & -1 & . \end{bmatrix} \begin{bmatrix} v_S \\ v_\Sigma \\ i_\Delta \end{bmatrix}, \; \lambda_1 = \frac{v_\Sigma}{v_S} \right\}. \tag{7.57}$$

We see on figure 7.22b that increasing the power supply voltage v_S increases the nullor region ($v_\Sigma = v_\Delta \iff v_I = 0$), whereas in saturation ($|v_\Sigma| > |v_S|$) the output v_Δ is reflected about $\pm v_S$.



(a) 2D $v_\Sigma \mapsto v_\Delta$ **(b)** 3D $(v_\Sigma, v_S) \mapsto v_\Delta$

Figure 7.22 – (FDA) causal map in input-output Σ-Δ coordinates.

13. Common and differential modes in electronics are usually associated with positive and negative symmetries such as power supply or input ports in traditional OPA. Here we consider input-ouput variable changes.

14. Co-incidently, in the final stage of redaction, we found that "across-ports" wave-variable changes have just been proposed in [BMS20], precisely to handle operational amplifiers in WDF.

7.3 Towards a grey-box passive model of the OPA

In previous sections, we have considered nonlinear but idealised black-box OPA model with conservative memoryless saturating behaviour. To increase realism, additional non-ideal behaviours should be accounted for, such as those quantified in datasheets: finite gain-bandwidth product, slew-rate, internal dissipation, finite input impedance, non-zero output resistance, etc. As an alternative to a full physical modelling (of a dozen of transistors), this section opens a way towards a grey-box oriented pH modelling with an affordable simulation cost.

Some phenomena (such as input and output impedance, power-supply voltage drop, etc) can be modelled by composing the ideal OPA models with resistors, diode and capacitors (see [BPCS74, WDR+16]). However bandwidth, slew-rate and internal dissipation, require a finer level of description. A possible approach (common in the literature [SS98, CDK87]), is to use a 3-stage model (see figure 7.23): first a differential amplifier behaving like a (saturating) voltage-controlled current source; second a dynamic stage responsible for bandwidth (in linear mode) and slew rate (in saturation); and third a unity gain push-pull output distributing power from the supply port to the output load. A main difference with common modelling approaches in [Chu75, p.111] or [BPCS74] is that our proposition does not make use of voltage or current controlled sources to model sub-components but explicitly models power-supply ports and passivity. Due to time constraint, our full modelling is not complete. We propose to use OPA building blocks as shown in figure 7.24 and an explicit model of a BJT push-pull for large-signals is detailed in appendix D.9.3 p.302. Minimal pH models of these blocks will be completed in future work.

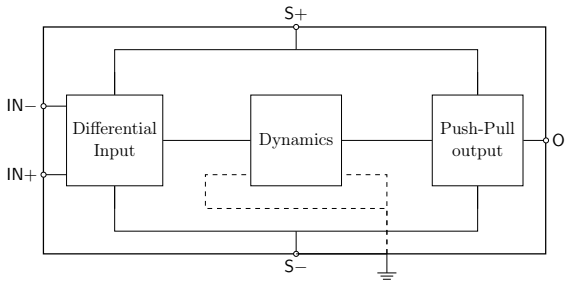
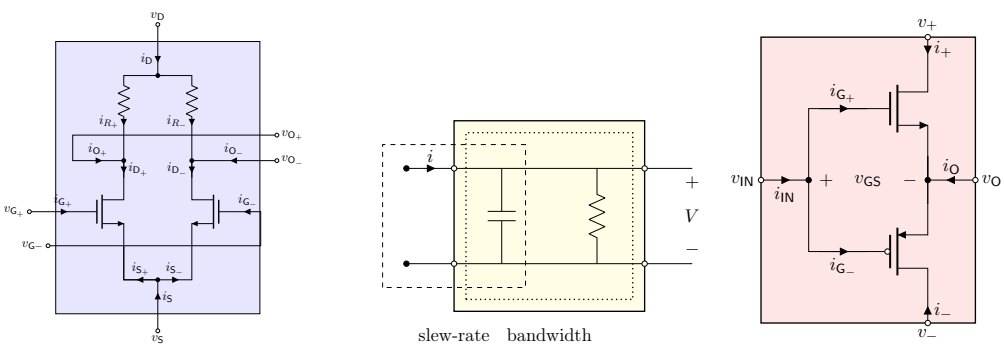


Figure 7.23 – (OPA, grey box model) structure of the macro model. Terminals are considered as ports by referencing them to the ground (not necessarily connected to the OPA).



(a) Differential input (no current source) **(b)** Dynamics **(c)** CMOS Push-Pull

Figure 7.24 – (OPA, grey box model) building blocks candidates.

Conclusion

In this chapter, we have proposed a minimal memoryless non-energetic model of the operational amplifier compatible with the the pH formalism. Surprisingly, despite the amount of (more advanced) publications on the subject and the abundant usage of OPA in electronics, we have not found in the literature such a nonlinear model, that is both energy-balanced and simple enough for standard use in most circuits. In order to stay within the PHS modelling framework, we had to propose a new model. Explicit modelling of power supply ports and saturation is a key ingredient to derive passive models and allows the modelling of non-ideal power-supply circuits (possibly modulated by the current of the output load).

As a further simplification and an alternative to pure nullors, we propose a 3-port fully differential amplifier with infinite gain. It includes (i) both nullor and saturation modes as special cases of a general relation (ii) a non-energetic memoryless modelling with an explicit port to model the power supply. To avoid the use of multi-valued relations, we propose a 3-dimensional implicit parametrisation of the component relation. This parametrisation is directly compatible with the simulation framework proposed in this thesis (chapter 5 p. 117), and in particular the fully implicit approach that we proposed in [MH20]. Other applications and simulations can be found in chapter 8.

Finally, the outline of a 3-stages grey-box pH model including slew-rate, finite gain-bandwidth and dissipation is sketched in section 7.3. The first steps to achieve this work have been developed: a common structure, candidate circuits for building blocks and an exact explicit input-output relation for a simplified BJT push-pull for large signals (see the technical details in appendix D.9.3 p.302). This preliminary result shows that an exact white-box modelling, although achievable, can quickly become overwhelmingly complex and does not scale with a high number of algebraic components. Due to time constraints, the derivation of simple and efficient pH realisations of the passive OPA building blocks from figure 7.23 (keeping the minimalist approach of [MH19]) is left for future research. Finally applications and simulation of circuits containing OPA are detailed in the next chapter.

Chapter 8

Circuits case studies

Contents

8.1 Fuzz Face (NPN variant) 199
8.2 Big Muff tone clipper 204
8.3 Tube Screamer drive stage 208
8.4 Korg MS-20 Filter 212
8.4.1 Overdrive amplifier 213
8.4.2 Filter 214
8.5 FitzHugh–Nagumo relaxation oscillator 217
8.6 Passive peaking equalizer (beyond the Nyquist frequency) 221
8.6.1 High-order RPM discretisation of a linear state-space system 224
8.6.2 Frequency response and Legendre filterbank interpretation 225

In this chapter, we consider a number of electronic audio circuits, chosen as representatives of the common situations and difficulties encountered when trying to simulate virtual analog audio circuits. All circuits are analysed and modelled systematically as pHS using the tools from chapters 1 and 2 (using both pH-DAE et pH-ODE formulations). We repeat the same process for each example in order to exhibit the common modelling steps as well as the different modelling and simulation strategies. The nonlinear systems are then discretised using the power-balanced projection methods from chapter 5 p.117 and solved using Newton iteration.

In section 8.1, we address the simulation of *stiff* pH-DAE with a variant of the classical FuzzFace circuit, a canonical design for *fuzz* guitar sounds.

In section 8.2, we merge the diode clipper circuit (already studied in chapters 2 and 5) with the tone-stack of the BigMuff Pi guitar pedal to produce a nonlinear tonestack (pH-ODE).

In section 8.3, we simulate the drive stage of the Tube Screamer guitar pedal. This is the occasion to consider a typical pattern used by electronic designers, namely *overdrive amplifiers* which saturates the feedback path of amplifiers. This is also the occasion to revisit the op amp model from chapter 7 in a different context.

In section 8.4, we consider a building block of analog synthesizers: we revisit the Sallen-Key filter topology from chapter 7, in this variant, the circuit uses 3 operational amplifiers to buffer stages and a nonlinear overdrive saturation in the feedback path (similar to the one of the TubeScreamer). These slight modifications can yield drastic changes to the sound and salient features of the filter such as *self-oscillations* and *inter-modulations*.

In section 8.5, we consider the FitzHugh-Nagumo relaxation oscillator which exhibit a limit cycle. With this circuit, we look more closely at the tunnel diode. This is an example of passive

component with a non-monotonous characteristic. The locally *negative incremental resistance* is the key ingredient used to favour the emergence of a *limit cycle* with both stable and non-stable equilibrium points. This is also the occasion to look at a system combining a *slow dynamic* (determining the period of oscillations) and *fast relaxations* when switching between stable states.

Finally in section 8.6, we consider a classical passive peaking equalizer whose resonance frequency *is much higher than the sampling rate*. Such a situation is traditionally solved through oversampling. By contrast, this use case is an opportunity to study the spectral properties of high-order projection methods from chapter 5 p.117. In particular, we look at their *extended bandwidth* using generalised sampling theory and compare with the oversampling approach.

Remark 8.1. All examples [a] in this chapter follow the same systematic derivation process $\mathsf{schematic} \rightarrow \mathsf{netlist} \rightarrow \mathsf{semi-explicit\ hybrid\ dirac\ structure} \rightarrow \mathsf{reduced\ dissipative\ structure}$. This process is detailed in figure 2.1 p.44. In step 3, to emphasize the sparse block-structure of $\mathbf{J}$ matrices, port-Hamiltonian systems are standardized under the following semi-explicit tree / cotree form (see (2.18) p.55)

$$\begin{bmatrix} \mathbf{i}_T \\ \mathbf{v}_L \end{bmatrix} = \begin{bmatrix} \mathbf{0} & -\mathbf{C}_L \\ \mathbf{C}_L^{\mathsf{T}} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{V}_T(\mathbf{i}_T) \\ \boldsymbol{I}_L(\mathbf{v}_T) \end{bmatrix} \qquad \equiv \qquad \mathbf{J} = \begin{array}{c} \\ \mathbf{i}_T \\ \mathbf{v}_L \end{array} \begin{array}{c} \begin{array}{cc} \mathbf{v}_T & \mathbf{i}_T \end{array} \\ \begin{bmatrix} \mathbf{0} & -\mathbf{C}_L \\ \mathbf{C}_L^{\mathsf{T}} & \mathbf{0} \end{bmatrix} \end{array},$$

where algebro-differential operators $\boldsymbol{V}_T$, $\boldsymbol{I}_L$ respectively stand for component laws of current-controlled tree branches and voltage-controlled cotree branches (links) and $\mathbf{C}_L$ is the link cutset matrix obtained from circuit incidence matrices according to eq. (2.15) p.55.

As a further simplification, in step 4, linear resistive branches are pre-solved to canonically obtain the following resistive tree/cotree formulation

$$\begin{bmatrix} \mathbf{i}_T \\ \mathbf{v}_L \end{bmatrix} = \begin{bmatrix} \mathbf{G}_T & -\boldsymbol{\alpha}^{\mathsf{T}} \\ \boldsymbol{\alpha} & \mathbf{R}_L \end{bmatrix} \begin{bmatrix} \boldsymbol{V}_T(\mathbf{i}_T) \\ \boldsymbol{I}_L(\mathbf{v}_T) \end{bmatrix} \qquad \equiv \qquad \mathbf{M} = \begin{array}{c} \\ \mathbf{i}_T \\ \mathbf{v}_L \end{array} \begin{array}{c} \begin{array}{cc} \mathbf{v}_T & \mathbf{i}_T \end{array} \\ \begin{bmatrix} \mathbf{G}_T & -\boldsymbol{\alpha}^{\mathsf{T}} \\ \boldsymbol{\alpha} & \mathbf{R}_L \end{bmatrix} \end{array},$$

where $\mathbf{G}_T$ is the tree conductance matrix, $\mathbf{R}_L$ is the link resistance matrix and $\boldsymbol{\alpha}$ is a tree/cotree matrix transformer ratio (see subsection 2.3.4 p.60). These two forms can be directly simulated thanks to our passivity-preserving projection theorem 5.1 p.119. Finally, adhoc reduction to ODE or DAE subsets is performed where appropriate.

For simplicity of exposition, power-balanced simulations are obtained using discretisation by projection with RPM(1,0) [b] (see definitions 5.1 p.122 for pH-ODE and 5.2 p.123 for pH-DAE).

a. Except in the MS-20 example: due to the high number of branches (34), we use nodal analysis to jump straight to the most reduced formulation.

b. Projection order $p = 1$, regularity order $k = 0$ (equivalent to the average vector field method).

Remark 8.2 (Practical existence / uniqueness conditions and Newton convergence)**.** Existence / uniqueness conditions have been studied in 5.2.3 p.127 for pH-ODE and (partially) in 5.3.2 p.135 for pH-DAE. However sharp *practical conditions* are still missing. Indeed, while practical convergence is always observed in presented simulations, theoretical convergence bounds are either missing, or too restrictive, in particular for stiff systems. For this reason, convergence conditions will not be detailed in upcoming examples. This important but difficult topic is left for future research.

8.1 Fuzz Face (NPN variant)

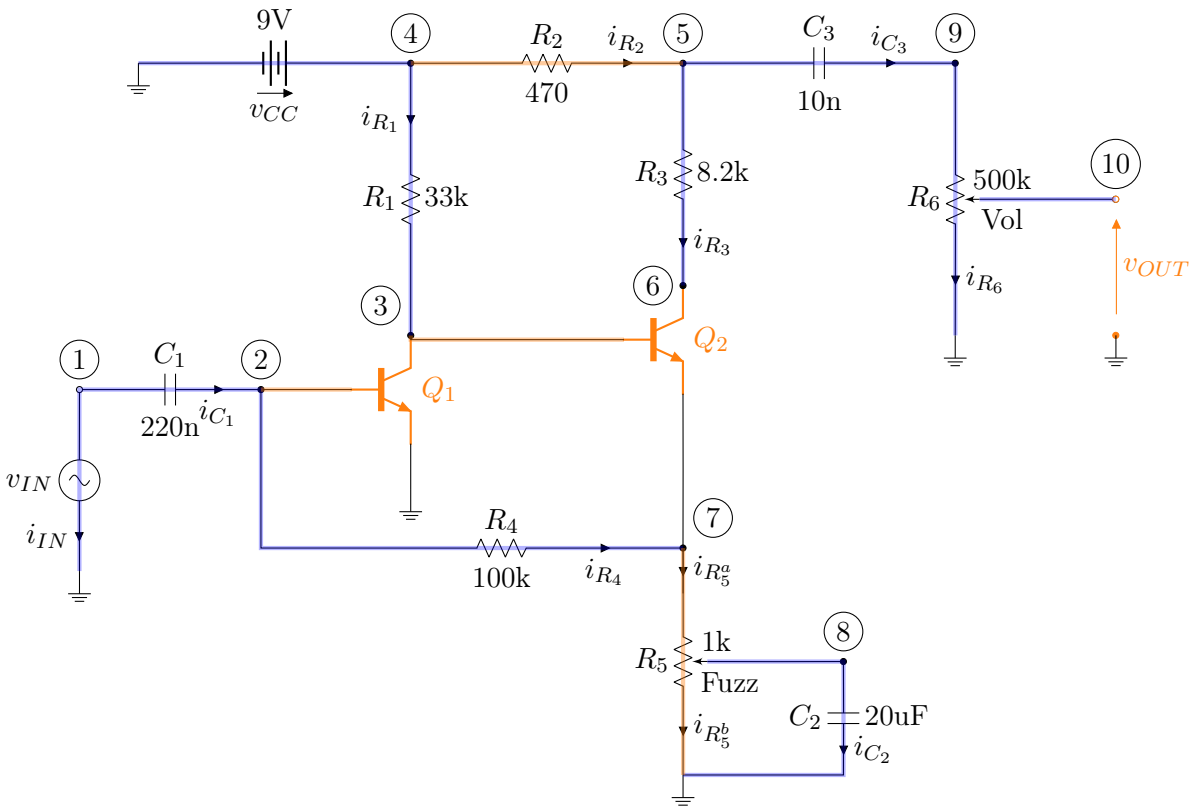


Figure 8.1 – (NPN Fuzz Face) Schematic. The chosen spanning tree T (current-controlled) is shown in blue. Complementarily, its cotree $\overline{T}$ (voltage-controlled) is shown in orange.

The *Fuzz Face* is an effect pedal for electric guitar designed to produce a distorted *fuzz* sound (reminiscent of the buzzing sound of damaged speakers [1]). It was conceived in 1966 by Arbiter Electronics Ltd and made famous by guitarists such as Jimi Hendrix (with custom modifications made by Roger Mayer), David Gilmour (Pink Floyd), Pete Townshend (The Who). The original design uses Germanium PNP transistors (positive ground, negative voltage source). A number of imitations, tribute and modifications have been proposed: Vox *Tone bender*, Mike Fuller's *'69 Fulltone* or more recently ZVEX *Woolly Mammoth*. The circuit has been studied in [COCR09, DZ11a, HHVW17, Hol19]. Here, we consider the NPN [2] variant of figure 8.1 which is obtained by replacing PNP by NPN transistors and inverting the power supply. For simulation, we use 2N3904 transistors with parameters $I_S = 10$ fA, $\beta_F = 300$ and $\beta_R = 4$ using the memoryless Ebers–Moll model. This circuit is an opportunity to see that in electronics, many components are resistors. But since the majority are linear, a significant reduction in the number of unknowns can be achieved by pre-solving linear constraints (the price to pay is denser matrices). As often in electronics, this circuit yields a pH-DAE that is not explicitly convertible to a pH-ODE. This is the occasion to look at the direct simulation of pH-DAE on a real circuit.

1. The song *Rocket 88* by Ike Turner and Jackie Brenston is often credited as the first "rock and roll" song featuring a damaged speaker. The songs *Rumble* by Link Wray and *You really got me* by The Kinks also feature speakers damaged on purpose to obtain a *fuzz* sound.
2. The Woolly Mammoth is also NPN.

Theory of operation As the behaviour and the design of the Fuzz Face are well documented, we only provide a short description. It can be roughly described as a (voluntarily badly biased) two stages common-emitter transistor amplifiers with feedback. The biasing is responsible for asymmetrical clipping and even harmonics generation. The cascade of two transistors was used (before OPA) to achieve a higher distortion gain. For more details, see reference [Ele20a].

Incidence matrix For the chosen orientation of branches[3], the incidence matrix (definition 2.12 p.49) of the graph corresponding to the fuzz face schematic (figure 8.1) is given by

$$\mathbf{A} = \begin{array}{c} \\ ⓪ \\ ① \\ ② \\ ③ \\ ④ \\ ⑤ \\ ⑥ \\ ⑦ \\ ⑧ \\ ⑨ \end{array} \overbrace{\begin{bmatrix} IN & CC & C_1 & C_2 & C_3 & R_1 & R_3 & R_4 & R_6 & R_2 & R_5^a & R_5^b & BC1 & BE1 & BC2 & BE2 & OUT \\ -1 & -1 & . & -1 & . & . & . & . & -1 & . & . & -1 & . & -1 & . & . & -1 \\ +1 & . & +1 & . & . & . & . & . & . & . & . & . & . & . & . & . & . \\ . & . & -1 & . & . & . & . & +1 & . & . & . & . & +1 & +1 & . & . & . \\ . & . & . & . & . & -1 & . & . & . & . & . & . & -1 & . & +1 & +1 & . \\ . & +1 & . & . & . & +1 & . & . & . & +1 & . & . & . & . & . & . & . \\ . & . & . & . & +1 & . & +1 & . & . & -1 & . & . & . & . & . & . & . \\ . & . & . & . & . & . & -1 & . & . & . & . & . & . & . & -1 & . & . \\ . & . & . & . & . & . & . & -1 & . & . & +1 & . & . & . & . & -1 & . \\ . & . & . & +1 & . & . & . & . & . & . & -1 & +1 & . & . & . & . & . \\ . & . & . & . & -1 & . & . & . & +1 & . & . & . & . & . & . & . & 1 \end{bmatrix}}^{\text{Branches } B}.$$

Dirac structure Using the causality assignment procedure detailed in subsection 2.3.3 p.57, we select the minimum spanning tree (def. 2.9 p.48) $T = \{IN, CC, C_1, C_2, C_3, R_1, R_3, R_4, R_6\}$, to split branches B into a current-controlled tree T and voltage-controlled cotree $\overline{T}$ (links). From the incidence matrix $\mathbf{A}$, using equation (2.15) p.55, we obtain the link cutset matrix $\mathbf{C}_L$ so that the circuit is described by the reduced hybrid Dirac structure (def. 2.21 p.55)

$$\mathbf{J} = \begin{array}{c} \\ \mathbf{i}_T \left\{ \begin{array}{c} i_{IN} \\ i_{CC} \\ i_{C_1} \\ i_{C_2} \\ i_{C_3} \\ i_{R_1} \\ i_{R_3} \\ i_{R_4} \\ i_{R_6} \end{array} \right. \\ \mathbf{v}_{\overline{T}} \left\{ \begin{array}{c} v_{R_2} \\ v_{R_5^a} \\ v_{R_5^b} \\ v_{BC1} \\ v_{BE1} \\ v_{BC2} \\ v_{BE2} \\ v_{OUT} \end{array} \right. \end{array} \overbrace{\begin{bmatrix} v_{IN} & v_{CC} & v_{C_1} & v_{C_2} & v_{C_3} & v_{R_1} & v_{R_3} & v_{R_4} & v_{R_6} \\ . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & -\mathbf{C}_L \\ . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . \\ 0 & +1 & 0 & 0 & -1 & 0 & 0 & 0 & -1 \\ +1 & 0 & -1 & -1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & +1 & 0 & 0 & 0 & 0 & 0 \\ +1 & -1 & -1 & 0 & 0 & +1 & 0 & 0 & 0 \\ +1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & +1 & 0 & 0 & -1 & -1 & +1 & 0 & -1 \\ -1 & +1 & +1 & 0 & 0 & -1 & 0 & +1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}}^{\mathbf{v}_T} \overbrace{\begin{bmatrix} i_{R_2} & i_{R_5^a} & i_{R_5^b} & i_{BC1} & i_{BE1} & i_{BC2} & i_{BE2} & i_{OUT} \\ 0 & -1 & 0 & -1 & -1 & 0 & +1 & 0 \\ -1 & 0 & 0 & +1 & 0 & -1 & -1 & 0 \\ 0 & +1 & 0 & +1 & +1 & 0 & -1 & 0 \\ 0 & +1 & -1 & 0 & 0 & 0 & 0 & 0 \\ +1 & 0 & 0 & 0 & 0 & +1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & +1 & +1 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & +1 & 0 & 0 & 0 & 0 & -1 & 0 \\ +1 & 0 & 0 & 0 & 0 & +1 & 0 & -1 \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \end{bmatrix}}^{\mathbf{i}_{\overline{T}}} \mathbf{C}_L^{\mathsf{T}} \end{array}.$$

Note that the canonical separation between tree and link/cotree variables has been emphasised by the ordering of component: tree currents $\mathbf{i}_T$ (left) can only exchange with cotree currents $\mathbf{i}_{\overline{T}}$ (right), while cotree voltages $\mathbf{v}_{\overline{T}}$ (left) can only exchange with tree voltages $\mathbf{v}_T$ (right).

3. Using the receiver convention, branch currents are oriented from positive nodes (+1) to negative nodes (−1).

Reduced dissipative structure To simplify simulation, we eliminate linear resistive branches $\left\{R_1, R_3, R_4, R_6, R_2, R_5^a, R_5^b\right\}$ by solving the corresponding linear resistive constraints, (see subsection 2.3.4, p.60 and [FH16a, Fal16, Lop16]). Reducing linear resistive relations, the Dirac structure matrix $\mathbf{J}$ is replaced by the (hybrid) linear dissipative structure[4] matrix

$$\mathbf{M} = \begin{array}{c} \\ i_{IN} \\ i_{CC} \\ i_{C_1} \\ i_{C_2} \\ i_{C_3} \\ v_{BC1} \\ v_{BE1} \\ v_{BC2} \\ v_{BE2} \\ v_{OUT} \end{array} \begin{array}{c} \begin{array}{cccccccccc} v_{IN} & v_{CC} & v_{C_1} & v_{C_2} & v_{C_3} & i_{BC1} & i_{BE1} & i_{BC2} & i_{BE2} & i_{OUT} \end{array} \\ \left[\begin{array}{cccccccccc} -G_{11} & . & G_{11} & G_{11} & . & -1 & -1 & . & \alpha_{14} & . \\ . & -G_{22} & . & . & G_{22} & +1 & . & -\alpha_{23} & -1 & -\alpha_{35} \\ G_{11} & . & -G_{11} & -G_{11} & . & +1 & +1 & . & -\alpha_{14} & . \\ G_{11} & . & -G_{11} & -G_{44} & . & . & . & . & \alpha_{45} & . \\ . & G_{22} & . & . & -G_{22} & . & . & \alpha_{23} & . & \alpha_{35} \\ +1 & -1 & -1 & . & . & -R_1 & . & R_1 & R_1 & . \\ +1 & . & -1 & . & . & . & . & . & . & . \\ . & \alpha_{23} & . & . & -\alpha_{23} & R_1 & . & -R_{33} & -R_1 & R_{35} \\ -\alpha_{14} & +1 & \alpha_{14} & -\alpha_{45} & . & R_1 & . & -R_1 & -R_{44} & . \\ . & \alpha_{35} & . & . & -\alpha_{35} & . & . & R_{35} & . & -R_{35} \end{array}\right] \end{array}, \tag{8.1}$$

where the conductances, gains and resistances parameters are

$$G_{11} = \frac{1}{R_4 + R_5^a}, \qquad G_{22} = \frac{1}{R_2 + R_6}, \qquad G_{44} = \frac{R_4 + R_5^a + R_5^b}{R_5^b\left(R_4 + R_5^a\right)},$$

$$\alpha_{14} = \frac{R_5^a}{R_4 + R_5^a}, \qquad \alpha_{23} = \frac{R_2}{R_2 + R_6}, \qquad \alpha_{35} = \frac{R_6}{R_2 + R_6}, \qquad \alpha_{45} = \frac{R_4}{R_4 + R_5^a},$$

$$R_{33} = \frac{R_2 R_6 + (R_1 + R_3)(R_2 + R_6)}{R_2 + R_6}, \quad R_{35} = \frac{R_2 R_6}{R_2 + R_6}, \quad R_{44} = \frac{R_1\left(R_4 + R_5^a\right) + R_4 R_5^a}{R_4 + R_5^a}.$$

Note that it is structured into a skew-symmetric part and a dissipative part of the form

$$\mathbf{M} = \begin{bmatrix} \mathbf{0} & -\boldsymbol{\alpha}^\mathsf{T} \\ \boldsymbol{\alpha} & \mathbf{0} \end{bmatrix} - \begin{bmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{bmatrix},$$

where $\mathbf{G} = \mathbf{G}^\mathsf{T} \succeq 0$ denotes the tree conductance matrix and $\mathbf{R} = \mathbf{R}^\mathsf{T} \succeq 0$ the cotree resistance matrix, while $\boldsymbol{\alpha}$ plays the role of adimensioned multi-dimensional transformer ratio (whose values have a magnitude less or equal to 1, see example 1.4 p.7). Since nonlinear transistor elements are coupled instantaneously through the (positive semi-definite) resistive matrix $\mathbf{R}$, further reduction to an explicit pH-ODE would require the implicit function theorem. Instead we use direct pH-DAE simulation implemented as follows.

pH-DAE Discretization We identify equations corresponding to implicitly defined variables

$$\mathbf{x} = \left[i_{C1}, i_{C2}, i_{C3}, v_{BC1}, v_{BE1}, v_{BC1}, v_{BE2}\right]^\mathsf{T}.$$

Once these variables are solved, then i_{IN}, i_{CC} and v_{OUT} are also determined (by rows 1,2,10 in (8.1)). To keep notation simple and for space reasons, we focus on the first order Average Vector Field (RPM methods with $p = 1$, $k = 0$) whose projector $\mathcal{P} : L^2(\Omega) \to \mathbb{P}^0(\Omega)$, denotes

4. Please refer to corollary 5.2 p.120 for the power balanced projection of linear dissipative structures.

projection on the space of constant functions [5]. We denote $\bar{u}$ the average projection coefficient of a function $u(t)$ over a time step $(t_0, t_0 + h)$ so that $(\mathcal{P}u)(t) = 1_\Omega(t) \cdot \bar{u}$. For linear capacitors and an affine temporal model of charge $q(t) = q_0 + \int_0^t \bar{i}(s)\,\mathrm{d}s$, the projected effort law $\overline{V}_C$, is

$$\overline{V}_C(q_0; \bar{i}_C) := \mathcal{P}\left(\frac{1}{C}\left(q_0 + h\int_0^t \bar{i}_C(s)\,\mathrm{d}s\right)\right) = \frac{q_0}{C} + \frac{h}{2C}\bar{i}_C. \tag{8.2}$$

For bipolar transistors (ex. 1.10 p.32), and (only for) piecewise constants signals $\bar{v}_{BC}, \bar{v}_{BE}$, the projected law equals the original nonlinearity (evaluated for the averaged voltages)

$$\begin{bmatrix} \overline{I}_{BC}(\bar{v}_{BC}, \bar{v}_{BE}) \\ \overline{I}_{BE}(\bar{v}_{BC}, \bar{v}_{BE}) \end{bmatrix} = \begin{bmatrix} \gamma_R & -1 \\ -1 & \gamma_F \end{bmatrix} \begin{bmatrix} \mathrm{pn}(\bar{v}_{BC}) \\ \mathrm{pn}(\bar{v}_{BE}) \end{bmatrix}.$$

Splitting $\mathbf{M}$, in equation (8.1), according to inputs $\bar{\mathbf{u}} = (\bar{v}_{IN}, \bar{v}_{CC})$ and unknown variables $\mathbf{x}$, and using the law of the output open circuit ($i_{OUT} = 0$ in fig. 8.1), we obtain the following discrete algebraic equations [6]

$$\bar{\mathbf{x}} = \tilde{\mathbf{A}}\bar{\boldsymbol{e}}(\bar{\mathbf{x}}) + \tilde{\mathbf{B}}\bar{\mathbf{u}}, \tag{8.3}$$

where matrices $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{B}}$ (extracted from $\mathbf{M}$ according to $\bar{\mathbf{x}}$ and $\bar{\mathbf{u}}$) are

$$\tilde{\mathbf{A}} = \left[\begin{array}{ccc|cccc} -G_{11} & -G_{11} & . & +1 & +1 & . & -A_{14} \\ -G_{11} & -G_{44} & . & . & . & . & A_{45} \\ . & . & -G_{22} & . & . & A_{23} & . \\ \hline -1 & . & . & -R_1 & . & R_1 & R_1 \\ -1 & . & . & . & . & . & . \\ . & . & -A_{23} & R_1 & . & -R_{33} & -R_1 \\ A_{14} & -A_{45} & . & R_1 & . & -R_1 & -R_{44} \end{array}\right], \qquad \tilde{\mathbf{B}} = \left[\begin{array}{cc} G_{11} & . \\ G_{11} & . \\ . & G_{22} \\ \hline +1 & -1 \\ +1 & . \\ . & A_{23} \\ -A_{14} & . \end{array}\right],$$

and where the projected variables $\bar{\mathbf{x}}$, $\bar{\mathbf{u}}$ and projected laws $\bar{\boldsymbol{e}}(\bar{\mathbf{x}})$ are

$$\bar{\mathbf{x}} = \left[\begin{array}{c} \bar{i}_{C_1} \\ \bar{i}_{C_2} \\ \bar{i}_{C_3} \\ \hline \bar{v}_{BC1} \\ \bar{v}_{BE1} \\ \bar{v}_{BC2} \\ \bar{v}_{BE2} \end{array}\right], \qquad \bar{\boldsymbol{e}}(\bar{\mathbf{x}}) = \left[\begin{array}{c} \overline{V}_{C_1}(q^0_{C_1}; \bar{i}_{C_1}) \\ \overline{V}_{C_2}(q^0_{C_2}; \bar{i}_{C_2}) \\ \overline{V}_{C_3}(q^0_{C_3}; \bar{i}_{C_3}) \\ \hline \overline{I}_{BC1}(\bar{v}_{BC1}, \bar{v}_{BE1}) \\ \overline{I}_{BE1}(\bar{v}_{BC1}, \bar{v}_{BE1}) \\ \overline{I}_{BC2}(\bar{v}_{BC2}, \bar{v}_{BE2}) \\ \overline{I}_{BE2}(\bar{v}_{BC2}, \bar{v}_{BE2}) \end{array}\right], \qquad \bar{\mathbf{u}} = \begin{bmatrix} \bar{v}_{IN} \\ \bar{v}_{CC} \end{bmatrix}.$$

First, (8.3) is solved using Newton iteration by looking for the root of $\boldsymbol{F}(\bar{\mathbf{x}}) = 0$, where

$$\boldsymbol{F}(\bar{\mathbf{x}}) = \bar{\mathbf{x}} - \tilde{\mathbf{A}}\bar{\boldsymbol{e}}(\bar{\mathbf{x}}) - \tilde{\mathbf{B}}\bar{\mathbf{u}}.$$

Then we compute $\bar{v}_{OUT}$ from $\bar{\mathbf{x}}$ and the observer equation (the last row of $\mathbf{M}$ in (8.1))

$$\bar{v}_{OUT} = A_{35}\left(\bar{v}_{CC} - \bar{v}_{IN}\right) + R_{35}\overline{I}_{BC2}(\bar{v}_{BC2}, \bar{v}_{BE2}).$$

5. See example 5.5.2 from chapter 5 for generalisations to higher projection order.

6. Note that, since capacitors are linear, one could further reduce the size of the algebraic equations to the four nonlinear transistor branches. We do not perform this reduction to show the interaction between (discretized) differential and algebraic equations.

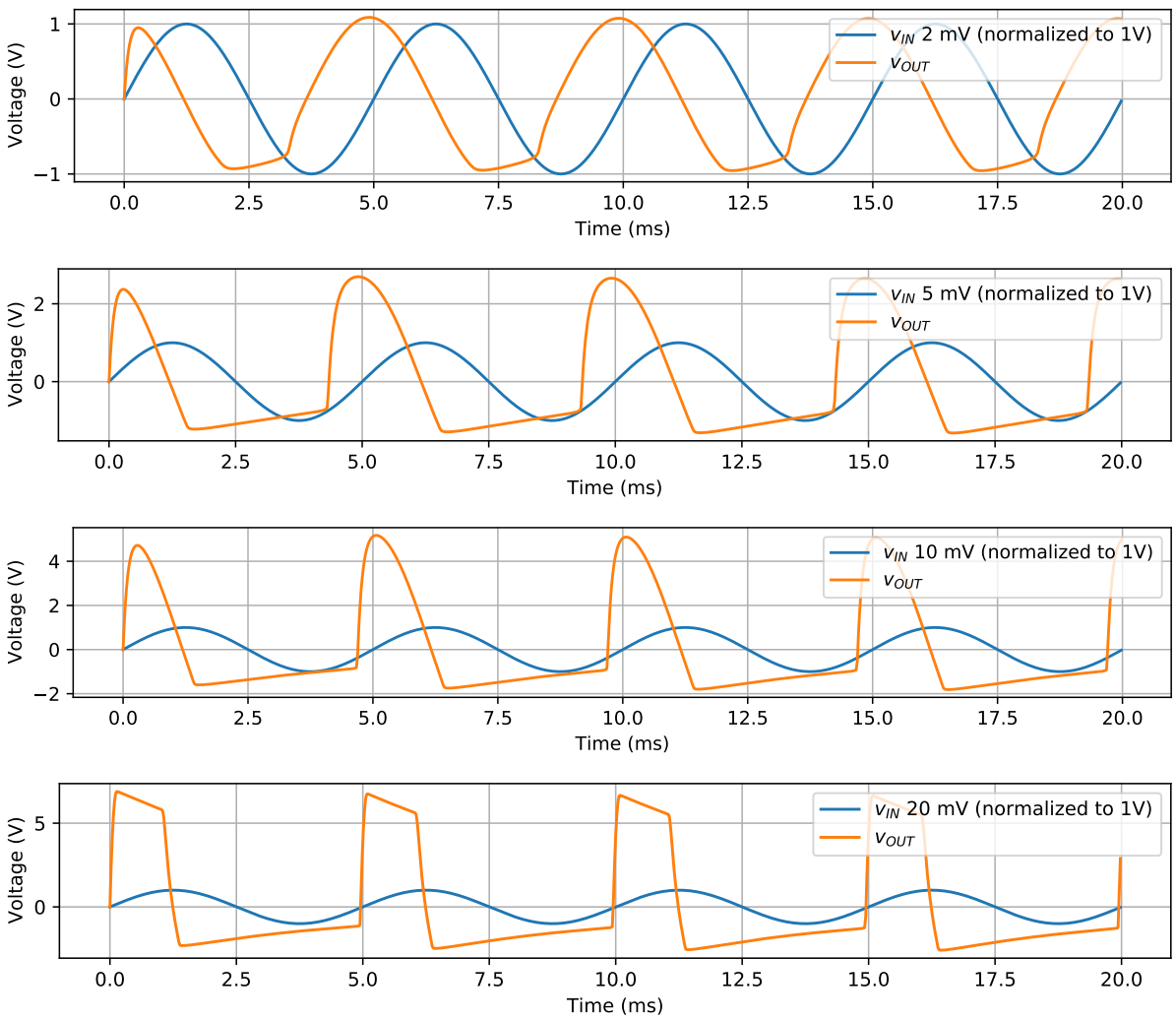


Figure 8.2 – (NPN Fuzz Face) simulation for a sine input with magnitudes $\{2, 5, 10, 20\}$ mV, frequency $f_0 = 200$ Hz and sampling rate $f_s = 44.1$ kHz. Note the asymmetrical distortion. The *fuzz* sound is roughly characterised by the transformation of the input into a (filtered) square wave with uneven pulse width. Convergence is reached after 1 to 5 iterations (1.671 on average).

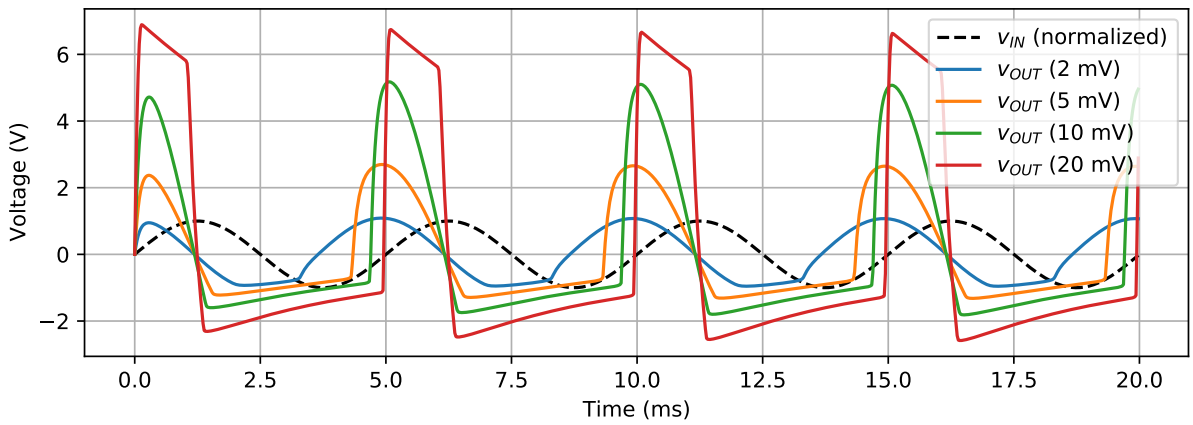


Figure 8.3 – (NPN Fuzz Face) Overlay of simulations from figure 8.2. As expected, we observe gradual asymmetrical clipping of the waveform as the gain is increased (consistent with SPICE).

8.2 Big Muff tone clipper

In this section we consider a nonlinear filter designed by simply merging the circuit of the original Big Muff π tone filter (red+blue) with the circuit of a diode clipper (green part on fig. 8.4). This non-trivial circuit is chosen for its relative simplicity, for the commonness of its constituent parts and because it can be reduced to a pH-ODE.

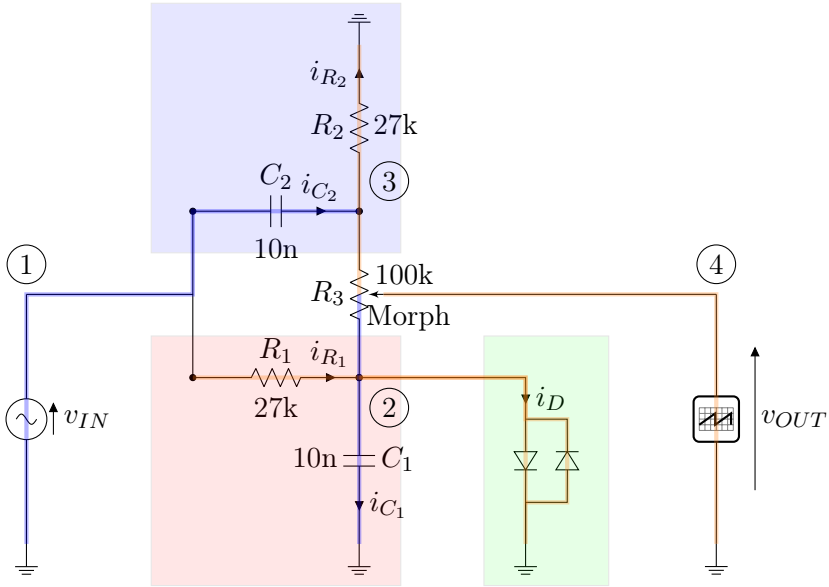


Figure 8.4 – (BMP Tone clipper) Schematics. Current-controlled spanning tree T shown in blue. Voltage-controlled cotree branches $\overline{T}$ in orange.

Theory of operation The BigMuff π tone circuit consists of a passive cross fade (through resistor R_3: $R_3^a = mR_3$, $R_3^b = (1-m)R_3$, $m \in [0,1]$) between a first order lowpass filter (R_1, C_1) (red block in fig. 8.4) and first order highpass filter (R_2, C_2) (blue block). As the combination of both circuits is unbuffered, the two filters interact. Moreover, the output voltage of the lowpass filter R_1, C_1 is clipped by diodes D_1, D_2 (in green) but since the circuit is passive, it also influences the high pass filter branch in a nonlinear way. As a result (see figure 8.5), the lowpass and highpass branches roughly produce smoothed square and triangular voltages respectively (for a sinusoidal input).

Incidence matrix The incidence matrix of the BMP graph shown on figure 8.4 is given by

$$\mathbf{A} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccccccccc} IN & C_1 & C_2 & R_3^b & R_1 & R_2 & R_3^a & D & OUT \end{array} \\ \begin{bmatrix} -1 & -1 & . & . & . & -1 & . & -1 & -1 \\ +1 & . & +1 & . & +1 & . & . & . & . \\ . & +1 & . & -1 & -1 & . & . & +1 & . \\ . & . & -1 & . & . & +1 & +1 & . & . \\ . & . & . & +1 & . & . & -1 & . & +1 \end{bmatrix} \end{array}$$

Dirac structure We select a spanning tree $T = \{IN, C_1, C_2, R_3^b\}$ to obtain the following Dirac structure matrix (encoding Kirchhoff laws)

$$\mathbf{J} = \begin{array}{c} \\ i_{IN} \\ i_{C1} \\ i_{C2} \\ i_{R_3^b} \\ v_{R1} \\ v_{R2} \\ v_{R_3^a} \\ v_D \\ v_{OUT} \end{array} \begin{array}{c} \begin{array}{ccccccccc} v_{IN} & v_{C1} & v_{C2} & v_{R_3^b} & i_{R1} & i_{R2} & i_{R_3^a} & i_D & i_{OUT} \end{array} \\ \begin{bmatrix} . & . & . & . & -1 & -1 & -1 & 0 & 0 \\ . & . & . & . & +1 & 0 & +1 & -1 & -1 \\ . & . & . & . & 0 & +1 & +1 & 0 & 0 \\ . & . & . & . & 0 & 0 & +1 & 0 & -1 \\ +1 & -1 & 0 & 0 & . & . & . & . & . \\ +1 & 0 & -1 & 0 & . & . & . & . & . \\ +1 & -1 & -1 & -1 & . & . & . & . & . \\ 0 & +1 & 0 & 0 & . & . & . & . & . \\ 0 & +1 & 0 & 1 & . & . & . & . & . \end{bmatrix} \end{array}.$$

Reduced dissipative structure The reduction of linear resistive relations yields the linear dissipative structure matrix

$$\mathbf{M} = \begin{array}{c} \\ i_{IN} \\ i_{C1} \\ i_{C2} \\ v_D \\ v_{OUT} \end{array} \begin{array}{c} \begin{array}{ccccc} v_{IN} & v_{C1} & v_{C2} & i_D & i_{OUT} \end{array} \\ \begin{bmatrix} -G_{11} & -G_{12} & -G_{13} & 0 & -\alpha_{21} \\ -G_{12} & -G_{22} & -G_{23} & -1 & -\alpha_{22} \\ -G_{13} & -G_{23} & -G_{33} & 0 & -\alpha_{23} \\ 0 & +1 & 0 & 0 & 0 \\ \alpha_{21} & \alpha_{22} & \alpha_{23} & 0 & -R_{22} \end{bmatrix} \end{array}, \tag{8.4}$$

where the tree conductance matrix $\mathbf{G}$, cotree resistance matrix $\mathbf{R}$ and transformation ratio $\boldsymbol{\alpha}$ are

$$\mathbf{G} = \begin{bmatrix} \frac{R_1R_2+(R_1+R_2)R_3}{R_1R_2R_3} & -\frac{R_1+R_3}{R_1R_3} & -\frac{R_2+R_3}{R_2R_3} \\ -\frac{R_1+R_3}{R_1R_3} & \frac{R_1+R_3}{R_1R_3} & \frac{1}{R_3} \\ -\frac{R_2+R_3}{R_2R_3} & \frac{1}{R_3} & \frac{R_2+R_3}{R_2R_3} \end{bmatrix}, \qquad \mathbf{R} = \begin{bmatrix} 0 & 0 \\ 0 & m(1-m)R_3 \end{bmatrix},$$

$$\boldsymbol{\alpha} = \begin{bmatrix} 0 & +1 & 0 \\ (1-m) & m & -(1-m) \end{bmatrix}.$$

ODE Notice in (8.4) that the diode voltage v_D does not depend implicitly on i_D ($M_{44} = 0$), so that we can easily solve the linear constraint $v_D = v_{C_1}$ (row 4). Furthermore, there is no load on output pin 4 so that the observer current vanishes ($i_{OUT} = 0$). Substituting the capacitor laws (see (8.2) in (8.4), we formulate the state-space ODE [7]

$$\begin{cases} \dot{\mathbf{x}} = -\mathbf{G_x}\nabla H(\mathbf{x}) - \boldsymbol{N}(\mathbf{x}) - \mathbf{G_u}\mathbf{u}, \\ \mathbf{y} = \mathbf{C}\nabla H(\mathbf{x}) + \mathbf{D}\mathbf{u}. \end{cases} \tag{8.5}$$

7. Here we removed the unobserved output variables, by consequence, the state space does not have the canonical form of a pH-ODE.

where

$$\mathbf{x} = \begin{bmatrix} q_{C1} \\ q_{C2} \end{bmatrix}, \quad \mathbf{G_u} = \begin{bmatrix} G_{12} \\ G_{13} \end{bmatrix}, \quad \mathbf{G_x} = \begin{bmatrix} G_{22} & G_{23} \\ G_{23} & G_{33} \end{bmatrix}, \quad \nabla H(\mathbf{x}) = \begin{bmatrix} \frac{1}{C_1} & 0 \\ 0 & \frac{1}{C_2} \end{bmatrix} \mathbf{x}, \quad \boldsymbol{N}(\mathbf{x}) = \begin{bmatrix} i_D\left(\frac{q_{C_1}}{C_1}\right) \\ 0 \end{bmatrix},$$

$$\mathbf{y} = v_{OUT}, \qquad \mathbf{u} = v_{IN}, \qquad \mathbf{C} = \begin{bmatrix} \alpha_{22} & \alpha_{23} \end{bmatrix}, \qquad \mathbf{D} = \alpha_{21}.$$

Discretisation by projection We consider the AVF discretisation. We use the averaged current variable $\bar{\mathbf{i}}_C = \left[\bar{i}_{C_1}, \bar{i}_{C_2}\right]^\mathsf{T}$ and the initial condition $\mathbf{x}_0 = \left[q^0_{C_1}, q^0_{C_2}\right]^\mathsf{T}$ to parametrize the trajectory

$$\boldsymbol{x}(\tau) = \mathbf{x}_0 + h \int_0^\tau \bar{\mathbf{i}}_C \, \mathrm{d}s.$$

By projection of (8.5) on the space of constant funtions, we obtain the algebraic equation on $\bar{\mathbf{i}}_C$

$$\boldsymbol{F}(\bar{\mathbf{i}}_C) = \bar{\mathbf{i}}_C + \mathbf{G_x}\overline{\nabla} H(\mathbf{x}_0, h\bar{\mathbf{i}}_C) + \overline{\boldsymbol{N}}(\mathbf{x}_0, h\bar{\mathbf{i}}_C) + \mathbf{G_u}\bar{\mathbf{u}} = 0, \tag{8.6}$$

where the AVF discrete gradient for linear capacitors is

$$\overline{\nabla} H(\mathbf{x}_0; \delta\mathbf{x}) = \begin{bmatrix} \frac{1}{C_1} & 0 \\ 0 & \frac{1}{C_2} \end{bmatrix} \left(\mathbf{x}_0 + \frac{1}{2}\delta\mathbf{x}\right),$$

where the averaged law of projected diodes is

$$\overline{\boldsymbol{N}}(\mathbf{x}_0; \delta\mathbf{x}) = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \overline{z}\left(\frac{q^0_{C_1}}{C_1}; \frac{\delta q_{C_1}}{C_1}\right), \quad \text{where} \quad \overline{z}\left(v_0; \delta v\right) = \begin{cases} \dfrac{Z\left(v_0 + \delta v\right) - Z(v_0)}{\delta v} & \delta v \neq 0, \\ z(v_0) & \delta v = 0. \end{cases}$$

and where the anti-parallel diode law z and its anti-derivative Z are given by

$$z(v) = 2I_S \sinh\left(\frac{v}{V_T}\right), \qquad Z(v) = 2I_S V_T \left(\cosh\left(\frac{v}{V_T}\right) - 1\right). \tag{8.7}$$

> **Remark 8.3.** We remind that projection is computed according to theorem (5.7) p.141 (see example 5.2). The quantity $\bar{z}$ plays the role of the *dissipative AVF discrete gradient* of the voltage potential Z. The average discrete gradient has been applied to dissipative potentials by the author in [MH18, (63)] where it is shown that the following closed-form expression holds
>
> $$\overline{z}(v_0; \delta v) = 2I_S \sinh\left(\frac{v + \frac{1}{2}\delta v}{V_T}\right) \operatorname{sinhc}\left(\frac{\delta v}{2V_T}\right), \text{ where } \operatorname{sinhc}(x) := \begin{cases} \sinh(x)/x & x \neq 0, \\ 1 & x = 0. \end{cases}.$$

Finally, the system (8.6) is solved using Newton iteration, where we use the result from equation (5.42) p.141 (also introduced in [MH18, (38)]) to compute the Jacobian of the AVF discrete gradients. Simulation results for varying values of the morph parameter are shown on figure 8.5.

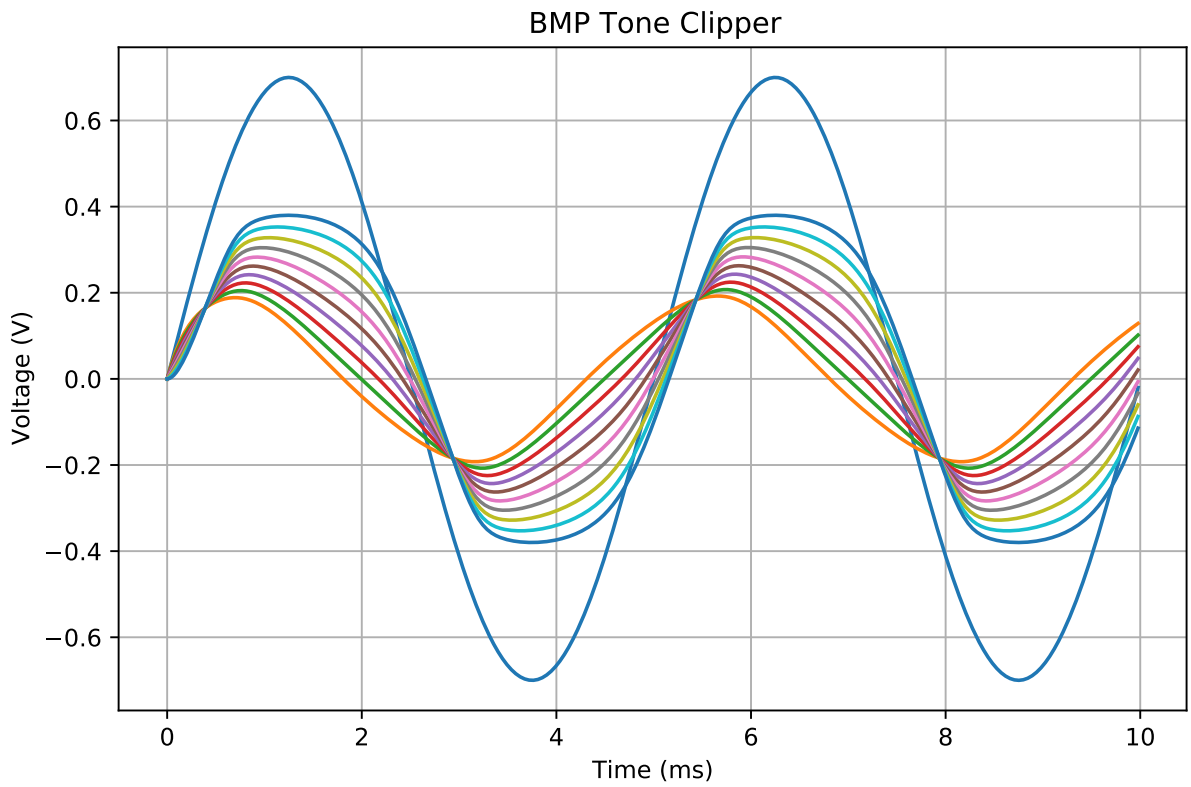


Figure 8.5 – (BMP Tone Clipper) Responses v_{OUT} (coloured curves) to a sinusoidal input v_{IN} in blue (amplitude 700 mV, fundamental frequency $f_0 = 200Hz$ and a sampling rate $f_s = 44.1$ kHz. Morph values are continuously selected for $m \in [0, 1]$.

As expected from the circuit design, in figure 8.5, the lowpass output (dark blue curve) is identical to that of a lowpass diode clipper circuit (i.e a damped saturated wave). As the morph potentiometer is moved in the opposite direction (orange curve), the waveform becomes progressively triangular (the diode limiting effect on voltage v_{C_1} in the lowpass branch, yields a quasi-constant current charge/discharge of capacitor C_2 on the highpass circuit side)

By consequence, the output waveform roughly changes from a damped saturated square (lowpass circuit branch) to a smooth triangular wave (high-pass circuit branch in orange) according to the morph potentiometer.

The interest of this circuit remains mostly pedagogical rather than practical [8]. It illustrates the design of new circuits from simpler subcircuits, and the (sometimes) unexpected consequences of unbuffered coupling. Indeed, "happy accidents" are not uncommon in the history of analog audio electronics (even more among guitarists). Nowadays, a popular branch of this trial and error approach to circuit design is commonly referred under the umbrella term of *circuit bending*.

8. In synthesizers, converters of sinusoidal waveforms to triangle and square waves use different and more complicated circuitry. See [EPPB17b, GEPP18] for more information about "west coast" waveshaping audio synthesis.

8.3 Tube Screamer drive stage

We consider the drive stage of the Tube Screamer (TS) guitar Pedal. The TS was manufactured by Ibanez in 1979 to emulate the saturation of tube amplifiers with solid-state circuitry. Notable users include Stevie Ray Vaughan, Carlos Santana and Steve Vai. This circuit is emblematic of the class of *overdrive* circuits (as opposed to *distortion* which is more agressive) and it can be found as a building block of many circuits (e.g. in the Boss OD-1, or in the feedback path of the Korg MS-20 Voltage-controlled filter shown in section 8.4). The main advantage of overdrive (compared to distortion) is that saturation applies to the difference $v_I - v_O$ instead of the direct signal v_I. This leads to a more subtle effect preserving the dynamics and expressivity of the input signal while enriching its harmonic content.

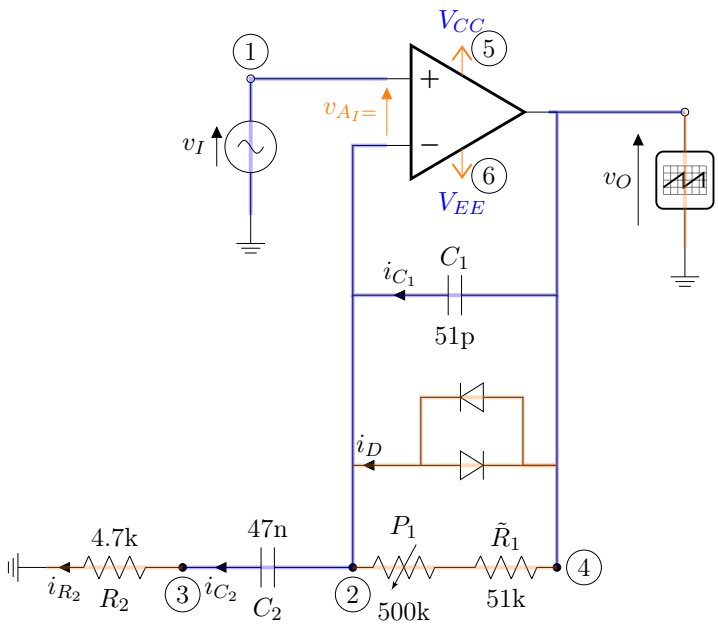


Figure 8.6 – (Tube screamer) Drive stage. In the original schematic, the virtual ground is set to $V_{bias} = 4.5V$, with $V_{CC} = 9$ V and $V_{EE} = 0$ grounded. For simplicity, we have chosen V_{bias} as the reference voltage and shifted V_{CC} and V_{EE} accordingly. Spanning tree T in blue.

Theory of operation Denote $R_1 = P_1 + \tilde{R}_1$. removing diodes and assuming that the OPA is in nullor mode, the circuit reduces to a non inverting amplifier with Laplace transfer function

$$H_{TS}(s) = 1 + \frac{Z_1(s)}{Z_2(s)} = 1 + \frac{R_1}{R_2}\underbrace{\left(\frac{1}{1+sR_1C_1}\right)}_{\text{low-pass}}\underbrace{\left(\frac{sR_2C_2}{1+sR_2C_2}\right)}_{\text{high-pass}}$$

where Z_1, Z_2 are respectively parallel and serial impedances corresponding to $R_1 \parallel C_1$ and R_2C_2. At high frequencies, $R_1 \parallel C_1$ act as a lowpass filter with cutoff frequency between 5.66 and 61.2 kHz, above which the gain reduces to unity. At low frequencies, R_2C_2 acts as a high-pass filter with cutoff frequency 720 Hz, below which the amplifier gain also reduces to unity. Between these two limits, the circuit behaves as a bandpass booster (see figure 8.7) where R_1 controls both the boost and the cutoff. Adding diodes to the circuit brings soft saturation and limits the voltage across diodes $v_D = v_O - v_I$ to approximately ± 700 mV. When diodes are conducting and the op amp is in nullor mode, the output voltage is approximately $v_O \approx v_I \pm 0.7$, so that the

effective gain also reduces to unity for large signals. For a typical guitar input signal (i.e. between 100 and 700 mV according to the type of pickups and playing intensity) and a 9 V battery as power supply, the headroom before the opamp enters saturation [9] is about 3 V. A more detailed analysis of the complete circuit can be found in [Ele20b].

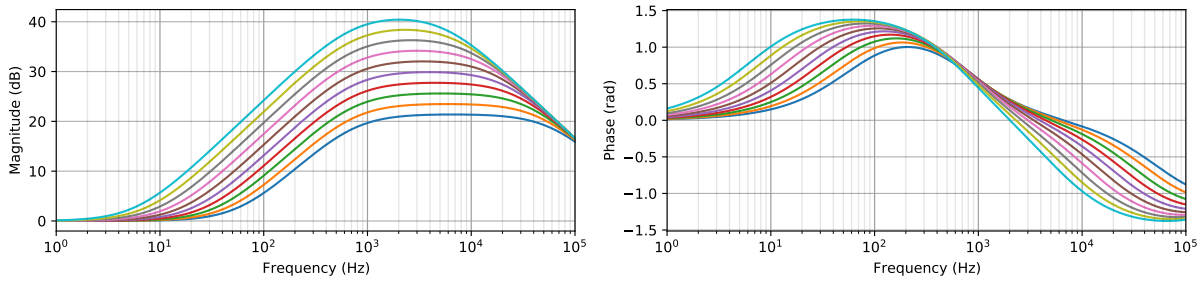


Figure 8.7 – (Tube screamer drive) Linearized frequency response for varying values of P_1.

Incidence matrix The incidence matrix of Tube screamer drive circuit shown on figure 8.6 is

$$\mathbf{A} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{array}{c} \begin{array}{ccccccccccccc} CC & EE & IN & C_1 & C_2 & A_O & R_1 & R_2 & D & A_I & A_{CC} & A_{EE} & OUT \end{array} \\ \begin{bmatrix} -1 & -1 & -1 & . & . & -1 & . & -1 & . & . & -1 & -1 & -1 \\ . & . & +1 & . & . & . & . & . & . & +1 & . & . & . \\ . & . & . & -1 & +1 & . & -1 & . & -1 & -1 & . & . & . \\ . & . & . & . & -1 & . & . & +1 & . & . & . & . & . \\ . & . & . & +1 & . & +1 & +1 & . & +1 & . & . & . & +1 \\ +1 & . & . & . & . & . & . & . & . & . & +1 & . & . \\ . & +1 & . & . & . & . & . & . & . & . & . & +1 & . \end{bmatrix} \end{array}.$$

Dirac structure We select the current-controlled spanning tree $T = \{CC, EE, IN, C_1, C_2, A_O\}$ with voltage-controlled co-tree $\overline{T} = \{R_1, R_2, D, A_I, A_{CC}, A_{EE}, OUT\}$ to obtain the following hybrid Dirac structure

$$\mathbf{J} = \begin{array}{c} \\ i_{CC} \\ i_{EE} \\ i_{IN} \\ i_{C_1} \\ i_{C_2} \\ i_{A_O} \\ v_{R_1} \\ v_{R_2} \\ v_D \\ v_{A_I} \\ v_{A_{CC}} \\ v_{A_{EE}} \\ v_{OUT} \end{array} \begin{array}{c} \begin{array}{ccccccccccccc} v_{CC} & v_{EE} & v_{IN} & v_{C_1} & v_{C_2} & v_{A_O} & i_{R_1} & i_{R_2} & i_D & i_{A_I} & i_{A_{CC}} & i_{A_{EE}} & i_{OUT} \end{array} \\ \begin{bmatrix} . & . & . & . & . & . & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ . & . & . & . & . & . & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ . & . & . & . & . & . & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ . & . & . & . & . & . & -1 & +1 & -1 & -1 & 0 & 0 & 0 \\ . & . & . & . & . & . & 0 & +1 & 0 & 0 & 0 & 0 & 0 \\ . & . & . & . & . & . & 0 & -1 & 0 & +1 & 0 & 0 & -1 \\ 0 & 0 & 0 & +1 & 0 & 0 & . & . & . & . & . & . & . \\ 0 & 0 & 0 & -1 & -1 & +1 & . & . & . & . & . & . & . \\ 0 & 0 & 0 & +1 & 0 & 0 & . & . & . & . & . & . & . \\ 0 & 0 & +1 & +1 & 0 & -1 & . & . & . & . & . & . & . \\ +1 & 0 & 0 & 0 & 0 & 0 & . & . & . & . & . & . & . \\ 0 & +1 & 0 & 0 & 0 & 0 & . & . & . & . & . & . & . \\ 0 & 0 & 0 & 0 & 0 & 1 & . & . & . & . & . & . & . \end{bmatrix} \end{array}.$$

9. For completeness, op amp clipping is handled in the simulation code. However, op amp clipping is too far from standard behaviour, so that it is not pertinent to show on simulation results.

Reduced linear resistive structure Reducing the resistive branches $\{R_1, R_2\}$ yields the linear dissipative structure

$$\mathbf{M} = \begin{array}{c} \\ i_{CC} \\ i_{EE} \\ i_{IN} \\ i_{C_1} \\ i_{C_2} \\ i_{A_O} \\ v_D \\ v_{A_I} \\ v_{A_{CC}} \\ v_{A_{EE}} \\ v_{OUT} \end{array} \begin{array}{c} \begin{array}{ccccccccccc} v_{CC} & v_{EE} & v_{IN} & v_{C_1} & v_{C_2} & v_{A_O} & i_D & i_{A_I} & i_{A_{CC}} & i_{A_{EE}} & i_{OUT} \end{array} \\ \begin{bmatrix} . & . & . & . & . & . & 0 & 0 & -1 & 0 & 0 \\ . & . & . & . & . & . & 0 & 0 & 0 & -1 & 0 \\ . & . & . & . & . & . & 0 & -1 & 0 & 0 & 0 \\ . & . & . & -G_{12} & -G_2 & G_2 & -1 & -1 & 0 & 0 & 0 \\ . & . & . & -G_2 & -G_2 & G_2 & 0 & 0 & 0 & 0 & 0 \\ . & . & . & G_2 & G_2 & -G_2 & 0 & +1 & 0 & 0 & -1 \\ 0 & 0 & 0 & +1 & 0 & 0 & . & . & . & . & . \\ 0 & 0 & +1 & +1 & 0 & -1 & . & . & . & . & . \\ +1 & 0 & 0 & 0 & 0 & 0 & . & . & . & . & . \\ 0 & +1 & 0 & 0 & 0 & 0 & . & . & . & . & . \\ 0 & 0 & 0 & 0 & 0 & +1 & . & . & . & . & . \end{bmatrix} \end{array},$$

with conductances $G_1 = 1/R_2$, $G_{12} = \frac{R_1+R_2}{R_1 R_2}$.

Reduced DAE and ODE To solve the system, we remove variables corresponding to trivial constraints in matrix $\mathbf{M}$. It is enough to consider the implicit DAE defined by the following submatrix of $\mathbf{M}$ (all other variables of the system can be retrieved from $i_{C_1}, i_{C_2}, v_{A_I}, v_D$ using $\mathbf{M}$ and component laws).

$$\mathbf{M}_r = \begin{array}{c} \\ i_{C_1} \\ i_{C_2} \\ v_{A_I} \\ v_D \end{array} \begin{array}{c} \begin{array}{ccccc} v_{IN} & v_{C_1} & v_{C_2} & v_{A_O} & i_D \end{array} \\ \begin{bmatrix} 0 & -G_{12} & -G_2 & G_2 & -1 \\ 0 & -G_2 & -G_2 & G_2 & 0 \\ +1 & +1 & 0 & -1 & . \\ 0 & +1 & 0 & 0 & . \end{bmatrix} \end{array}.$$

To handle the OPA, we have to consider the third row of $\mathbf{M}_r$ with special care:

- *Nullor mode* (see subsection 7.2.1 p.190): we have $v_{A_I}(\lambda) = 0$. This yields the linear constraint $v_{A_O} = v_{IN} + v_{C_1}$
- *Saturation mode*: we have the constraint $v_{A_I}(\lambda) = v_I + v_{C_1} - v_{A_O}(\lambda)$. Furthermore, if $v_{A_I} > 0$ then $v_{A_0} = v_{CC}$ and if $v_{A_I} < 0$, then $v_{A_O} = v_{EE}$.

Here, we unify both modes by solving for v_{A_O} and introduce the function

$$v_{A_O}(v) := \begin{cases} v_{EE} & v < v_{EE}, \\ v & v \in [v_{EE}, v_{CC}], \\ v_{CC} & v > v_{CC}. \end{cases} \tag{8.8}$$

Substituting capacitor and diode laws ($v_C = q/C$ and $i_D(\cdot) = z(\cdot)$ from (8.7)), in the first two rows of $\mathbf{M}_r$, and using $v_{A_I} = v_{IN} + v_{C_1}$ (row 3), we finally obtain the reduced ODE

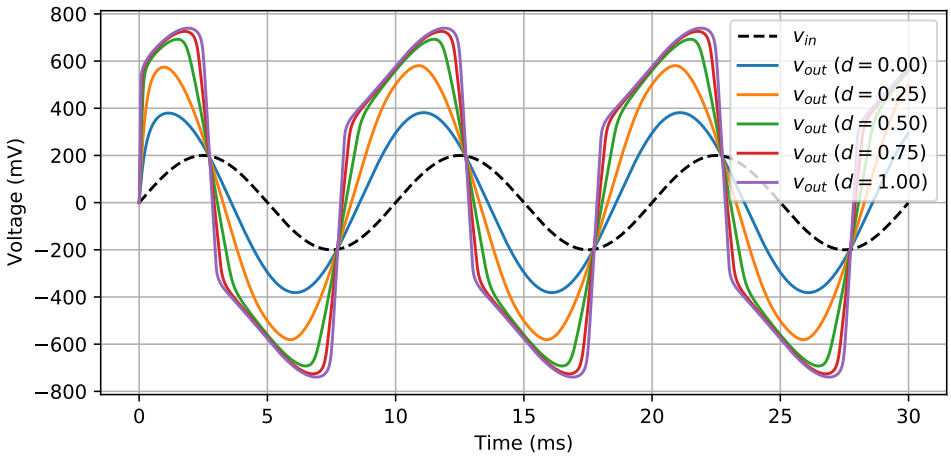
$$\begin{bmatrix} \dot{q}_1 \\ \dot{q}_2 \end{bmatrix} = \begin{bmatrix} -G_{12} & -G_2 \\ -G_2 & -G_2 \end{bmatrix} \begin{bmatrix} q_1/C_1 \\ q_2/C_2 \end{bmatrix} + \begin{bmatrix} G_2 \\ G_2 \end{bmatrix} v_{A_O}\left(v_{IN} + \frac{q_1}{C_1}\right) - \begin{bmatrix} 1 \\ 0 \end{bmatrix} i_D\left(\frac{q_1}{C_1}\right) \tag{8.9}$$

Discretization Using the Average Vector Field discretisation method with $q_1(\tau) = q_1^0 + \tau\delta q_1$, $q_2(\tau) = q_2^0 + \tau\delta q_2$ yields the algebraic equation $\boldsymbol{F}(\delta\mathbf{x}) = 0$ for the variables $\delta\mathbf{x} = (\delta q_1, \delta q_2)$ where

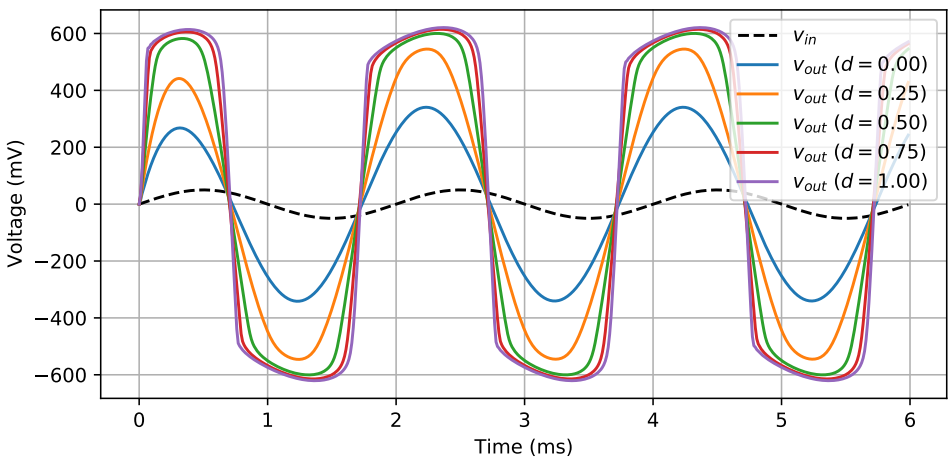
$$\boldsymbol{F}(\delta\mathbf{x}) = \begin{bmatrix}\delta q_1\\ \delta q_2\end{bmatrix} - h\left(\begin{bmatrix}-G_{12} & -G_2\\ -G_2 & -G_2\end{bmatrix}\begin{bmatrix}\bar{q}_1/C_1\\ \bar{q}_2/C_2\end{bmatrix} + \begin{bmatrix}G_2\\ G_2\end{bmatrix}\bar{v}_{A_O}\left(v_{IN} + \frac{q_1}{C_1}\right) - \begin{bmatrix}1\\ 0\end{bmatrix}\bar{i}_D\left(\frac{q_1}{C_1}\right)\right),$$

where $\bar{v}_{A_0}(v) = \int_0^1 v_{A_0}(v(\tau))\,\mathrm{d}\tau$, $\bar{i}_D(v) = \int_0^1 v_{A_0}(v(\tau))\,\mathrm{d}\tau$ and $\bar{q}_C = \int_0^1 q_C(\tau)\,\mathrm{d}\tau = q_C^0 + \delta q_C/2$. denote the average vector field projection of component laws in the right hand side of (8.9). As for other examples, the system $\boldsymbol{F}(\delta\mathbf{x}) = 0$ is solved using Newton method.

Simulation results for a sampling rate of $f_s = 44.1$ kHz are shown on figure 8.8. For simulated examples, convergence is reached after 1 to 3 iterationés (1.52 on average) for absolute and relative Newton errors respectively of 10 μV and a 10^{-10}. Note that exhaustive energy and power plots are not reproduced for each example for brievety (see [MH18, fig. 2 and 4], reproduced in appendix G p.323, for similar plots, see also figure 5.13 p.152).



(a) $f_0 = 100$ Hz, $G = 200$ mV

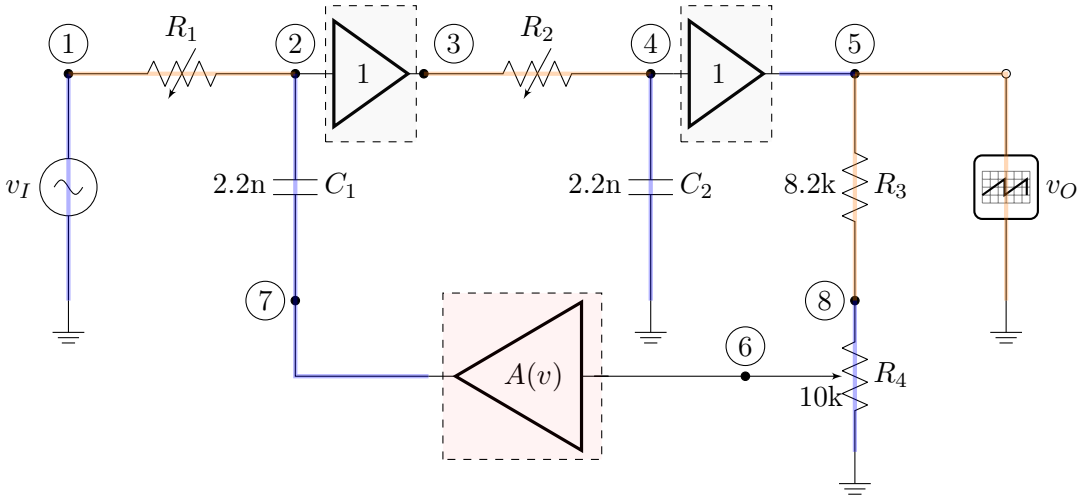


(b) $f_0 = 500$ Hz, $G = 50$ mV

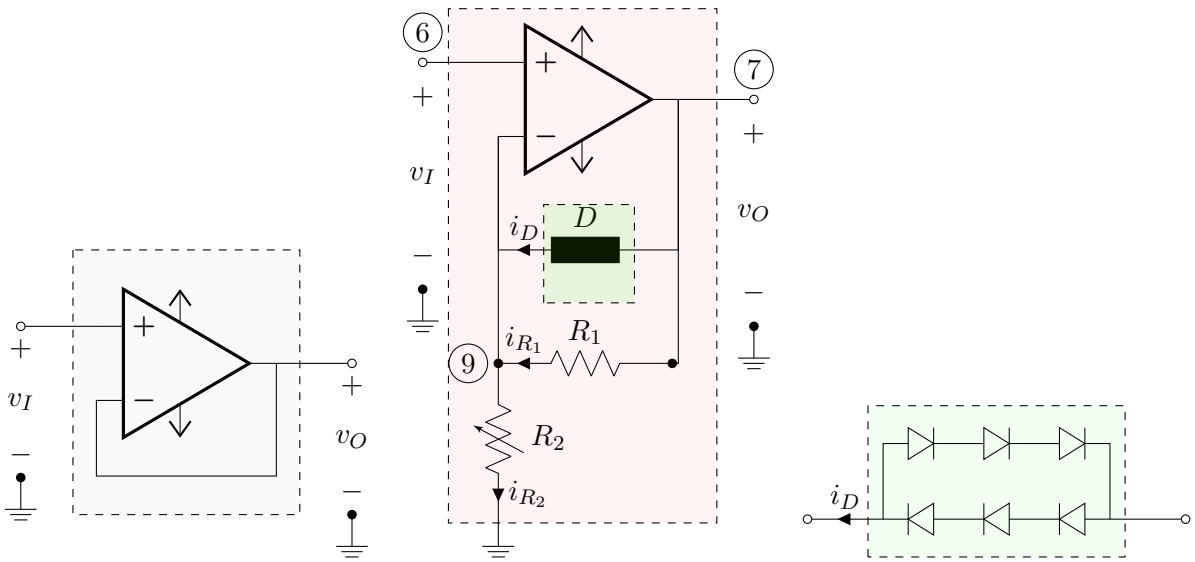
Figure 8.8 – (Tube screamer drive) simulation for the series resistance R_1 exponentially distributed in $[51, 551]$ kΩ according to the drive parameter $d \in [0, 1]$. The input signal is a sinusoid with frequency f_0 and amplitude G simulated at $f_s = 44.1$ kHz.

8.4 Korg MS-20 Filter

The filter of the Korg MS-20 synthesizer is (with the Moog filter) one of the most famous synthesizer filter. It has been studied in the references [Sti06, Pir13]. This filter is closely related to the Sallen–Key filter from section 7.1.4 with the following differences: the lowpass filter stages are buffered from each other by (ideal) voltage followers [10]; the feedback path contains a nonlinear overdrive amplifier (see figure 8.9b) and a voltage divider to control the resonance of the filter. Resistors R_3 and R_4 (voltage divider) controls the feedback gain of the filter. Furthermore, the nonlinear amplifier also features a calibration gain. The combination of both gains with nonlinearities allows the filter to reach self-oscillation.



(a) Overall filter schematic.



(b) Voltage buffer $v_O = v_I$ **(c)** Overdrive amplifier $v_O = A(v_I)$ **(d)** Clipping diodes (element D)

Figure 8.9 – (MS-20 filter) Simplified overall schematic (a) and its sub-components (b-d). In (a), the chosen spanning tree T is shown in blue and its complimentary cotree $\overline{T}$ in orange.

10. in this example, contrary to the Sallen-Key example of section 7.1.4, we assume that the power supply voltages are large enough to not enter saturation.

8.4.1 Overdrive amplifier

The overdrive amplifier (figure 8.9c-d) is a non-inverting amplifier (as in section 7.1.4) with negative feedback diodes to limit the voltage difference between inputs and outputs. This situation is similar to the TubeScreamer saturation in section 8.3 but without capacitor filtering.

Algebraic modelling The stage is composed of resistors, diodes and OPA, all considered memoryless. To avoid solving such a stiff system iteratively, we choose to pre-solve this sub-circuit as an equivalent algebraic component. We assume that the power supply voltages are large enough to maintain the OPA in nullor mode. Applying nodal analysis at node 9 (see figure 8.9c) yields $i_{R_2} = i_D + i_{R_1}$. This leads to to the voltage equation $v_I/R_2 = (v_O - v_I)/R_1 + i_D(v_O - v_I)$, that we reformulate as an implicit equation on the output voltage

$$v_O = \left(1 + \frac{R_1}{R_2}\right) v_I - R_1 i_D(v_O - v_I), \tag{8.10}$$

where the clipping diodes law i_D is given by

$$i_D(v) = I \sinh\left(\frac{v}{V}\right) \qquad \text{with} \qquad I = 2I_S, \qquad V = 3V_T. \tag{8.11}$$

Analysis for small and large signals: For small signals ($i_D \approx 0$), diodes are not conducting, so that the non-inverting amplifier is governed by $v_O \approx G v_I$ with $G = 1 + \frac{R_1}{R_2}$. For $R_1 = 10\text{k}$ and $R_2 = 2200(1 + (1 - \kappa))$ with $\kappa \in [0, 1]$, the small signal gain of the amplifier belongs to $[3.2, 5.54]$. Conversely, as soon as diodes conducts (large signals), the signal is soft-clipped. Assuming that we know an explicit mapping $v_O = A(v_I)$, (lumping the power supply ports [11]), we can replace the circuit by the nonlinear amplifier two-port defined by

$$\left\{(v_I, v_O, i_I, i_O) \in \mathbb{R}^2 \times \mathbb{R}^2 \mid i_I = 0,\ v_O = A(v_I)\right\}.$$

Explicit formulation and approximation Using the implicit function theorem, one can prove that there exists a unique function $A : v_I \mapsto v_O = A(v_I)$ solution of (8.10) that can be tabulated (see figure 8.10). Going further, we look for a closed-form approximation of A. To this end, we invert the hyperbolic sine in (8.11) to obtain the equivalent formulation of (8.10)

$$v_O = v_I + V \operatorname{asinh}\left(\frac{G v_I - v_O}{R_1 I}\right).$$

This form suggests a candidate approximation model $A(v) \approx g(v; \alpha, \beta, \gamma)$ parametrized by (α, β, γ)

$$g(v; \alpha, \beta, \gamma) = v + \alpha \operatorname{sign}(\mathrm{v}) \left(\operatorname{asinh}\left(\beta |v|^\gamma\right)\right)^{1/\gamma}. \tag{8.12}$$

A very accurate approximation can be obtained from (8.12) (see figure 8.10). For $V = 3 \cdot 26$ mV A, $I = 2$ fA, nonlinear least squares optimisation yields the optimal parameters

κ	α	β	γ
0	1.7	1.33	14.56
0.5	1.7	1.78	14.28
1.0	1.69	2.69	13.71

.

11. If required it is still possible to recover the power supply currents to express the power balance using the OPA model from subsection 7.2.1, but we do not detail this further.

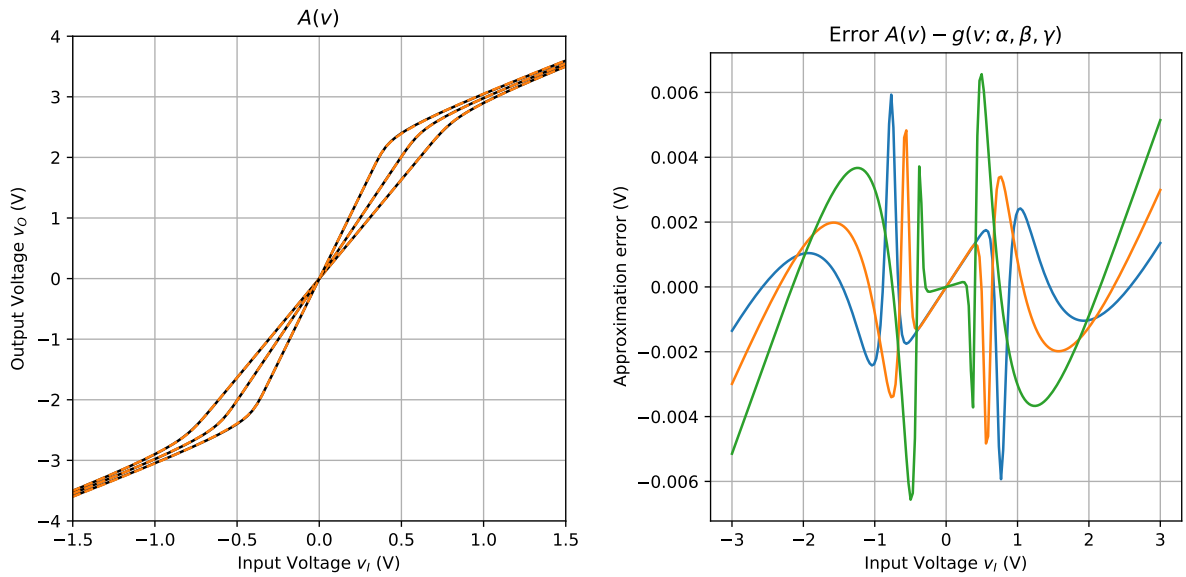


Figure 8.10 – Explicit overdrive amplifier mapping and its approximation for $\kappa \in [0, 1]$. Exact relation $A(v)$ in black, and its approximation $g(v)$ in dashed orange.

8.4.2 Filter

For this circuit, thanks to buffering, it is simpler to use Nodal analysis (at nodes 2, 4) to directly obtain the ODE: using Kirchhoff laws, we have $i_{C_1} = i_{R_1}$ and $i_{C_2} = i_{R_2}$ and using the node voltages $e_2 = v_{AO} + v_{C_1}$, $e_3 = v_2$, $v_5 = v_4$ one gets

$$i_{C_1} = \frac{v_I - v_{AO} - v_{C1}}{R_1}, \qquad i_{C_2} = \frac{v_{AO} + v_{C_1} - v_{C2}}{R_2}.$$

The nonlinear state space system is obtained using (i) the amplifier law $v_{AO} = A(kv_{C_2})$ where $k = \frac{\rho R_4}{R_3+R_4} \in [0, 0.55]$ corresponds to the voltage divider, (ii) introducing co-energy variables $x_1 := v_{C_1}$, $x_2 := v_{C_2}$ for the linear capacitor law $i_C = C\dot{v}_C$ and (iii) defining the cutoff pulsation $\omega_c := 1/(RC)$ for equal resistances $R_1 = R_2 = R$, and capacitances $C_1 = C_2 = C$.

$$\frac{1}{\omega_c}\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\end{bmatrix} = \begin{bmatrix}-1 & 0\\ 1 & -1\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix} + \begin{bmatrix}-1\\ 1\end{bmatrix} A(kx_2) + \begin{bmatrix}1\\ 0\end{bmatrix} v_I, \qquad \text{where} \qquad k = \frac{\rho R_4}{R_3 + R_4}, \tag{8.13a}$$

$$v_O = x_2. \tag{8.13b}$$

Small signals analysis For small signals, we have the linear approximation $A(v) \approx Kv$ with overall feedback gain $K = Gk$ (remind that $G = 1 + \frac{R_1}{R_2}$). Then equation (8.13a) simplifies to

$$\frac{1}{\omega_c}\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\end{bmatrix} \approx \begin{bmatrix}-1 & -K\\ 1 & K-1\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix} + \begin{bmatrix}1\\ 0\end{bmatrix} v_I. \tag{8.14}$$

From this linearized state-space system, we can obtain the following Laplace transfer function, which corresponds to a resonant lowpass filter with Q-factor $Q = \frac{1}{2-K}$ and cutoff pulsation ω_c (see section 7.1.4 for more details on resonant lowpass filters and their frequency response).

$$H_{\text{MS20}}(s) = \frac{1}{\left(\frac{s}{\omega_c}\right)^2 + (2-K)\left(\frac{s}{\omega_c}\right) + 1}. \tag{8.15}$$

As noted in [Pir13] and in contrast with the unbuffered case from section 7.1.4, the filter reaches infinite Q (i.e self-oscillation) for $K = 2$ instead of $K = 3$. Furthermore, according to circuit parametrisation, the maximum feedback gain (for small signals) belongs to $[1.76, 3.05]$ for $\rho = 1$, $\kappa \in [0, 1]$, which is enough to reach self-oscillations.

Large signals analysis For large signals, the output of the overdrive amplifier can be approximated by $v_{AO} \approx \pm 2.1 + v_I$ so that the direct gain of the circuit is bounded by $K = \rho < 0.55$. Despite the absence of rail-to-rail hard clipping as in section 8.3, the clipping diodes are still strong enough to stabilise the system.

Discretization To simulate this filter, we use the Average Vector Field discretization. Projection of equations (8.13a), (8.13b) for affine state trajectories of the form $v(\tau) = v^0 + \tau\delta v$ with $\tau \in [0, 1]$ yields the discrete state space

$$\begin{bmatrix} \delta x_1 \\ \delta x_2 \end{bmatrix} = \omega_d \left(\begin{bmatrix} -1 & 0 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} x_1^0 + \frac{1}{2}\delta x_1 \\ x_2^0 + \frac{1}{2}\delta x_2 \end{bmatrix} + \begin{bmatrix} -1 \\ 1 \end{bmatrix} \overline{A}(kv_2^0, k\delta v_2) + \begin{bmatrix} 1 \\ 0 \end{bmatrix} \bar{v}_I \right), \tag{8.16a}$$

$$v_O = x_2 + \frac{1}{2}\delta x. \tag{8.16b}$$

where the normalised pulsation is $\omega_d = h\omega_c$ and $\overline{A}(v_0, \delta v) := \langle 1, A(v_0 + \tau\delta v) \rangle$ denotes the AVF projection of the feedback nonlinearity. The algebraic system (8.16a) is rewritten as $\boldsymbol{F}(\delta\mathbf{x}) = 0$ with $\delta\mathbf{x} = [\delta x_1, \delta x_2]^\mathsf{T}$ and solved using Newton iteration, where the Jacobian of $\boldsymbol{F}$ is

$$\boldsymbol{F}'(\delta\mathbf{x}) = \mathbf{I} - \omega_d \left(\frac{1}{2} \begin{bmatrix} -1 & 0 \\ 1 & -1 \end{bmatrix} + k \begin{bmatrix} 0 & -1 \\ 0 & 1 \end{bmatrix} \frac{\partial}{\partial \delta x_2} \overline{A}(kx_2^0, k\delta x_2) \right).$$

Simulation results Simulation results are displayed below on figures 8.11, 8.12 and 8.13 for square and saw inputs at various amplitudes to exhibit the nonlinear behaviour of this filter.

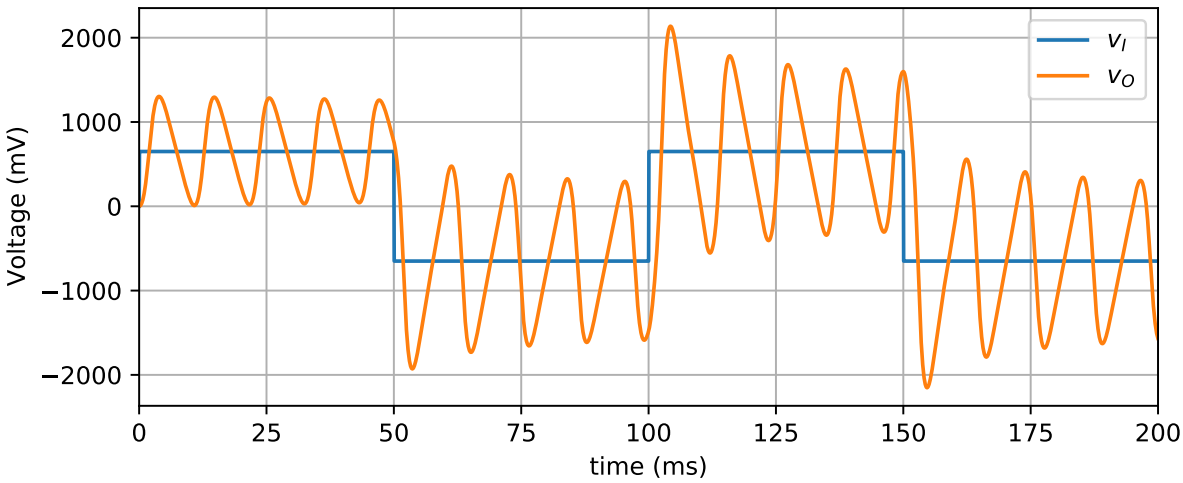


Figure 8.11 – MS-20 filter response to a square wave input with peak volage 650 mV, for a cutoff frequency of 100 Hz and a resonance $k = 0.9$, $\kappa = 1$. The nonlinear self-oscillation is clearly visible, with an asymmetrical waveshape modulated by the square wave input signal.

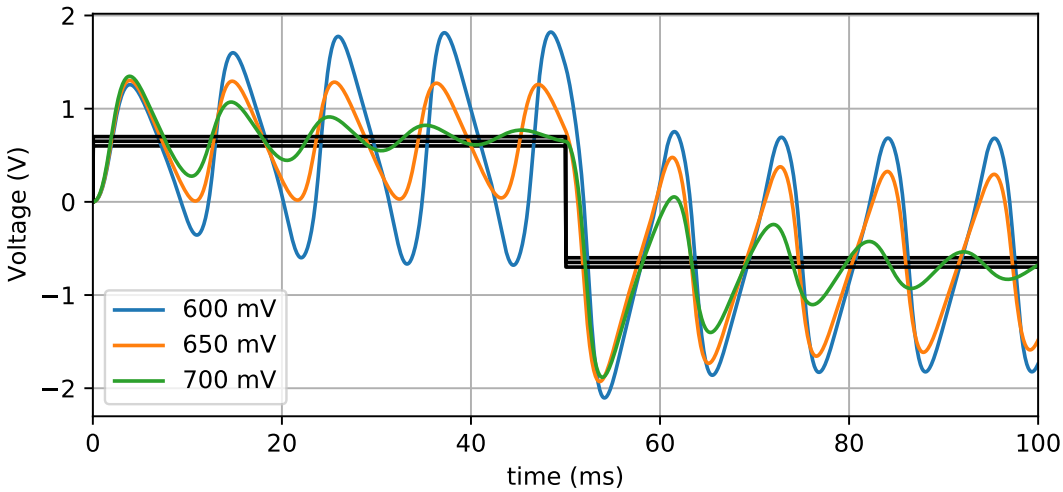


Figure 8.12 – (MS-20 filter) same simulation as figure 8.11 with input levels $\{600, 650, 700\}$ mV. We observe that the input amplitude influences the amplitude of self-oscillation, its frequency, its damping and its shape. The higher the input, the higher the damping. The lower the oscillation amplitude, the higher the resonance frequency.

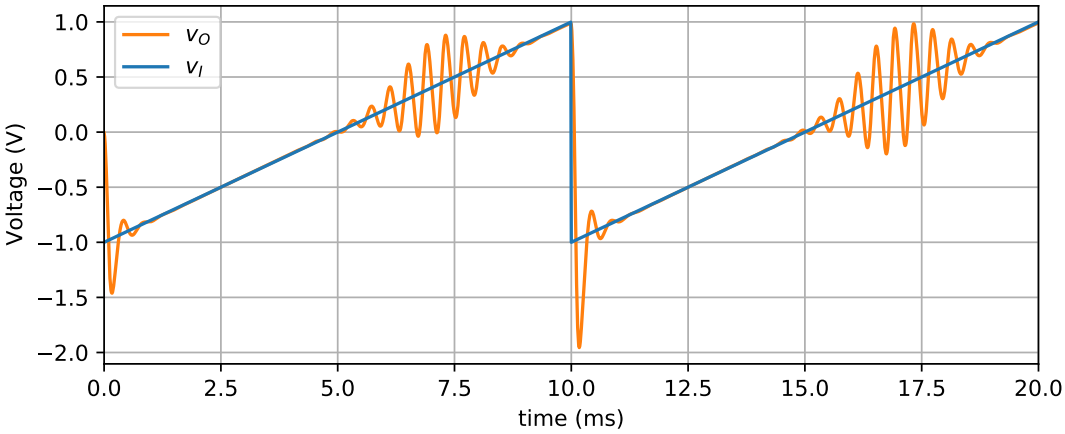


Figure 8.13 – (MS-20) response to a 1V sawtooth signal with fundamental frequency $f_0 = 100$ Hz. The cutoff frequency is set to 2.5 kHz for a resonance $k = 0.68$. Bursts of self-oscillation in the middle of the ramp are typical of this filter and allowed by the temporarily lower input level.

Simulation results are consistent with SPICE simulation and measurements. The expected behaviour of this filter and its salient features are reproduced. Note that comparing results with the ones of the Sallen–Key filter in figure 7.13 p.187, we observe that small topological changes (buffering stages and a nonlinear feedback path) yield significant modifications to the behaviour of this filter (an thus to its sonic character). Important differences are: (i) filter oscillations are saw-like rather than sinusoidal (fig. 8.11), (ii) the behaviour is more progressive according to input level (fig. 8.12), (iii) self-oscillation can happen near zero-crossings (fig. 8.13).

8.5 FitzHugh–Nagumo relaxation oscillator

In this section, we consider the electronic realisation of a FitzHugh–Nagumo (FHN) (see [Fit55, NAY62]) relaxation oscillator. The FitzHugh–Nagumo model was originally proposed by FitzHugh as modification of the Van der Pol system to model neurons. It uses a cubic nonlinearity with negative incremental resistance to achieve self-excitation. The electronic circuit realisation of fig. 8.14 was proposed by Nagumo and uses a tunnel diode (see ex. 1.9 p.31) to implement a nonlinearity with negative incremental resistance. In music, FHN oscillators have been used for sound synthesis purposes in [Col08, SBM] and for beat/tempo synchronisation in [Eck02, AOI07].

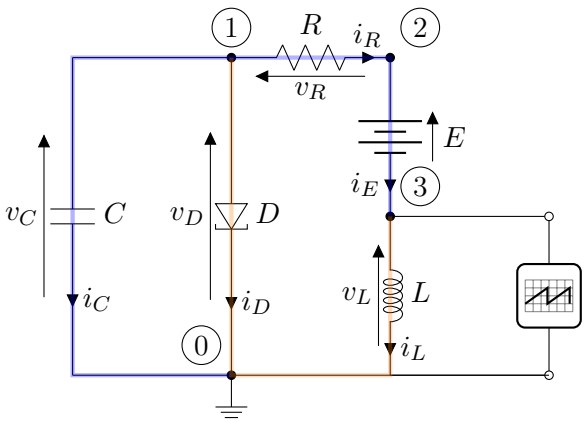


Figure 8.14 – Electronic realisation of a Fitzhugh–Nagumo relaxation oscillator.

Theory of operation At static operating point, the capacitor C can be considered as an open circuit and the inductor L as a short-circuit. The tunnel diode D is biased by the combination of the voltage source E and resistor R by the load line $v_D = E - Ri_D$ (for $v_L = \dot{\phi} = 0$ and $i_R = -i_D$). It can exhibit astable, monostable or bi-stable behaviour according to the choice of E and R (see [RCA63, p.36-44] and figure 8.15). The inductor controls the slow dynamics by modulating the bias point current. This roughly determines the period of relaxation oscillations. The capacitor acts as a stiffness controller by smoothing the fast jumps occuring when the trajectory is in the unstable negative incremental resistance region. Indeed, in the limit $C \to 0$, the diode becomes current-controlled by the inductor. Its characteristics is current-controlled and multi-valued (see fig. 8.15), but only the positive incremental resistance points are stable points of the system.

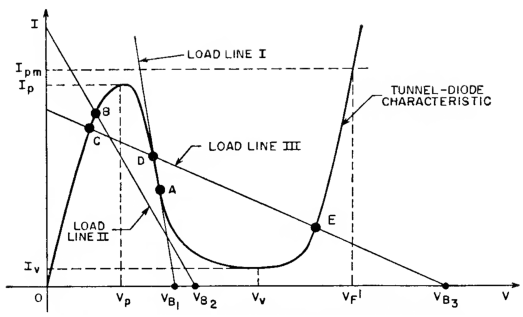


Figure 8.15 – Different biasing scenarios for a tunnel diode multivibrator. Figure extracted from the RCA tunnel diode manual [RCA63].

Incidence matrix The incidence matrix of the graph corresponding to the FHN schematic is

$$\mathbf{A} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{ccccc} E & C & R & D & L \end{array} \\ \begin{bmatrix} -1 & . & . & . & -1 \\ 1 & . & 1 & . & . \\ . & 1 & -1 & 1 & . \\ . & -1 & . & -1 & 1 \end{bmatrix} \end{array}.$$

Dirac structure From the incidence matrix $\mathbf{A}$, we select the current-controlled spanning tree $T = \{R, C, E\}$, with voltage-controlled cotree $\overline{T} = \{D, L\}$, to obtain the hybrid Dirac structure

$$\mathbf{J} = \begin{array}{c} \\ v_D \\ v_L \\ i_R \\ i_C \\ i_E \end{array} \begin{array}{c} \begin{array}{ccccc} i_D & i_L & v_R & v_C & v_E \end{array} \\ \begin{bmatrix} . & . & 0 & 1 & 0 \\ . & . & -1 & 1 & -1 \\ 0 & 1 & . & . & . \\ -1 & -1 & . & . & . \\ 0 & 1 & . & . & . \end{bmatrix} \end{array}.$$

Reduced Linear resistive structure Eliminating the linear resistor branch R and solving the trival constraint $v_D = v_C$ yields the linear dissipative structure

$$\mathbf{M} = \begin{array}{c} \\ v_D \\ v_L \\ i_C \\ i_E \end{array} \begin{array}{c} \begin{array}{cccc} i_D & i_L & v_C & v_E \end{array} \\ \begin{bmatrix} . & . & 1 & 0 \\ . & -R & 1 & -1 \\ -1 & -1 & . & . \\ 0 & 1 & . & . \end{bmatrix} \end{array} \longrightarrow \tilde{\mathbf{M}} = \begin{array}{c} \\ v_L \\ i_C \\ i_E \end{array} \begin{array}{c} \begin{array}{cccc} i_D(v_C) & i_L & v_C & v_E \end{array} \\ \begin{bmatrix} . & -R & 1 & -1 \\ -1 & -1 & . & . \\ 0 & 1 & . & . \end{bmatrix} \end{array}.$$

pH-ODE Finally substituting the laws of the components, one obtains the dissipative pH-ODE

$$\dot{\mathbf{x}} = -\boldsymbol{r}(\mathbf{x}) + \mathbf{J}\nabla H(\mathbf{x}) + \mathbf{G}v_E, \tag{8.17a}$$

$$i_E = -\mathbf{G}^{\mathsf{T}}\nabla H(\mathbf{x}). \tag{8.17b}$$

where the state $\mathbf{x}$, skew-symmetric matrix $\mathbf{J}$, Hamiltonian H, resistive function $\boldsymbol{r}$ and port matrix $\mathbf{G}$ are given by

$$\mathbf{x} = \begin{bmatrix} \phi \\ q \end{bmatrix}, \quad \mathbf{J} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \quad H(\mathbf{x}) = \frac{1}{2}\left(\frac{\phi^2}{L} + \frac{q^2}{C}\right), \quad \boldsymbol{r}(\mathbf{x}) = \begin{bmatrix} R\nabla_\phi H(\mathbf{x}) \\ z\left(\nabla_q H(\mathbf{x})\right) \end{bmatrix}, \quad \mathbf{G} = \begin{bmatrix} -1 \\ 0 \end{bmatrix}.$$

We use as default values $E = 400$ mV, $R = 20\ \Omega$, $C = 10\ \mu$F, $L = 300$ mH. For the tunnel diode, we use the model of the tunnel diode from example 1.9 p.31

$$z(v) = I_S\left(\exp\left(\frac{v}{V_T}\right) - 1\right) + I_P\left(\frac{v}{V_P}\right)\exp\left(-\left(\frac{v - V_P}{V_P}\right)\right),$$

with parameters $I_S = 1$ fA, $V_T = \frac{kT}{q_e} \approx 26$ mV, $I_P = 4.7$ mA, $V_P = 100$ mV.

Simulation The system is solved using AVF projection and Newton iteration. Simulation results are shown on figure 8.16 with time series corresponding to different values of the bias voltages E and the capacitor C. Phase plots are shown on figure 8.17.

The q-nullcline[12] ($\dot{q} = 0$) corresponding to the tunnel diode is an attractor for the slow dynamics. When its intersection with the ϕ-nullcline ($\dot{\phi} = 0$) happens in the negative incremental resistance region, the equilibrium point is unstable, leading to a limit cycle. On the contrary, when the intersection happens in the region of positive incremental resistance, the equilibrium point is stable and all trajectories converge to it (red trajectory).

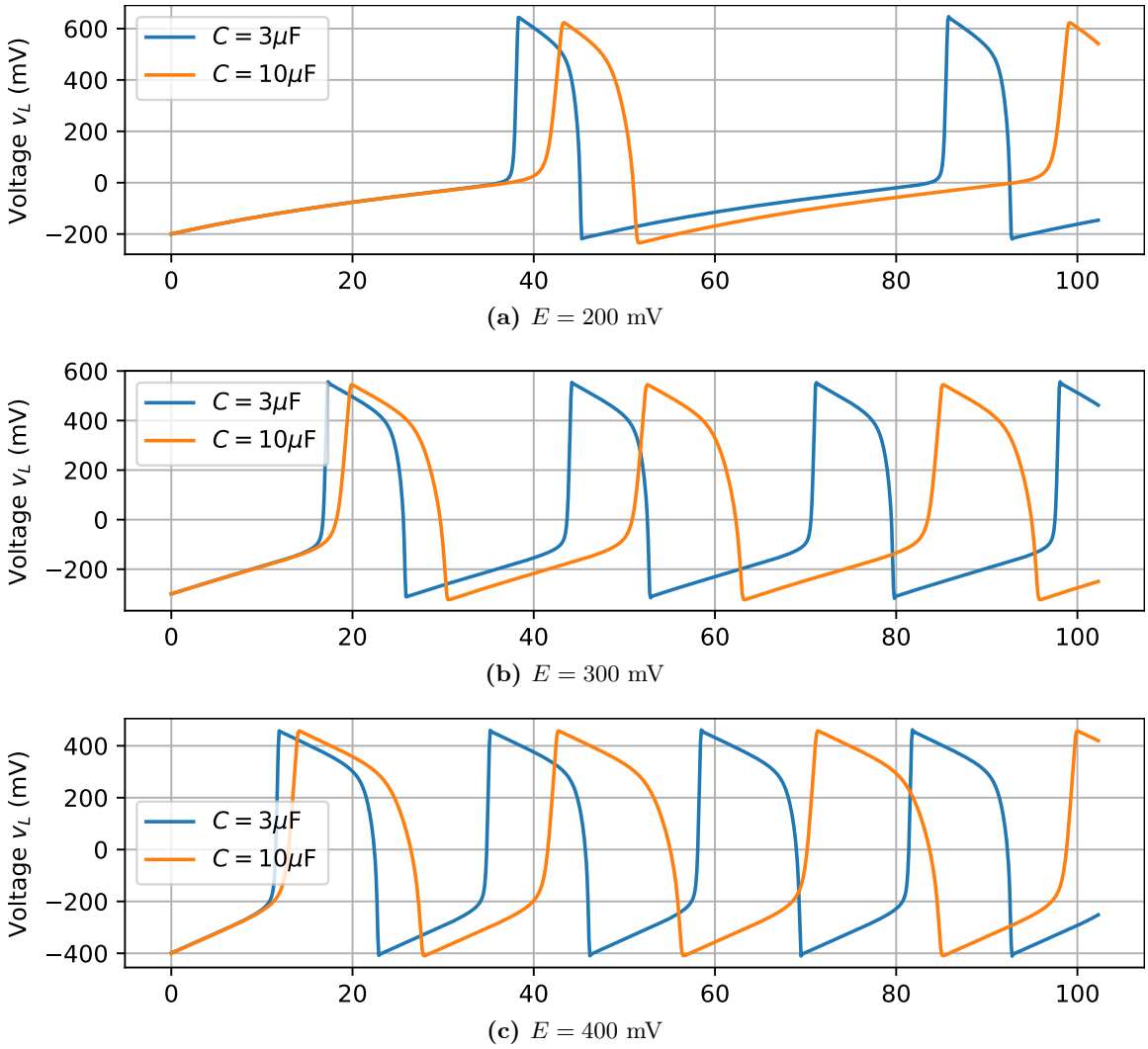


Figure 8.16 – FitzHugh–Nagumo relaxation oscillations, varying values of the offset voltage $E \in \{200, 300, 400\}$ mV and capacitance $C \in \{3, 10\}$ μF.

In figure 8.16, the frequency of relaxations increases with the bias voltages E while the period increases with higher values of capacitance C. The smoothing effect of the capacitance is noticeable by reducing the slope of the relaxation. Time is displayed in milliseconds.

12. For a system of ODE $\dot{\mathbf{x}} = f(\mathbf{x})$, the i-th nullcline is the geometric shape such that $\dot{\mathbf{x}}_i = 0$. The equilibrium points of the system are located where all of the nullclines intersect (i.e. $\dot{\mathbf{x}} = 0$).

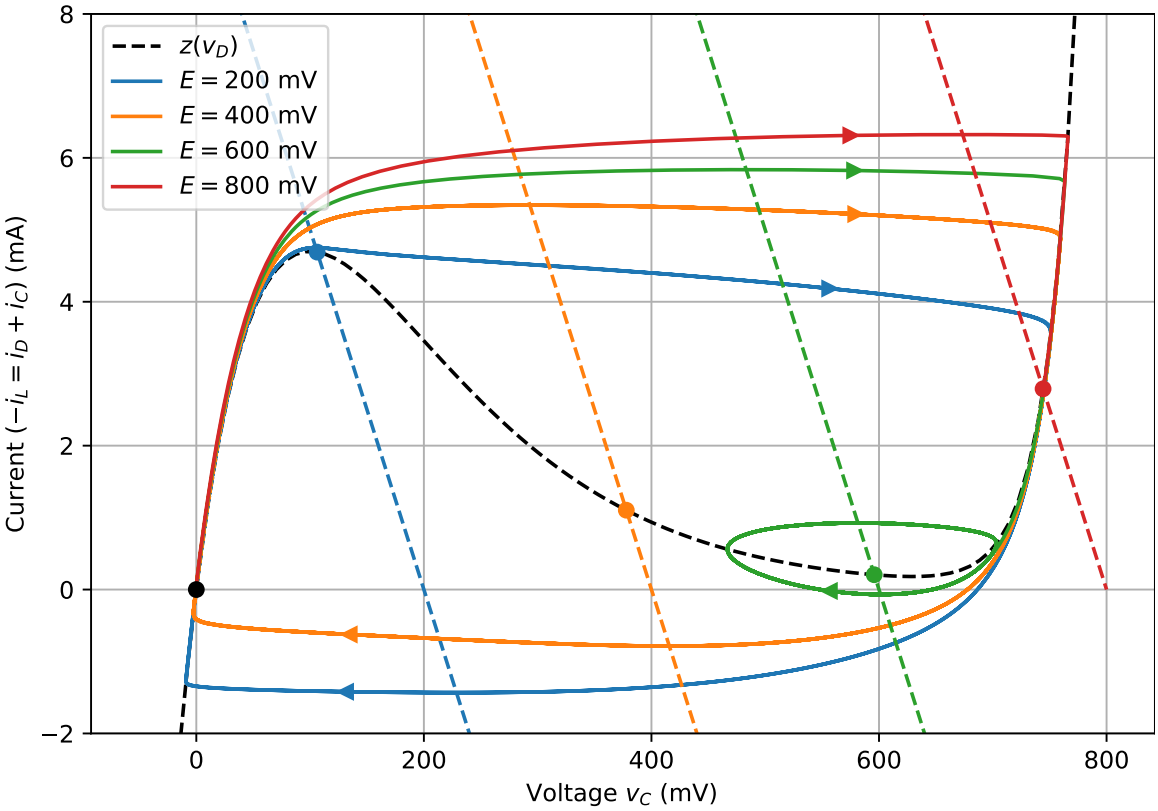


Figure 8.17 – (FitzHugh–Nagumo) Phase plot, for varying values of E.

In figure 8.17, the simulated orbits trajectories are displayed in the $(v_C, -i_L)$-space of co-energy variables. Inductor nullclines are shown as load lines corresponding to each bias point (dashed colored curves, $\dot{\phi} = 0 \iff v_C = E + Ri_L$). Conversely, the capacitor nullcline ($\dot{q} = 0 \iff i_L = -z(v_C)$ and $v_C = v_D$) corresponds to the tunnel diode characteristic. It is an attractor for the slow dynamic (dashed black). Note that the red curve converges to a stable equilibrium point (positive incremental resistance) at the intersection of the (dashed red) load line and the (dashed black) tunnel diode characteristic (red dot) while other curves converge to limit cycles about unstable equilibrium points (blue, orange and green dots).

8.6 Passive peaking equalizer (beyond the Nyquist frequency)

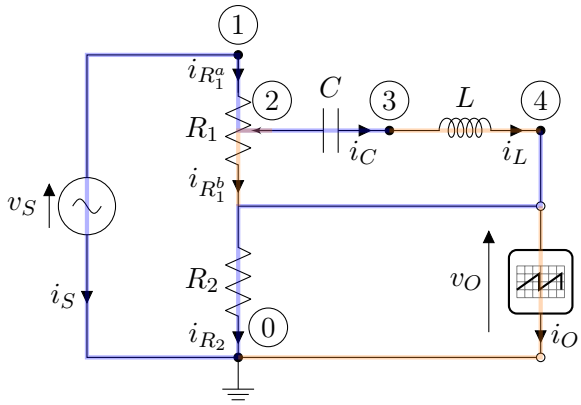


Figure 8.18 – (Passive Peaking EQ). Spanning tree T in blue, cotree $\overline{T}$ in orange.

We consider a passive peaking equaliser circuit (the only linear example in this chapter) to study the effect of high order RPM methods on frequency warping and spectral accuracy for open systems. Indeed, *in the linear case*, the stability function for projection order $p = 1$ is identical to the mid-point and bilinear ones (sharing the same numerical dispersion).

Reminder on the bilinear method Artefacts of the bilinear method on the frequency response of systems are well known. Let $H_a(s)$ denote the Laplace transfer function of a continuous-time system, its discrete-time approximation $H_d(z)$ is obtain by substituting s by

$$\tilde{s}(z) = \frac{\Delta}{M} = \frac{2}{h}\frac{z-1}{z+1}, \qquad \text{where} \qquad \Delta = \frac{z-1}{h}, \qquad \text{and} \qquad M = \frac{z+1}{2} \tag{8.18}$$

in $H_a(s)$ so that $H_d(z) := H_a(\tilde{s}(z))$. Operators Δ and M are finite differences approximation of the time derivative and identity centred at $h/2$ (to compensate the time shift induced by Δ) where $z = e^{hs}$ denotes the Laplace transform of the positive time-shift operator (see [Bil09, p.35]). Substituting $z = e^{hs}$ in (8.18) one can show that bilinear discretization acts as the mapping [13]

$$\frac{h\tilde{s}}{2} = \tanh\left(\frac{hs}{2}\right), \qquad \text{so that} \qquad \frac{h\tilde{\omega}}{2} = \tan\left(\frac{h\omega}{2}\right) \qquad \text{for} \qquad s = i\omega. \tag{8.19}$$

The principal value of this mapping warps the frequency axis $\tilde{s} \in i\mathbb{R}$ to the range $s \in i(-h\pi, h\pi)$ severely distorting the frequency response at high frequencies (see fig. 8.19b and D.4 p.299).

> **Remark 8.4.** To link the AVF/RPM(1,0) method with the bilinear scheme, note that, for an affine trajectory $x(t) = x_0 + (t/h)(x_1 - x_0)$, M is the discrete equivalent of the average vector field projection $\bar{x} = (x_1 + x_0)/2$ and Δ of the average slope $\dot{\bar{x}} = (x_1 - x_0)/h$.

Goals To challenge high-order RPM schemes (def. 5.1 p.122), we consider the case where the peaking equalizer has a resonance frequency *beyond the Nyquist frequency*. This situation is in fact common in electronic audio circuits: several analog equalisers use a peaking EQ between 20 kHz and 100 kHz with a large bandwidth to implement high frequency boost (instead of a shelving filter). Note that for audio use, we are not interested in the frequency response above 20 khz (beyond human hearing). Nevertheless, the action of a 50 kHz resonance on input signals below 20 kHz is significant (see fig. 8.19a) and should be faithfully reproduced.

13. see also the frequency warping graphs shown in figure D.2 p.298 for several values of projection order p.

Theory of operation The potentiometer is parametrised by $\gamma \in [0,1]$ according to the law $R_1^a = (1-\gamma)R_1$, $R_1^b = \gamma R_1$. When the potentiometer R_1 is down ($\gamma = 0$ $R_1^a = R_1$, $R_1^b = 0$), the RLC network is short-circuited ($v_2 = v_4$) so that the remaining circuit is a simple voltage divider with static gain $a_0 = \frac{R_2}{R_1+R_2}$. When γ is increased, the RLC network acts as a bandpass filter whose contribution is added to the output to yield a peaking EQ. Its resonance frequency is controlled by L, C and its bandwidth by R_2.

Incidence matrix The incidence matrix of the circuit shown on figure 8.18 is given by

$$\tilde{\mathbf{A}} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccccccc} S & C & R_1^a & R_2 & R_1^b & L & O \end{array} \\ \begin{bmatrix} -1 & . & . & -1 & . & . & -1 \\ +1 & . & +1 & . & . & . & . \\ . & +1 & -1 & . & +1 & . & . \\ . & -1 & . & . & . & +1 & . \\ . & . & . & +1 & -1 & -1 & +1 \end{bmatrix} \end{array}.$$

Dirac structure We select the current-controlled tree $T = \{S, C, R_1^a, R_2\}$ with cotree $\overline{T} = \{R_1^b, L, O\}$ to obtain the Dirac structure

$$\mathbf{J} = \begin{array}{c} \\ i_S \\ i_C \\ i_{R_1^a} \\ i_{R_2} \\ v_{R_1^b} \\ v_L \\ v_O \end{array} \begin{array}{c} \begin{array}{ccccccc} v_S & v_C & v_{R_1^a} & v_{R_2} & i_{R_1^b} & i_L & i_O \end{array} \\ \begin{bmatrix} . & . & . & . & -1 & -1 & 0 \\ . & . & . & . & 0 & +1 & 0 \\ . & . & . & . & +1 & +1 & 0 \\ . & . & . & . & +1 & +1 & -1 \\ +1 & 0 & -1 & -1 & . & . & . \\ +1 & -1 & -1 & -1 & . & . & . \\ 0 & 0 & 0 & +1 & . & . & . \end{bmatrix} \end{array}.$$

Reduced linear resistive structure Reducing linear resistive branches $\left\{R_1^a, R_2, R_1^b\right\}$, with the potentiometer relation $R_1^a = (1-\gamma)R_1$, $R_1^b = \gamma R_1$, yields the resistive structure

$$\mathbf{M} = \begin{array}{c} \\ i_S \\ i_C \\ v_L \\ v_O \end{array} \begin{array}{c} \begin{array}{cccc} v_S & v_C & i_L & i_O \end{array} \\ \begin{bmatrix} -G_{11} & 0 & -\alpha_{11} & -\alpha_{12} \\ 0 & 0 & +1 & 0 \\ \alpha_{11} & -1 & -R_{11} & R_{12} \\ \alpha_{12} & 0 & R_{12} & -R_{22} \end{bmatrix} \end{array}. \tag{8.20}$$

where

$$G_{11} = \frac{1}{R_1 + R_2}, \qquad \alpha_{11} = \gamma G_{11} R_1, \qquad \alpha_{12} = G_{11} R_2,$$
$$R_{22} = \frac{R_1 R_2}{R_1 + R_2}, \qquad R_{12} = \gamma R_{22}, \qquad R_{11} = \left(1 + \frac{R_1}{R_2}(1-\gamma)\right) R_{12}.$$

pH-ODE and state-space formulations The pH-ODE is built from (8.20) by (i) choosing the state $\boldsymbol{x} = [q, \phi]^\mathsf{T}$, input $\boldsymbol{u} = [v_S, i_O]$, and output $\boldsymbol{y} = [i_S, v_O]$, (ii) substituting component laws $v_C = q/C$, $i_C = \dot{q}$, $i_L = \phi/L$, $v_L = \dot{\phi}$ in (8.20) with energy $H(q, \phi) = \frac{q^2}{2C} + \frac{\phi^2}{2L}$. In practice, we use an open circuit load $i_O = 0$ (reduced input $u = v_S$) and neglect i_S (reduced output $y = v_O$). Then (8.20) can be formulated as a pH-ODE (left) with reduced state-space system (right)

$$\dot{\boldsymbol{x}} = (\mathbf{J_x} - \mathbf{R_x})\mathbf{Q}\boldsymbol{x} + \mathbf{G}\boldsymbol{u}, \qquad \rightarrow \qquad \dot{\boldsymbol{x}} = \mathbf{A}\boldsymbol{x} + \mathbf{B}u, \tag{8.21a}$$

$$\boldsymbol{y} = -\mathbf{G}^\mathsf{T}\mathbf{Q}\boldsymbol{x} + (\mathbf{J_u} - \mathbf{R_u})\boldsymbol{u}, \qquad \rightarrow \qquad y = \mathbf{C}\boldsymbol{x} + \mathbf{D}u. \tag{8.21b}$$

where systems matrices are respectively

$$\mathbf{J_x} - \mathbf{R_x} = \begin{bmatrix} 0 & 1 \\ -1 & -R_{11} \end{bmatrix}, \quad \mathbf{Q} = \begin{bmatrix} 1/C & 0 \\ 0 & 1/L \end{bmatrix}, \quad \mathbf{G} = \begin{bmatrix} 0 & 0 \\ \alpha_{11} & R_{12} \end{bmatrix}, \quad \mathbf{J_u} - \mathbf{R_u} = \begin{bmatrix} -G_{11} & -\alpha_{12} \\ \alpha_{12} & -R_{22} \end{bmatrix},$$

$$\mathbf{A} = (\mathbf{J_x} - \mathbf{R_x})\mathbf{Q}, \quad \mathbf{B} = \begin{bmatrix} 0 \\ \alpha_{11}, \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} 0 & R_{12}/L \end{bmatrix}, \quad \mathbf{D} = \begin{bmatrix} \alpha_{12} \end{bmatrix}.$$

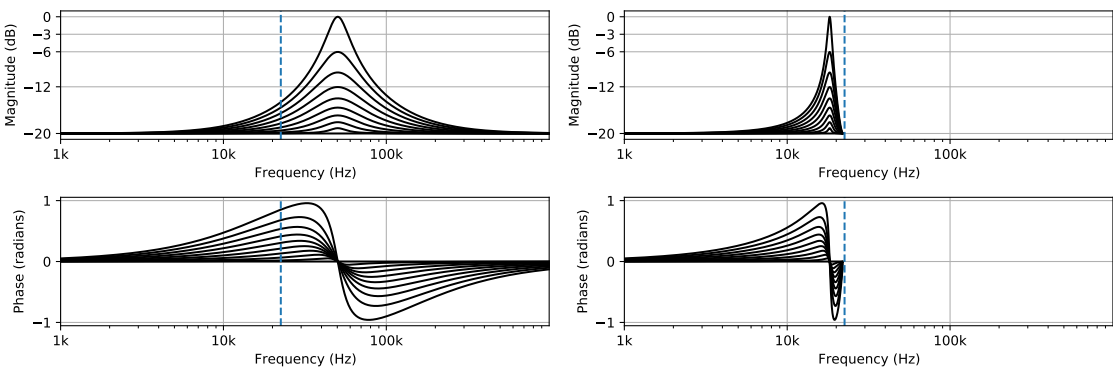
Laplace transfer function Computing the Laplace transfer function using the formula $H_{\mathsf{EQ}}(s) = \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D}$ yields the standard form of a peaking equalizer

$$H_{\mathsf{EQ}}(s) = \left(\frac{R_2}{R_1 + R_2}\right) \cdot \frac{LCs^2 + \gamma R_1 C s + 1}{LCs^2 + \gamma R_1 C \left(\frac{(1-\gamma)R_1 + R_2}{R_1 + R_2}\right) s + 1} = K \cdot \frac{\left(\frac{s}{\omega_0}\right)^2 + B(\gamma)\left(\frac{s}{\omega_0}\right) + 1}{\left(\frac{s}{\omega_0}\right)^2 + \frac{B(\gamma)}{G(\gamma)}\left(\frac{s}{\omega_0}\right) + 1},$$

where the direct gain K, pulsation ω_0, damping/bandwidth B and resonance gain G are

$$K = \frac{R_2}{R_1 + R_2}, \qquad \omega_0 = \frac{1}{\sqrt{LC}}, \qquad B(\gamma) = \gamma R_1 \sqrt{\frac{C}{L}}, \qquad G(\gamma) = \frac{R_1 + R_2}{(1-\gamma)R_1 + R_2}. \tag{8.22}$$

Note that this peaking EQ is neither constant-Q nor exactly proportional-Q (see [Boh88]). The quality factor $Q = 1/B$ is modulated by γ, so that the higher the boost, the larger the bandwidth.



(a) Analog response **(b)** AVF/Mid-point approximation

Figure 8.19 – (Peaking EQ) Frequency Response for $L = 10$ mH, $C = 1$ nF, $R_1 = 9k$, $R_2 = 1k$, $\gamma \in [0, 1]$. This yields $f_0 \approx 50$ kHz and $Q(1) \approx 2.8$. (a) continuous-time response, (b) warped frequency response of second order mid-point/AVF discretisation (see eq.(8.19)) for a sampling rate $f_s = 44.1$ kHz (Nyquist frequency $f_s/2$ in dashed blue). The main drawback is that the resonance peak is warped by several kHz into the audible frequency band. Note that the frequency response is also periodised above the Nyquist frequency by sampling, but is not shown here.

8.6.1 High-order RPM discretisation of a linear state-space system

Definitions To discretize the state-space (8.21a)-(8.21b), we use [14] RPM(p,0). Denote [15]

$$\vec{\mathbf{x}} := [\langle P_i \,|\, \dot{\boldsymbol{x}}\rangle]_{i=0}^{p-1}, \qquad \vec{\mathbf{u}} := [\langle P_i \,|\, u\rangle]_{i=0}^{p-1}, \qquad \vec{\mathbf{y}} := [\langle P_i \,|\, y\rangle]_{i=0}^{p-1}, \tag{8.23a}$$

the projection coefficients of functions $\dot{\boldsymbol{x}}(\tau), u(\tau), y(\tau)$ in the Legendre basis over a unit time step $\Omega = (0,1)$. Using the Kronecker product (see appendix D.10.1) and $n \times n$ identity $\mathbf{I}_n$, denote by

$$\vec{\mathbf{A}} = \mathbf{I}_p \otimes \mathbf{A}, \qquad \vec{\mathbf{B}} = \mathbf{I}_p \otimes \mathbf{B}, \qquad \vec{\mathbf{C}} = \mathbf{I}_p \otimes \mathbf{C}, \qquad \vec{\mathbf{D}} = \mathbf{I}_p \otimes \mathbf{D}, \qquad \vec{\mathbf{I}} = \mathbf{I}_p \otimes \mathbf{I}_n, \tag{8.23b}$$

expanded state-space and identity matrices. Moreover, denote respectively

$$\vec{\mathbf{1}} = \mathbf{e}_0 \otimes \mathbf{I}_n \text{ with } \mathbf{e}_0 = \left[\langle P_i \,|\, 1\rangle\right]_{i=0}^{p-1}, \quad \text{and} \quad \vec{\mathbf{V}} = \mathbf{V}_p \otimes \mathbf{I}_n \text{ with } \mathbf{V}_p = \left[\langle P_i \,|\, \int_0^\tau P_j\rangle\right]_{i,j=0}^{p-1}, \tag{8.23c}$$

the matrix representation of the constant function $|\,1\rangle$ and the operational matrix of integration ($\mathbf{V}_p \equiv \int_0^\tau$ extended to $\mathbb{R}^n$) (see (C.17) p.286). Introduce the discrete integration operator

$$\vec{\mathcal{V}} : (\mathbf{x}_0, \vec{\mathbf{x}}) \mapsto \vec{\mathbf{x}} = \vec{\mathbf{1}}\mathbf{x}_0 + h\vec{\mathbf{V}}\vec{\mathbf{x}}. \tag{8.23d}$$

Projected state-space Using these notations, Legendre projection of the continuous-time state-space system (8.21a)-(8.21b) yields the projected linear system of *algebraic* equations

$$\begin{cases} \vec{\mathbf{x}} = \vec{\mathbf{A}}\vec{\mathbf{x}} + \vec{\mathbf{B}}\vec{\mathbf{u}}, \\ \vec{\mathbf{y}} = \vec{\mathbf{C}}\vec{\mathbf{x}} + \vec{\mathbf{D}}\vec{\mathbf{u}} \end{cases}, \qquad \text{where} \qquad \vec{\mathbf{x}} = \vec{\mathcal{V}}(\mathbf{x}_0, \vec{\mathbf{x}}). \tag{8.24}$$

Explicit solution Solving (8.24) for $\vec{\mathbf{x}}$, yields the coefficients of the projected vector field

$$\vec{\mathbf{x}} = \left(\vec{\mathbf{I}} - h\vec{\mathbf{A}}\vec{\mathbf{V}}\right)^{-1}\left(\vec{\mathbf{A}}\vec{\mathbf{V}}\vec{\mathbf{1}}\mathbf{x}_0 + \vec{\mathbf{B}}\vec{\mathbf{u}}\right), \tag{8.25}$$

where $\vec{\mathbf{A}}\vec{\mathbf{V}} = \mathbf{V}_p \otimes \mathbf{A}$ (by properties of kronecker products, see appendix D.10.1). As the state increment $\mathbf{x}_1 - \mathbf{x}_0$ is proportional to the average ($\int_0^1$) of the vector field $\dot{\boldsymbol{x}}$, projecting (8.25) on $\langle 1\,|$, (equivalent to the transposed matrix $(\vec{\mathbf{1}})^\mathsf{T}$), we deduce the discrete time-stepping scheme

$$\mathbf{x}_1 = \mathbf{x}_0 + h(\vec{\mathbf{1}})^\mathsf{T}\left(\vec{\mathbf{I}} - h\vec{\mathbf{A}}\vec{\mathbf{V}}\right)^{-1}\left(\vec{\mathbf{A}}\vec{\mathbf{V}}\vec{\mathbf{1}}\mathbf{x}_0 + \vec{\mathbf{B}}\vec{\mathbf{u}}\right). \tag{8.26}$$

From (8.23d)-(8.25), we get the explicit input to output map (in term of Legendre coefficients)

$$\vec{\mathcal{H}}_{\mathbf{x}_0,h} : \vec{\mathbf{u}} \mapsto \vec{\mathbf{y}} = \vec{\mathbf{C}}\left(\vec{\mathbf{1}}\mathbf{x}_0 + h\vec{\mathbf{V}}\left(\vec{\mathbf{I}} - h\vec{\mathbf{A}}\vec{\mathbf{V}}\right)^{-1}\left(\vec{\mathbf{A}}\vec{\mathbf{V}}\vec{\mathbf{1}}\mathbf{x}_0 + \vec{\mathbf{B}}\vec{\mathbf{u}}\right)\right) + \vec{\mathbf{D}}\vec{\mathbf{u}}. \tag{8.27}$$

Remark 8.5. The Jacobian of the mapping (8.27) with respect to $\vec{\mathbf{u}}$ is

$$\vec{\mathbf{C}}(h\vec{\mathbf{V}}(\vec{\mathbf{I}} - h\vec{\mathbf{A}}\vec{\mathbf{V}})^{-1}\vec{\mathbf{B}}) + \vec{\mathbf{D}} \tag{8.28}$$

This is analog to $H(s) = \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D} = \mathbf{C}\frac{1}{s}(\mathbf{I} - \mathbf{A}\frac{1}{s})^{-1}\mathbf{B} + \mathbf{D}$, the Laplace transfer function of a state space system. Note that in (8.28) the operational matrix of integration $h\vec{\mathbf{V}}$ plays the role of the Laplace integration operator $\frac{1}{s}$.

14. See def.5.1 p.122. For simplicity, we only consider regularity order $k = 0$
15. Note that $\dot{\vec{\mathbf{x}}}$ denotes the coefficients of the projected vector field, here the dot is a label, not an operator.

8.6.2 Frequency response and Legendre filterbank interpretation

We want to study the quality of the RPM(p,0) high-order projection scheme (8.27) on the continuous-time frequency response. First, we establish the continuous-time system corresponding to the discrete-time one (8.27). Second, we derive its Laplace transfer function [16].

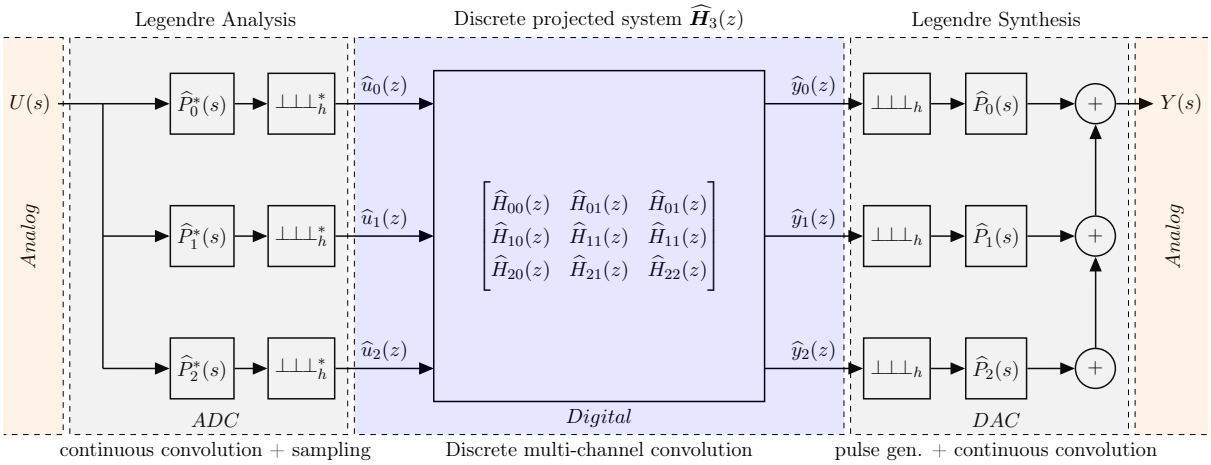


Figure 8.20 – Interpretation of RPM(p=3,0) as a mixed Legendre filterbank. We remind that the discrete Z-domain is embedded into the continuous Laplace domain through $z = e^s$.

Legendre filterbank interpretation Step 1) The discrete-time system (8.27) governs the discrete-time Legendre coefficients mapping $\vec{\mathbf{u}}[n] \mapsto \vec{\mathbf{y}}[n]$ (blue block on fig. 8.20). In this step, we formalize its Z-domain matrix transfer function $\widehat{\boldsymbol{H}}_p(z)$. Step 2) Legendre coefficients result from a frame-synchronous analysis/projection process $u(t) \mapsto \vec{\mathbf{u}}[n]$. This can be reformulated as convolution with the mirrored impulse responses $P_k(-t)$ followed by sampling (gray block). The continuous-time output results from the dual synthesis process, $\vec{\mathbf{y}}[n] \mapsto y(t)$ (reversing the order of operations): impulse synthesis followed by convolution with Legendre poynomials $P_k(t)$ (figs. C.1 C.2 p.287). In this step, we obtain their Laplace transfer function. Step 3) The complete system (analysis, discrete system, synthesis) can be represented by the cascade in figure 8.20. In this step, we obtain its frequency response $Y(s)$ for a zero order hold input $U(s)$.

Step 1: Z-domain transfer function To obtain the Z-domain transfer function of the projected state-space ($\widehat{\boldsymbol{H}}_p(z)$ in the middle of the filterbank), we propose the following result

Proposition 8.1 (Z-transform of Legendre projected state-space)**.** *Consider the continuous state-space system (8.21a)-(8.21b) discretised by RPM(p,0), according to (8.23a)-(8.27). Then, the Z-domain transfer function $\widehat{\boldsymbol{H}}_p$, of dimension $p \times p$, satisfying $\widehat{\boldsymbol{y}}(z) = \widehat{\boldsymbol{H}}_p(z)\widehat{\boldsymbol{u}}(z)$, is*

$$\widehat{\boldsymbol{H}}_p(z) = \left(\left(\frac{\mathbf{e}_0\mathbf{e}_0^\mathsf{T}}{z-1} + \mathbf{V}_p\right) \otimes \mathbf{C}\right)\left(\vec{\mathbf{I}} - \left(\frac{\mathbf{e}_0\mathbf{e}_0^\mathsf{T}}{z-1} + \mathbf{V}_p\right) \otimes \mathbf{A}\right)^{-1} \vec{\mathbf{B}} + \vec{\mathbf{D}}. \tag{8.29}$$

The proof of this proposition is detailed in appendix D.10 p.305.

16. Under the condition that input signals already belong to projection space. Note that the continuous-time system of figure 8.20 is not shift-invariant hence its Laplace transfer function is not defined in general.

Step 2: Laplace transform of Legendre operator For continuous-time analysis and synthesis, we need the Laplace transform of the Legendre polynomials restricted to $\tau \in (0,1)$.

A) Unrestricted transfer functions: We first introduce the one-side Laplace transform of shifted orthonormal Legendre polynomials extended to $[0,\infty)$

$$\widehat{B}_k(s) := \int_0^\infty e^{-\tau s} P_k(\tau)\,\mathrm{d}\tau. \tag{8.30}$$

Symbolic computation up to degree 3 yields

k	0	1	2	3
$\widehat{B}_k(s)$	$\frac{1}{s}$	$\frac{\sqrt{3}(2-s)}{s^2}$	$\frac{\sqrt{5}\left(12-6s+s^2\right)}{s^3}$	$\frac{\sqrt{7}\left(120-60s+12s^2-s^3\right)}{s^4}$

(8.31)

Remark 8.6 (Legendre polynomials and Padé approximations of the exponential)**.** The numerators of $\widehat{B}_k(s)$ are proportional to the denominators of the (k,k) Padé [a] approximation of the exponential (see [Ehl69]) while numerators of $\widehat{B}_k(-s)$ corresponds to the Padé numerators so that

$$\mathrm{Pade}_{(k,k)}[\exp](s) = (-1)^{k+1}\frac{\widehat{B}_k(-s)}{\widehat{B}_k(s)} = e^s + \mathcal{O}(s^{2k}). \tag{8.32}$$

a. We have already seen Padé approximations of the exponential when considering the stability function of RPM (see section D.7)

B) Time-limited transfer functions: Restricting shifted orthonormal Legendre polynomials to the unit time interval, their Laplace transform is

$$\widehat{P}_k(s) := \int_0^1 e^{-\tau s} P_k(\tau)\,\mathrm{d}\tau, \tag{8.33}$$

Symbolic computation yields the results shown in table 8.1.

k	$\widehat{P}_k(s)$
0	$\dfrac{1-e^{-s}}{s}$
1	$\dfrac{\sqrt{3}}{s^2}\left((2-s)-(2+s)e^{-s}\right)$
2	$\dfrac{\sqrt{5}}{s^3}\left(\left(12-6s+s^2\right)-(12+6s+s^2)e^{-s}\right)$
3	$\dfrac{\sqrt{7}}{s^4}\left((120-60s+12s^2-s^3)-(120+60s+12s^2+s^3)e^{-s}\right)$

Table 8.1 – Laplace transforms of Legendre polynomials restricted to $(0,1)$.

Then, we introduce the Legendre convolution operator of order p in the Laplace domain by

$$\widehat{\boldsymbol{P}}_p(s) = \begin{bmatrix}\widehat{P}_0(s) & \dots & \widehat{P}_{p-1}(s)\end{bmatrix}. \tag{8.34}$$

Remark 8.7 (Laplace exponential approximation error)**.** The Laplace transforms of unrestricted and time-limited polynomials are linked by the identity

$$\widehat{P}_k(s) = \widehat{B}_k(s) - (-1)^{k+1}\widehat{B}_k(-s)e^{-s}. \tag{8.35}$$

Furthermore, dividing (8.35) by $\widehat{B}_k(s)$ and using equation (8.32), one can form the error

$$\widehat{E}_k(s) = \frac{\widehat{P}_k(s)}{\widehat{B}_k(s)} = 1 - \frac{\mathrm{Pade}_{(k,k)}[\exp](s)}{\exp(s)} = -\mathcal{O}(s^{2k})e^{-s}. \tag{8.36}$$

This error quantifies regions, in the Laplace domain, where the time shift operator e^s is well approximated by the projection methods (shown on figures 8.21 and 8.22). This error measure is also closely related to the stability theory of order stars (see [WHN78]).

k	0	1	2	3
$\widehat{E}_k(s)$	$(1-e^{-s})$	$\left(1-\frac{2+s}{2-s}e^{-s}\right)$	$\left(1-\frac{s^2+6s+12}{s^2-6s+12}e^{-s}\right)$	$\left(1-\frac{120+60s+12s^2+s^3}{120-60s+12s^2-s^3}e^{-s}\right)$

Table 8.2 – Laplace exponential approximation error for the Legendre polynomials.

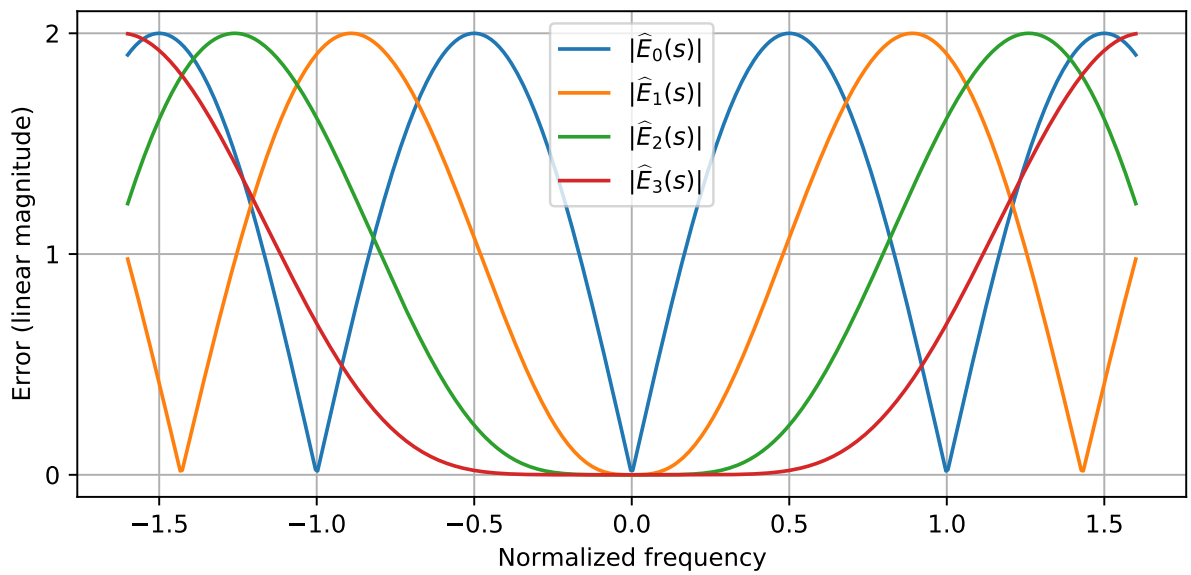


Figure 8.21 – Legendre exponential approximation error in the frequency domain. Note the manifestation of Strang–Fix conditions in the spectral domain (see eq. (3.22) p.87 and appendix C.3 p.285): the order of accuracy increases with the number of zeros of the error $\widehat{E}_k(s)$ at the origin $s = 0$, which in turn increases the width of the maximally flat approximation region.

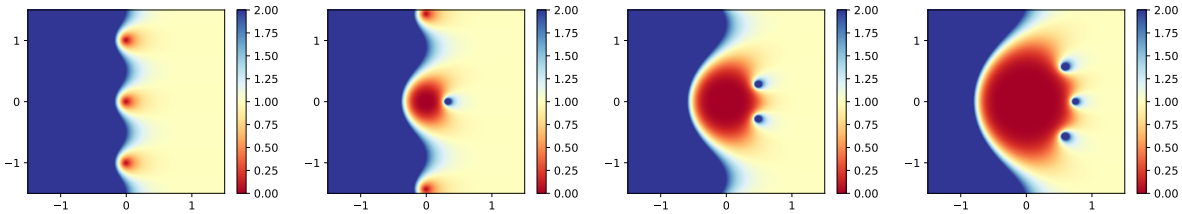


Figure 8.22 – Exponential approximation error $\left|\widehat{E}_k(s)\right|$ in the Laplace plane for Legendre polynomials for $k = 0, 1, 2, 3$ (from left to right). We observe that the accurate region (in red) increases with the order p. Furthermore, the periodicity of oscillations gets slower on the Fourrier axis $i\mathbb{R}$ as a mark of increased bandwidth.

Step3: Laplace transfer function Remind that because of (frame-synchronous) projection, the linear system is h-shift-invariant [17] but not continuous-shift-invariant: for time shifts $\tau = kh$, $k \in \mathbb{Z}$, a delayed input yields a delayed output $\mathcal{Y}(e^{-\tau s}U(s)) = e^{-\tau s}\mathcal{Y}(U(s))$. Hence its Laplace transfer function is generally not defined.

For simplicity, we restrict our study to a frame-synchronous zero-order-hold input $u(t) = \sum_n P_0(t/h - n)u[n]$ with samples $u[n]$, which *already belongs to the projection space*. Its Laplace transform is $U(s) = \widehat{P}_0(hs)\widehat{u}(z = e^{hs})$ where $\widehat{u}(z)$ denotes the Z-transform of sequence $u[n]$, so that the Z-domain input of the discrete filterbank is $\widehat{\boldsymbol{u}}(z) = \begin{bmatrix} 1 \\ \mathbf{0}_{p-1} \end{bmatrix} \widehat{u}(z)$. Then, the Laplace transform of the continuous output of order p is

$$Y_p(s) = \widehat{\boldsymbol{P}}_p(hs)\widehat{\boldsymbol{H}}_p(z)\left(\begin{bmatrix} 1 \\ \mathbf{0}_{p-1} \end{bmatrix} \widehat{u}(z)\right), \quad \text{for} \quad z = e^{hs}.$$

For *this particular (frame-synchronous) input*, dividing Y_p by U and cancelling $\widehat{u}(z)$ finally yields the Laplace transfer function

$$H_p(s) := \frac{Y_p(s)}{U(s)} = \widehat{\boldsymbol{P}}_p(hs)\widehat{\boldsymbol{H}}_p(e^{hs}) \begin{bmatrix} 1 \\ \mathbf{0}_{p-1} \end{bmatrix} \frac{1}{\widehat{P}_0(hs)}. \tag{8.37}$$

Observations The magnitude and phases responses are displayed on figures 8.23 8.24. We make the following observations:

a) Starting with order $p \geq 2$, it is possible to simulate a pole *above the Nyquist frequency*,

b) Such a pole is subject to frequency warping, but the warping error gets lower when increasing either the sampling rate f_s or the projection order p.

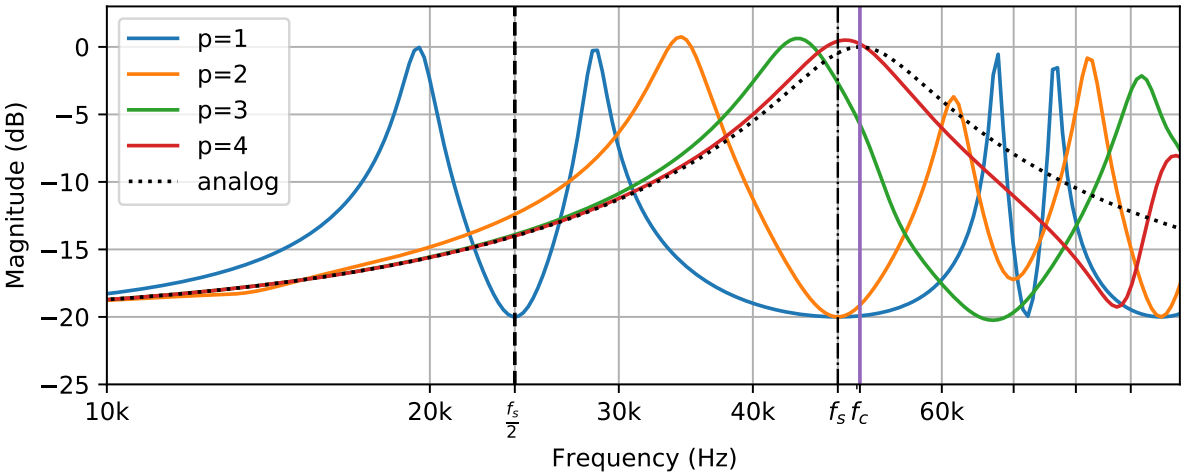
c) Starting with order $p \geq 2$ the frequency response below 20 kHz [18] is qualitatively very similar to the analog one [19].

d) For $p \geq 3$ the response is very close to the analog one, even for low sampling rates $f_s \ll f_c$. For $p = 2$, a small amount of oversampling is beneficial, while for $p = 1$, it is necessary to use the classical Shannon-Nyquist condition $f_s > 2f_c$ to obtain a good match below 20 kHz.

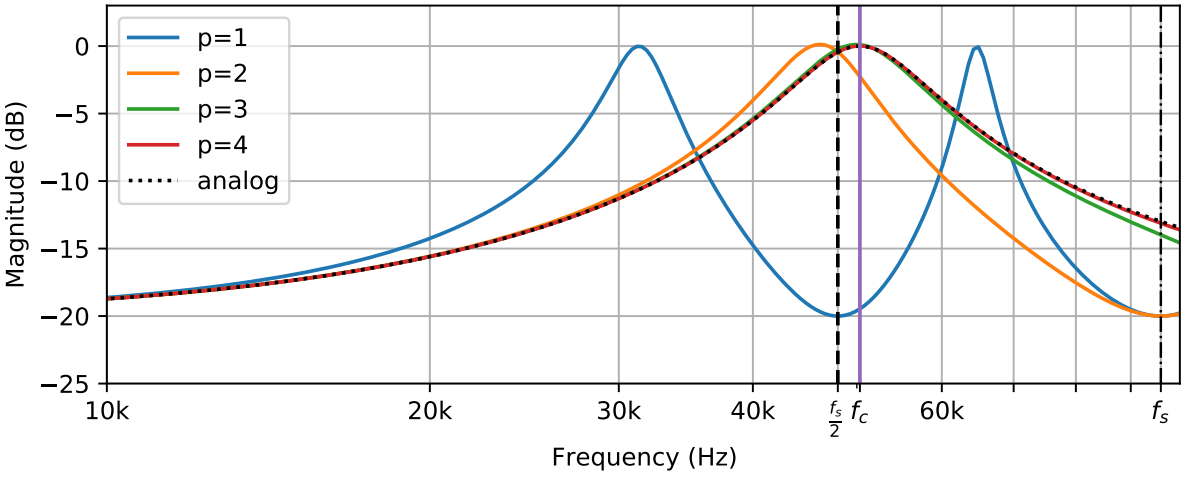
17. Using the lifting isomorphism $\breve{u}[n](\tau) = u(h(n + \tau))$, it can be transformed to an equivalent discrete shift-invariant system with an infinite number of "phases" τ between sampling instants n (see [MM10]).

18. For audio use, we are not interested in frequencies above 20kHz i.e. the limit of audible frequencies.

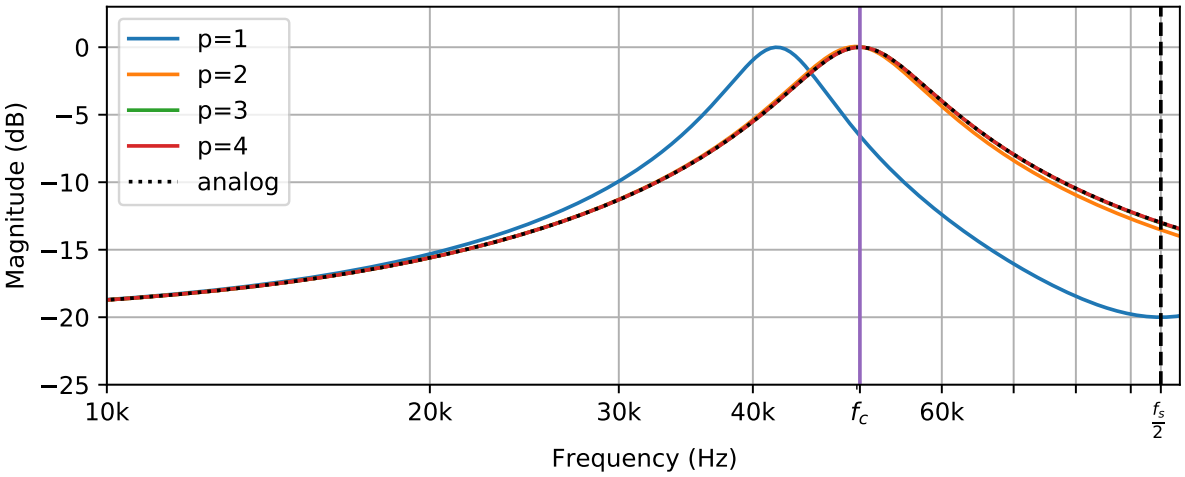
19. We get rid of the compression of the analog frequency axis $[0, \infty)$ to the digital one $[0, f_s/2)$ that is typical of the mid-point and bilinear schemes.



(a) $f_s = 48$ kHz

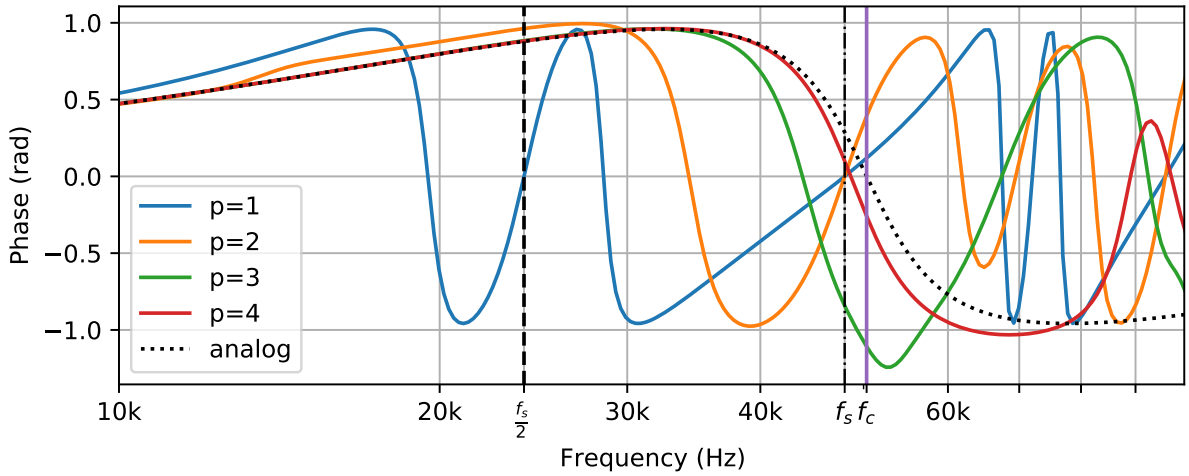


(b) $f_s = 96$ kHz

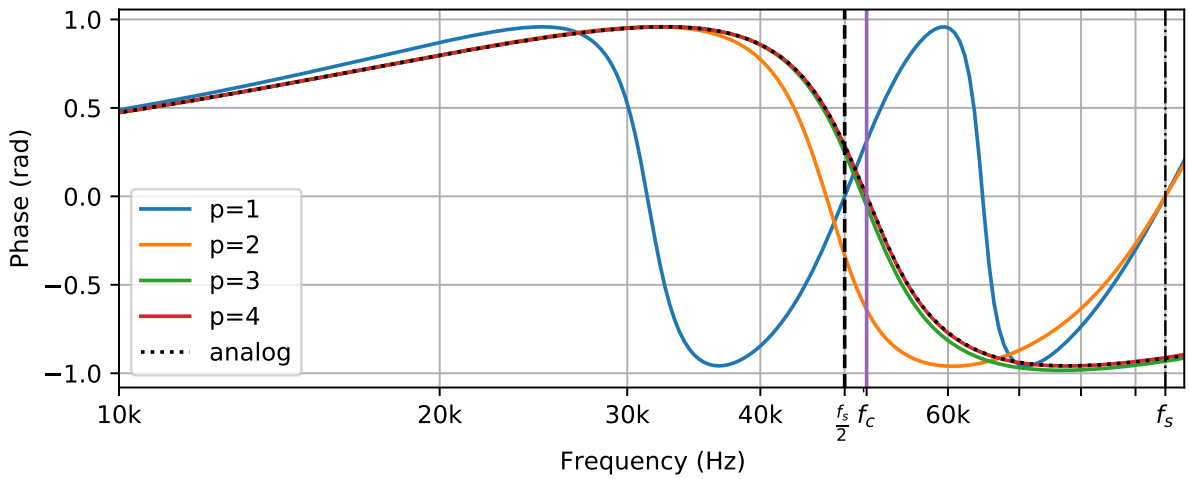


(c) $f_s = 192$ kHz

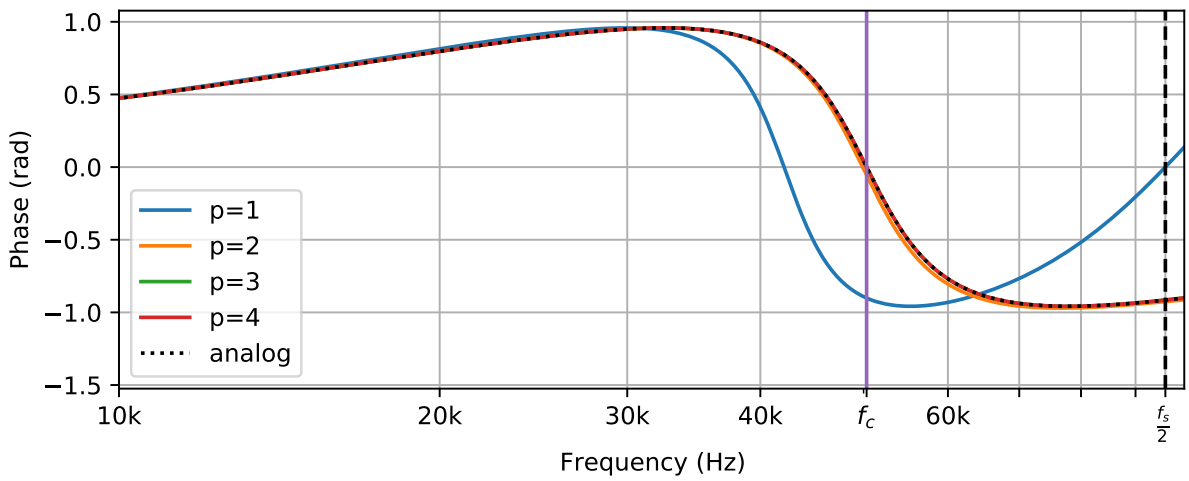
Figure 8.23 – (Peaking EQ) Magnitude response of the projected system with cutoff frequency $f_c \approx 50$ kHz for common audio sampling rates $f_s \in [48, 96, 192]$ kHz and projection orders $p = 1, 2, 3, 4$. (No prewarping has been applied to observe the effects of frequency warping).



(a) $f_s = 48$ kHz



(b) $f_s = 96$ kHz



(c) $f_s = 192$ kHz

Figure 8.24 – (Peaking EQ) Phase response of the projected system with cutoff frequency $f_c \approx 50$ kHz for common audio sampling rates $f_s \in [48, 96, 192]$ kHz and projection orders $p = 1, 2, 3, 4$.

Oversampling vs increasing order

Building on the previous observations, it is natural to ask the following question: How does oversampling by a factor q (i.e lowering the step size $h_q = h/q$) compares to raising the projection order p for *the same number of parameters* pq *by time-step*? i.e. we compare simulations that have the same *rate of innovation* $B_{p,q} = pq/h$ (generalized bandwidth). To measure both magnitude and phase innaccuracies, we introduce the following relative error in the Fourrier domain

$$\epsilon_{p,q}(f) = \left| \frac{H_{\mathsf{EQ}}(s) - H_p\left(\frac{hs}{q}\right)}{H_{\mathsf{EQ}}(s)} \right|_{s=j2\pi f} . \tag{8.38}$$

For a base audio sampling rate $f_s = 1/h = 48$ kHz, we compare the error $\epsilon_{1,q}$ to $\epsilon_{p,1}$ for $pq = 2, 3, 4$, that is pure oversampling $\epsilon_{q,1}$ versus pure order increase $\epsilon_{p,1}$ strategies

Results are shown in table 8.3 and on figure 8.25. Considering the audible frequency band below the Nyquist frequency $f_s/2$, we remark that the higher order approximation error $\epsilon_{p,1}$ is always lower than the oversampled approximation error $\epsilon_{1,q}$ by at least 10 dB. This is confirmed by the results in table 8.3. Furthermore, thanks to the higher accuracy, the error drops much faster for sub-Nyquist frequencies (see footnote 20). Above the Nyquist frequency, we remark that the maximum errors for each approximations are comparable, but the high order error $\epsilon_{p,1}$ is lower most of the time. In summary, we observe that:

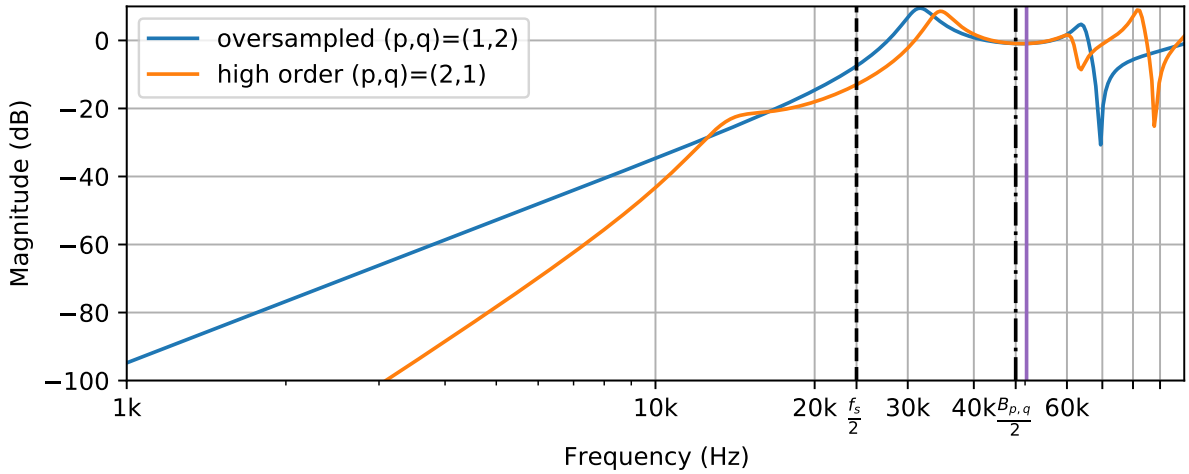
increasing the projection order p improves the error much faster than oversampling by q,

even when the pole is not small compared to the frame rate $1/h$ [20]. We conjecture that this increased domain of accuracy must be limited to a region within or close to the generalized bandwidth $B_{p,q}$ (see fig. 8.22). This issue would require a dedicated study and is left for further research. As another perspective, the L^2-orthogonal V-system [MQSW07] is a generalization of Legendre polynomials and Haar Wavelets which can both reproduce polynomials up to order p and cover multiple time scales. This way, different trade-offs between high-order accuracy and frequency resolution than the ones presented here could be considered.

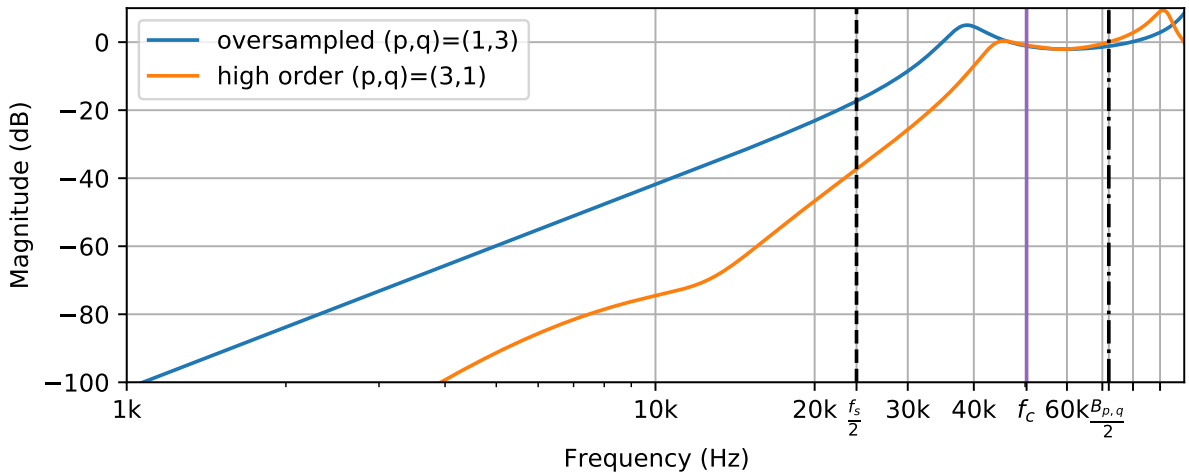
comparison on 20 Hz - 20 kHz	$pq = 2$	$pq = 3$	$pq = 4$
oversampled: maximum error $\left\|\epsilon_{1,q}\right\|_\infty$	$185 \cdot 10^{-3}$	$70.0 \cdot 10^{-3}$	$37.3 \cdot 10^{-3}$
high order: maximum error $\left\|\epsilon_{p,1}\right\|_\infty$	$\mathbf{125 \cdot 10^{-3}}$	$\mathbf{4.63 \cdot 10^{-3}}$	$\mathbf{1.21 \cdot 10^{-3}}$
oversampled: mean abs error $\left\|\epsilon_{1,q}\right\|_1$	$7.91 \cdot 10^{-3}$	$3.22 \cdot 10^{-3}$	$1.76 \cdot 10^{-3}$
high order: mean abs error $\left\|\epsilon_{p,1}\right\|_1$	$\mathbf{6.57 \cdot 10^{-3}}$	$\mathbf{1.25 \cdot 10^{-4}}$	$\mathbf{3.54 \cdot 10^{-5}}$

Table 8.3 – (Peaking EQ) comparison of the transfer function approximation error $\epsilon_{1,q}$ (oversampling) and $\epsilon_{p,1}$ (high order) over the audible range $20 - 20000$ Hz. The frequency domain error of high order discretisation is systematically lower than the oversampled one for the same degrees of freedom per time step $pq = 2, 3, 4$.

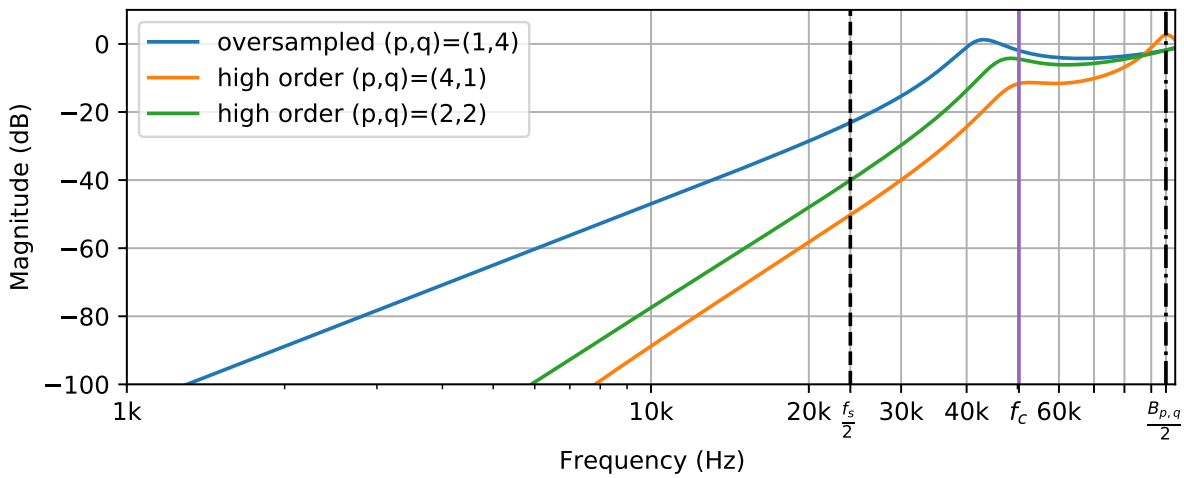
20. We remind that our test uses a pole *above the Nyquist frequency* (also above the base sampling rate) to challenge the numerical method. Otherwise, for sub-Nyquist poles such that $|h\lambda| \ll 1$, it is already obvious from accuracy analysis that non-oversampled high-order methods have an error in $\mathcal{O}(|h\lambda|^{2p})$ which drops exponentially with p, much faster than the error of oversampled second-order methods in $\mathcal{O}(|h\lambda/q|^2)$ i.e. polynomial in q.



(a) $pq = 2$



(b) $pq = 3$



(c) $pq = 4$

Figure 8.25 – (Peaking EQ) Comparison of transfer function approximation errors $\epsilon_{p,q}(f)$ (in decibels) for a constant number of parameters pq (see eq. (8.38)). The oversampling error $\epsilon_{1,q}$ (blue) is compared to the high order error $\epsilon_{p,1}$ (orange) for $pq = 2, 3, 4$.

Conclusion

In this chapter, we have reviewed a number of representative electronic audio circuits. Circuits have been modeled using the PHS framework with a systematic transformational approach from the circuit graph to continuous and discrete time simulation equations using the tools of chapter 2 and 5. We have considered bipolar transistors in section 8.1, diode clipping and filtering in section 8.2, operational amplifiers with feedback saturation in section 8.3 and 8.4, a self-oscillating resonant filter in section 8.4, a passive equalizer with a resonance above the Nyquist frequency in section 8.6 and a relaxation oscillator using a tunnel diode as non-monotonous negative-resistance element in section 8.5. The FuzzFace circuit had to be simulated as a pH-DAE because of the algebraic coupling between transistors, while others like the MS-20 or the relaxation oscillator could be simulated as ODEs. For the MS-20, we chose to pre-solve the algebraic feedback nonlinearity offline as an equivalent component, rather than having to solve a stiff DAE. This approach considerably reduces the complexity of simulation at the price of more preparation work. All of the circuits were nonlinear except the peaking equaliser. For this circuit, we exploited linearity to study the accuracy and increased bandwidth of high order projection methods in the spectral domain. We confirmed that high-order methods have faster convergence than oversampling for open linear systems, even more when the frequency region of interest is below the Nyquist frequency.

Part IV

Towards Geometric Algebra

Chapter 9

Geometric Algebra for PHS

Contents

9.1 Introduction to Geometric Algebra . . . 240
9.1.1 Linear geometric transforms . . . 245
9.1.2 Sub-algebras . . . 247
9.2 Motivating examples and invariants . . . 249
9.2.1 Harmonic oscillator . . . 249
9.2.2 Dissipative oscillator . . . 250
9.2.3 Maxwell equations (in empty space) . . . 251
9.3 Port-Hamiltonian systems using Geometric Algebra . . . 252
9.3.1 Going further: unifying transforms . . . 253
9.4 Representing Dirac structures with Geometric Algebra . . . 254
9.4.1 Dirac structures . . . 256
9.5 Exploring the geometry of R(n,n) with GA . . . 258
9.6 Rotor description of the flow-effort to wave variables change . . . 260
9.6.1 Hyperbolic squeeze mapping . . . 260
9.6.2 Rotation by $\pi/4$ . . . 261

This chapter is dedicated to *Geometric Algebra* (GA) and attempts to highlight its potentialities for port-Hamiltonian System modelling.

A complete overview of GA is clearly out of the scope of this chapter, Geometric Algebra is at the same time very simple and elementary in its construction, making a perfect fit for undergraduates, and very far reaching, unifying concepts as diverse as complex numbers, split complex numbers, quaternions, octonions, Pauli an Dirac matrices, projective, conformal and non-euclidean geometries within a unifying framework. A main difficulty to its wider adoption is related to the fact that it requires *unlearning* to fully grasp its full potential. In particular, it is necessary to get rid of the three dimensional cross product [1] (which does not generalises to an arbitrary number of dimension). A second learning barrier, which I found more difficult in practice, is to stop identifying General Linear transforms with their matrix representation. This chapter describes my personal journey towards using geometric algebra with port-Hamiltonian systems.

Section 9.1, is a brief introduction to Geometric Algebra. In section 9.2, we show some motivating examples where Geometric Algebra is a key tool to simplify the representation of physical problems allowing to extract their invariants. In section 9.3, we use GA to represent

1. The 3-dimensional cross product can be defined as the Hodge dual of the exterior product of two vectors.

General Linear transforms uniformly as (parabolic, hyberbolic) rotations using elements of *the same algebra*[2]. In section 9.4, we use GA to describe Dirac structures, revisiting the content of section 1.3.1 p.20.

In sections 9.3 and 9.4 we use non-euclidean geometry which is required to describe the duality pairing of Dirac structures and hyperbolic transformations in general linear transforms. Section 9.3 and 9.4 present some initial work that needs to be further devlopped and matured. This work shows how to technically represent Dirac structures and General linear transforms with Geometric Algebra. However it still lacks the simplifying elegance usually associated with GA. One of the main difficulty is that intuitions from euclidean geometry are no longer valid in non-euclidean spaces[3]. I hope that this chapter motivates more people to adopt Geometric Algebra and find more satisfying answers to these questions.

2. A powerful property of complex numbers is that a complex number can represent both a point of the 2D space and a scaling/rotation. In GA, we can generalize this property. Another common example from computer graphics is that 3D geometry is significantly simplified by using quaternions to represent affine 3D transformations.

3. We note that reference [Hes93] avoids non-euclidean metrics by identifying the configuration space with its dual: the momentum space. Conversely, in [DHSVA93] non-euclidean signatures are key to represent general linear transforms $GL(n)$ as orthogonal transforms $O(n, n)$.

Why use Geometric Algebra for PHS?

Without diving into details yet [4], my original motivation for trying to encode the physics of PHS using the language of geometric algebra arises from the following observations:

1) For conservative systems of the form $\dot{\mathbf{x}} = \mathbf{J}\nabla H(\mathbf{x})$, the skew-symmetric matrix $\mathbf{J} = -\mathbf{J}^\mathsf{T}$, is an infinitesimal generator of rotations. It defines an *anti-commutative Poisson bracket* [5] [Olv00, p.390] $\{f, g\}_\mathbf{J} = -\{g, f\}_\mathbf{J}$. In the language of Grassmann algebras, this is intimately linked to the notions of *exterior product* $\wedge$ and *bivector* so that the dynamic of Poisson/Hamiltonian systems can be described by the Poisson bracket

$$\dot{\mathbf{x}} = \{\mathbf{x}, H\}_\mathbf{J}\,.$$

2) For dissipative gradient systems of the form $\dot{\mathbf{x}} = -\mathbf{R}\nabla H(\mathbf{x})$, a symmetric positive semi-definite dissipation matrix $\mathbf{R} = \mathbf{R}^\mathsf{T} \succeq 0$ is used to encode dissipation. In turn, this induces a *metric bracket* [6] $(f, g)_\mathbf{R} = (g, f)_\mathbf{R}$. The dynamic of purely dissipative gradient systems can be written using the metric bracket as

$$\dot{\mathbf{x}} = -(\mathbf{x}, H)_\mathbf{R}.$$

3) For dissipative PHS of the form $\dot{\mathbf{x}} = (\mathbf{J} - \mathbf{R})\nabla H(\mathbf{x})$, both rotation and dissipation happen at the same time. This is unified in the geometry of *metriplectic* systems [Mor86, BMBM18], by introducing the notion of a *metriplectic bracket* $[[f, g]] = \{f, g\}_\mathbf{J} - (f, g)_\mathbf{R}$ to combine purely conservative and purely dissipative geometries.

4) From the geometric algebra viewpoint [7], the geometric product $\mathbf{uv}$ of two vectors $\mathbf{u}, \mathbf{v}$ is equal to the sum of the inner product $\mathbf{u} \cdot \mathbf{v}$ (a scalar) and the exterior product $\mathbf{u} \wedge \mathbf{v}$ (a bivector). Furthermore, while the cosine of the angle between vectors $\mathbf{u}, \mathbf{v}$ is naturally encoded by the inner product into the scalar part [8] $\mathbf{1}$ of the algebra, the exterior product completes the picture by encoding the sine of the angle into the bivector part $\boldsymbol{i}$ [9] (generalizing complex numbers in any dimensions). This is summarized by the following identity

$$\mathbf{uv} = \mathbf{u} \cdot \mathbf{v} + \mathbf{u} \wedge \mathbf{v} = |\mathbf{u}||\mathbf{v}|\,(\mathbf{1}\cos\theta + \boldsymbol{i}\sin\theta)\,, \quad \text{where} \quad \mathbf{1} := \frac{\mathbf{u} \cdot \mathbf{v}}{|\mathbf{u}||\mathbf{v}|}, \quad \boldsymbol{i} := \frac{\mathbf{u} \wedge \mathbf{v}}{|\mathbf{u}||\mathbf{v}|}.$$

Since geometric algebra has the power to unify inner and exterior products into a single geometric product, it seems natural to embrace this formalism and study its consequences for PHS modelling.

This chapter is a personal take on the subject and the reflect of my current understanding (far from being complete). The proposed approach is to put aside our knowledge of matrix algebra and to exclusively use GA constructs to reintroduce, step by step, the PH modelling tools from chapter 1 p.7. For the formulation of Hamiltonian mechanics using GA see [Hes93] and [DGL+03, p. 432]. For Lagrangian mechanics see [DGL+03, p. 420].

4. See [Olv00, p.390] for a definition of the Poisson bracket and [Mor86, BMBM18] for metriplectic geometry.

5. In euclidean coordinates the Poisson bracket is $\{f, g\}_\mathbf{J} = \sum_{i,j} \frac{\partial f}{\partial x_i} \mathbf{J}_{ij} \frac{\partial g}{\partial x_j}$ so that $\{\mathbf{x}, H\}_\mathbf{J} = \mathbf{J}\nabla H(\mathbf{x})$.

6. In euclidean coordinates the metric bracket is $(f, g)_\mathbf{R} = \sum_{ij} \frac{\partial f}{\partial x_i} \mathbf{R}_{ij} \frac{\partial g}{\partial x_j}$ so that $(\mathbf{x}, H)_\mathbf{R} = \mathbf{R}\nabla H(\mathbf{x})$.

7. An introduction is detailed in section 9.1

8. Geometric algebra is a *graded* algebra, i.e. is has 0-vectors, 1-vectors, 2-vectors, etc. It is a common notation to denote $\mathbf{1}$ the basis element representing the scalar part of the algebra (a 0-vector).

9. We use the symbol $\boldsymbol{i}$ to emphasize its role as a complex number, *in the plane spanned by vectors* $\mathbf{u}$, $\mathbf{v}$. But, it is embedded and can be oriented arbitrarily in dimension n.

9.1 Introduction to Geometric Algebra

> **Remark 9.1** (Reading advice)**.** Introducing Geometric Algebra (GA) in just a few pages is not an easy task. For a self-taught introduction to GA, I recommend starting from the basics by reading reference [Mac10] (taking the time to do the exercises) followed by [Mac12b] on Geometric Calculus (GC). For more advanced topics and physical applications, the book [DGL+03] is a very good starting point. For a quick course on GA, see [Mac17, GLD93] and [Hes14, Hes86] see also [Hit01]. For the relation between GA and differential geometry refer to [Hes11]. For minimal and axiomatic constructions of GA see [Mac02, Art06], see also [DGL+03, p.84]. In this manuscript, I will deliberately skip some of the hallmarks of GA such as Space-time Algebra, and GA representations of Dirac and Pauli matrices.

Modern Geometric Algebra was initiated by David Hestenes building on the work of Hamilton, Grassmann and Clifford. A main difference with Clifford Algebras is in the simpler notations [10] and the stronger focus on geometry (hence the name). The main concept of GA is the introduction of the *geometric product*. This makes the product of two (multi-)vectors a well-defined mathematical object. It also gives rise to the introduction of mathematical objects such as the inverse of a (nonzero) vector, blades, multi-vectors, pseudo-scalars, spinors, etc (introduced below). To see this, we start from well known concepts such as the inner product and the exterior product before introducing the (graded) *geometric algebra*.

Inner product (of vectors) The inner product, denoted $\mathbf{u}\cdot\mathbf{v}$, of two vectors $\mathbf{u}$, $\mathbf{v}$ is a *scalar* number with magnitude $|\mathbf{u}||\mathbf{v}|\cos\theta$ where $|\mathbf{u}| \equiv \sqrt{\mathbf{u}\cdot\mathbf{u}}$ denotes the length (norm) of $\mathbf{u}$ and θ is the angle from $\mathbf{u}$ to $\mathbf{v}$. It satisfies the symmetric relation

$$\mathbf{u}\cdot\mathbf{v} = \mathbf{v}\cdot\mathbf{u}.$$

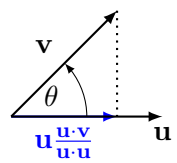
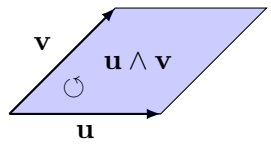


Figure 9.1 – Inner product of vectors.

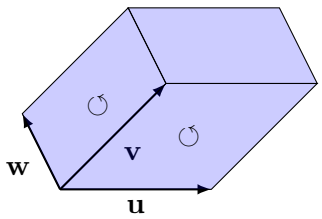
Exterior product (of vectors) The inner product only gives a partial information regarding vectors $\mathbf{u}$, $\mathbf{v}$. Traditionally, in 3 dimension it is customary to use the cross product $\mathbf{u}\times\mathbf{v}$, however such a construct is only valid in 3-dimensional space. Instead, Grassmann introduced the exterior product $\wedge$ and the associated exterior algebras. The exterior product $\mathbf{u}\wedge\mathbf{v}$ of two vectors $\mathbf{u}$, $\mathbf{v}$ has magnitude $|\mathbf{u}||\mathbf{v}|\sin\theta$ but it is not a scalar or a vector: it is an *oriented area* (or bivector or 2-vector) from $\mathbf{u}$ to $\mathbf{v}$. It satisfies the anti-commutative relation

$$\mathbf{u}\wedge\mathbf{v} = -\mathbf{v}\wedge\mathbf{u}.$$

A geometric interpretation of the exterior product $\mathbf{u}\wedge\mathbf{v}$ is the oriented area corresponding to the parallelogram formed by vectors $\mathbf{u}$, $\mathbf{v}$. This construction can be generalised to any number of vectors leading to the notion of *k-blades* [11] representing *oriented volumes* between vectors [12]. For example in 3-dimension the volume of highest grade is a 3-volume represented by the 3-vector (or 3-blade) $\mathbf{u}\wedge\mathbf{v}\wedge\mathbf{w}$.



(a) oriented area



(b) oriented volume

Figure 9.2 – Exterior product of vectors.

10. which makes it more approachable by non mathematicians.
11. A blade is equal to the product of nonzero orthogonal vectors $\mathbf{B} = \mathbf{e}_1\wedge\mathbf{e}_2\wedge\mathbf{e}_k$ so that its norm $|\mathbf{B}| = |\mathbf{e}_1|\ldots|\mathbf{e}_k|$ is equivalent to the volume of the rectangular parallelogram with edges $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_k$.
12. if some vectors are co-linear then their oriented volume is zero.

Geometric product (of vectors) We can think of the inner and outer products as the symmetric and antisymmetric parts of a new product called the geometric product [13] below.

$$\mathbf{uv} \equiv \mathbf{u} \cdot \mathbf{v} + \mathbf{u} \wedge \mathbf{v}$$

We remark that the inner product and the exterior product respectively lower and rise the grade of their operands. The product of *parallel* vectors is a pure *scalar* and the product of *orthogonal* vectors and is a *bivector*. A more axiomatic approach (detailed below) is to reverse the situation and extract the inner product and exterior product respectively as the symmetric and skew-symmetric parts of the geometric product

$$\mathbf{u} \cdot \mathbf{v} = \frac{1}{2}(\mathbf{uv} + \mathbf{vu}), \qquad \mathbf{u} \wedge \mathbf{v} = \frac{1}{2}(\mathbf{uv} - \mathbf{vu}). \tag{9.1}$$

Geometric algebra We reproduce the following definition of geometric algebra.

Definition 9.1 (Geometric algebra [Mac17])**.** The *geometric algebra* $\mathbb{G}^n$ is an extension of the inner product space $\mathbb{R}^n$ noted $\mathbb{G}^n := \mathcal{G}(\mathbb{R}^n)$. It is an associative algebra with scalar identity element $\mathbf{1}$. That is, it is a vector space with a product satisfying properties P1-P4 for all scalars a and elements $A, B, C \in \mathbb{G}^n$.

P1. $A(B + C) = AB + AC$ and $(B + C)A = BA + CA$ (left and right distributivity) ,

P2. $(aA)B = A(aB) = a(AB)$ (Compatibility of scalar and geometric multiplication),

P3. $(AB)C = A(BC)$ (Associativity)

P4. $\mathbf{1}A = A\mathbf{1}$ (Commuting left and right multiplicative identity)

the product is called the *geometric product*. Members of $\mathbb{G}^n$ are called *multi-vectors*. We list two more properties.

P5. The geometric product of $\mathbb{G}^n$ is linked to the algebraic structure of $\mathbb{R}^n$ by

$$\mathbf{u}^2 = \mathbf{uu} = \mathbf{u} \cdot \mathbf{u} = \mathbf{1}|\mathbf{u}|^2 \qquad \forall \mathbf{u} \in \mathbb{R}^n$$

P6. Every orthonormal basis of $\mathbb{R}^n$ determines a canonical basis of the vector space $\mathbb{G}^n$ (see table 9.1 p.242).

Property P5 yields that nonzero vectors have a *multiplicative inverse* in $\mathbb{G}^n$ noted $\mathbf{u}^{-1} = \mathbf{u}/|\mathbf{u}|^2$.

Notations GA is a *graded algebra*. In the general setting, an element A of the GA is a mixed-grade *multivector*. It can be decomposed as a direct sum of graded k-vectors (a sum of k-blades) noted $\langle A \rangle_k$ where $\langle \cdot \rangle_k$ is the *grade extracting operator* of order k, so that

$$A = \langle A \rangle_0 + \langle A \rangle_1 + \langle A \rangle_2 + \ldots + \langle A \rangle_n .$$

The *geometric product* of two multivectors M, N is denoted MN. An important operation in GA is called *reversion* which reverses the order of its operands. It is defined and denoted by

$$(MN)^\dagger = N^\dagger M^\dagger, \qquad \langle M \rangle_1^\dagger = \langle M \rangle_1 . \tag{9.2}$$

In this thesis, we use the lower case bold notation $\mathbf{u}$ for vectors, uppercase bold $\mathbf{B}$ for bivectors, and lower case standard font a for scalars. As an exception, the neutral element of GA is often denoted $\mathbf{1}$ to higlight its role as a basis for elements of grade 0 (scalars) as in $a \equiv \mathbf{1}a$.

13. Note that the axiomatic definition (def. 9.1) of the geometric product is preferable to manipulate multi-vectors of mixed grade. The identity $\mathbf{uv} = \mathbf{u} \cdot \mathbf{v} + \mathbf{u} \wedge \mathbf{v}$ is only valid for vectors. See equations (9.3)-(9.5)

Canonical basis of $\mathbb{G}^n$ Let $\{\mathbf{e}_i\}_{i=1}^n$ be an orthonormal basis of $\mathbb{R}^n$ with signature $\mathbf{e}_i^2 = \mathbf{e}_i \cdot \mathbf{e}_i = \mathbf{1}$ (by definition $\mathbf{e}_i \cdot \mathbf{e}_j = \mathbf{1}\delta_{ij}$). The vector space $\mathbb{G}^n = \mathcal{G}(\mathbb{R}^n)$ has a canonical basis of dimension 2^n. Its subspaces (of grade k) have dimension $\binom{n}{k}$. Examples for $\mathbb{G}^2, \mathbb{G}^3, \mathbb{G}^4$ are given in table 9.1.

Grade k	basis	denomination	cardinality $\binom{n}{k}$
0	$\mathbf{1}$	0-vectors (scalars)	1
1	$\mathbf{e}_1$, $\mathbf{e}_2$	1-vectors (vectors)	2
2	$\mathbf{e}_1\mathbf{e}_2$	2-vectors (bivectors)	1

(a) $\mathbb{G}^2$, $\dim(\mathbb{G}^2) = 4$

Grade k	basis	denomination	cardinality $\binom{n}{k}$
0	$\mathbf{1}$	0-vectors (scalars)	1
1	$\mathbf{e}_1$, $\mathbf{e}_2$, $\mathbf{e}_3$	1-vectors (vectors)	3
2	$\mathbf{e}_2\mathbf{e}_3$, $\mathbf{e}_3\mathbf{e}_1$, $\mathbf{e}_1\mathbf{e}_2$	2-vectors (bivectors)	3
3	$\mathbf{e}_1\mathbf{e}_2\mathbf{e}_3$	3-vectors (trivectors)	1

(b) $\mathbb{G}^3$, $\dim(\mathbb{G}^3) = 8$

Grade k	basis	denomination	cardinality $\binom{n}{k}$
0	$\mathbf{1}$	0-vectors (scalars)	1
1	$\mathbf{e}_1$, $\mathbf{e}_2$, $\mathbf{e}_3$, $\mathbf{e}_4$	1-vectors (vectors)	4
2	$\mathbf{e}_1\mathbf{e}_2$, $\mathbf{e}_1\mathbf{e}_3$, $\mathbf{e}_1\mathbf{e}_4$, $\mathbf{e}_2\mathbf{e}_3$, $\mathbf{e}_2\mathbf{e}_4$, $\mathbf{e}_3\mathbf{e}_4$	2-vectors (bivectors)	6
3	$\mathbf{e}_2\mathbf{e}_3\mathbf{e}_4$, $\mathbf{e}_3\mathbf{e}_4\mathbf{e}_1$, $\mathbf{e}_4\mathbf{e}_1\mathbf{e}_2$, $\mathbf{e}_1\mathbf{e}_2\mathbf{e}_3$	3-vectors (trivectors)	4
4	$\mathbf{e}_1\mathbf{e}_2\mathbf{e}_3\mathbf{e}_4$	4-vectors (quadrivectors)	1

(c) $\mathbb{G}^4$, $\dim(\mathbb{G}^4) = 16$

Table 9.1 – Canonical bases of $\mathbb{G}^2$, $\mathbb{G}^3$, $\mathbb{G}^4$.

Multiplication tables To get an understanding and some intuition of the algebra, one can obtain the multiplication tables [14] using the following properties

- By collinearity, orthonormal vectors in $\mathbb{R}^n$ square to one (since $\mathbf{e}_i \wedge \mathbf{e}_i = 0$)

$$\mathbf{e}_i^2 = \mathbf{e}_i\mathbf{e}_i = \mathbf{e}_i \cdot \mathbf{e}_i = \mathbf{1},$$

- By orthogonality, basis vectors anti-commute (because $\mathbf{e}_i \cdot \mathbf{e}_j = \delta_{ij}$)

$$\mathbf{e}_i\mathbf{e}_j = \mathbf{e}_i \wedge \mathbf{e}_j = -\mathbf{e}_j \wedge \mathbf{e}_i = -\mathbf{e}_j\mathbf{e}_i \quad i \neq j.$$

Then by reordering terms, according to anti-commutation rules, we obtain canonical basis elements (see table 9.1). The multiplication tables of $\mathbb{G}^2$ and $\mathbb{G}^3$ are shown in table 9.2.

14. Efficient numerical implementations of GA rely on fast encoding and realisation of these multiplication tables.

AB	$\mathbf{1}$	$\mathbf{e}_1$	$\mathbf{e}_2$	$\boldsymbol{i}$
$\mathbf{1}$	$\mathbf{1}$	$\mathbf{e}_1$	$\mathbf{e}_2$	$\boldsymbol{i}$
$\mathbf{e}_1$	$\mathbf{e}_1$	$\mathbf{1}$	$\boldsymbol{i}$	$\mathbf{e}_2$
$\mathbf{e}_2$	$\mathbf{e}_2$	$-\boldsymbol{i}$	$\mathbf{1}$	$-\mathbf{e}_1$
$\boldsymbol{i}$	$\boldsymbol{i}$	$-\mathbf{e}_2$	$\mathbf{e}_1$	$-\mathbf{1}$

(a) Multiplication table of $\mathbb{G}^2$. where $\boldsymbol{i} = \mathbf{e}_1\mathbf{e}_2$

AB	$\mathbf{1}$	$\mathbf{e}_1$	$\mathbf{e}_2$	$\mathbf{e}_3$	$\mathbf{B}_1$	$\mathbf{B}_2$	$\mathbf{B}_3$	$\mathbf{I}$
$\mathbf{1}$	$\mathbf{1}$	$\mathbf{e}_1$	$\mathbf{e}_2$	$\mathbf{e}_3$	$\mathbf{B}_1$	$\mathbf{B}_2$	$\mathbf{B}_3$	$\mathbf{I}$
$\mathbf{e}_1$	$\mathbf{e}_1$	$\mathbf{1}$	$\mathbf{B}_3$	$-\mathbf{B}_2$	$\mathbf{I}$	$-\mathbf{e}_3$	$\mathbf{e}_2$	$\mathbf{B}_1$
$\mathbf{e}_2$	$\mathbf{e}_2$	$-\mathbf{B}_3$	$\mathbf{1}$	$\mathbf{B}_1$	$\mathbf{e}_3$	$\mathbf{I}$	$-\mathbf{e}_1$	$\mathbf{B}_2$
$\mathbf{e}_3$	$\mathbf{e}_3$	$\mathbf{B}_2$	$-\mathbf{B}_1$	$\mathbf{1}$	$-\mathbf{e}_2$	$\mathbf{e}_1$	$\mathbf{I}$	$\mathbf{B}_3$
$\mathbf{B}_1$	$\mathbf{B}_1$	$\mathbf{I}$	$-\mathbf{e}_3$	$\mathbf{e}_2$	$-\mathbf{1}$	$-\mathbf{B}_3$	$\mathbf{B}_2$	$-\mathbf{e}_1$
$\mathbf{B}_2$	$\mathbf{B}_2$	$\mathbf{e}_3$	$\mathbf{I}$	$-\mathbf{e}_1$	$\mathbf{B}_3$	$-\mathbf{1}$	$-\mathbf{B}_1$	$-\mathbf{e}_2$
$\mathbf{B}_3$	$\mathbf{B}_3$	$-\mathbf{e}_2$	$\mathbf{e}_1$	$\mathbf{I}$	$-\mathbf{B}_2$	$\mathbf{B}_1$	$-\mathbf{1}$	$-\mathbf{e}_3$
$\mathbf{I}$	$\mathbf{I}$	$\mathbf{B}_1$	$\mathbf{B}_2$	$\mathbf{B}_3$	$-\mathbf{e}_1$	$-\mathbf{e}_2$	$-\mathbf{e}_3$	$\mathbf{1}$

(b) Multiplication table of $\mathbb{G}^3$. where $\mathbf{B}_1 = \mathbf{e}_2\mathbf{e}_3$, $\mathbf{B}_2 = \mathbf{e}_3\mathbf{e}_1$, $\mathbf{B}_3 = \mathbf{e}_1\mathbf{e}_2$, $\mathbf{I} = \mathbf{e}_1\mathbf{e}_2\mathbf{e}_3$.

Table 9.2 – (Geometric Algebra) Multiplication tables.

Extended definitions of inner and exterior product Let $\langle M\rangle_i$ denote the components of grade i (i-vectors) in M. Then, the inner product (here left-contraction [15]) and exterior product of a i-vector A with a j-vector B are respectively defined by [Mac10, p.101]

$$A \cdot B := \langle AB\rangle_{i-j}\,, \qquad\qquad A \wedge B := \langle AB\rangle_{i+j}\,, \tag{9.3}$$

where $A \wedge B = 0$ if $i + j > n$. We highlight some identities that are used in the following [16]. Note that, in the case of a vector $\mathbf{a}$ multiplied by a bivector $\mathbf{B}$, *signs are reversed compared to* (9.1)!

$$\mathbf{a} \cdot \mathbf{B} = \frac{1}{2}\left(\mathbf{aB} - \mathbf{Ba}\right), \qquad\qquad \mathbf{a} \wedge \mathbf{B} = \frac{1}{2}\left(\mathbf{aB} + \mathbf{Ba}\right). \tag{9.4}$$

More generally, for a k-vector A, the vector-blade formulae are given by

$$\mathbf{a} \cdot A = \frac{1}{2}\Big(\mathbf{a}A - (-1)^k A\mathbf{a}\Big), \qquad\qquad \mathbf{a} \wedge A = \frac{1}{2}\Big(\mathbf{a}A + (-1)^k A\mathbf{a}\Big). \tag{9.5}$$

For example, let $\mathbf{a} = \mathbf{e}_1$, $\mathbf{B} = \mathbf{e}_1\mathbf{e}_2$, then using (9.4) $\mathbf{a} \cdot \mathbf{B} = \frac{1}{2}(\mathbf{e}_1\mathbf{e}_1\mathbf{e}_2 - \mathbf{e}_1\mathbf{e}_2\mathbf{e}_1) = ((\mathbf{e}_1^2)\mathbf{e}_2 + (\mathbf{e}_1)^2\mathbf{e}_2) = \mathbf{e}_2$ and $\mathbf{a} \wedge \mathbf{B} = \frac{1}{2}(\mathbf{e}_1\mathbf{e}_1\mathbf{e}_2 + \mathbf{e}_1\mathbf{e}_2\mathbf{e}_1) = ((\mathbf{e}_1^2)\mathbf{e}_2 - (\mathbf{e}_1)^2\mathbf{e}_2) = 0$.

15. The litterature on Clifford algebras often uses the left contraction notation $A\rfloor B$ to denote $A \cdot B$.
16. We need the contraction of a vector with a bivector to implement skew-symmetric maps for Hamiltonian systems and Dirac structures.

Norm Expand a multivector A with respect to a canonical basis $\{\mathbf{e}_J\}$ [17] (of graded multivectors) as $A = \sum_J \mathbf{e}_J a_J$. Then, the norm [18] $|A|$ of A is defined by [19]

$$|A|^2 = \sum_J |a_J|^2 . \tag{9.6}$$

Inverse Generalizing the inverse of a vector (see definition 9.1 P5), let $\mathbf{B}$ be a k-blade $\mathbf{B} = \mathbf{u}_1 \wedge \ldots \wedge \mathbf{u}_k$. It can be written in an orthonormal basis $\{\mathbf{b}_i\}$ of the hyperplane spanned by $\mathbf{B}$ as $\mathbf{B} = |\mathbf{B}|\, \mathbf{b}_1 \ldots \mathbf{b}_k$. One can define its (right) inverse as the unique element $\mathbf{B}^{-1}$ such that $\mathbf{B}\mathbf{B}^{-1} = \mathbf{1}$. One can easily show that its inverse is given by the reversion

$$\mathbf{B}^{-1} = \mathbf{B}^\dagger / |\mathbf{B}| = \mathbf{b}_k \ldots \mathbf{b}_1 / |\mathbf{B}| . \tag{9.7}$$

Indeed, using $\mathbf{b}_i^2 = \mathbf{1}$ (in euclidean spaces), we have $\mathbf{B}\mathbf{B}^{-1} = |\mathbf{B}|\, \mathbf{b}_1 \ldots \mathbf{b}_k \mathbf{b}_k \ldots \mathbf{b}_1 / |\mathbf{B}| = \mathbf{1}$.

Duality The n-vectors in $\mathbb{G}^n$ are called *pseudo-scalars*. They have the property of commuting with all elements of the algebra (hence their name). For example, the unit pseudoscalar of $\mathbb{G}^3$ with orthonormal basis $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ is $\mathbf{I} = \mathbf{e}_1\mathbf{e}_2\mathbf{e}_3$ (sometimes denoted by $\mathbf{P}$ to avoid confusion). It has unit norm $|\mathbf{I}| = |\mathbf{e}_1||\mathbf{e}_2||\mathbf{e}_3| = 1$ and is unique up to a sign change when permuting the order of multiplication. Its inverse is $\mathbf{I}^{-1} = \mathbf{e}_3\mathbf{e}_2\mathbf{e}_1 = -\mathbf{I}$. In $\mathbb{G}^n$ we have $\mathbf{I}^{-1} = (-1)^{n!}\mathbf{I}$.

Definition 9.2 (Dual [Mac10])**.** The dual of a multivector A is $A^* := A/\mathbf{I}$.

For example, the dual of vector $\mathbf{e}_1$ is the bivector $\mathbf{e}_1^* = \mathbf{e}_1/\mathbf{I} = \mathbf{e}_1(\mathbf{e}_3\mathbf{e}_2\mathbf{e}_1) = -\mathbf{e}_1^2\mathbf{e}_2\mathbf{e}_3 = -\mathbf{e}_2\mathbf{e}_3$. Moreover, if a blade $\mathbf{A}$ represents the span $S_\mathbf{A} \subset \mathbf{R}^n$ of its vectors, then its dual $\mathbf{A}^*$ represents is orthogonal complements $S_\mathbf{A}^\perp$.

Theorem 9.1 (Duality [Mac10])**.** *The inner product and outer products are dual*

$$(A \cdot B)^* = A \wedge B^*, \qquad (A \wedge B)^* = A \cdot B^* \tag{9.8}$$

With this definition of the GA dual, the *Hodge dual* from exterior algebra can be defined *explicitly* by $\star(A) := -A^*$. So that in $\mathbb{G}^3$

$$\star(\mathbf{1}) = -\mathbf{I}, \qquad \star(\mathbf{e}_1) = \mathbf{e}_2 \wedge \mathbf{e}_3, \qquad \star(\mathbf{e}_2) = \mathbf{e}_3 \wedge \mathbf{e}_1, \qquad \star(\mathbf{e}_3) = \mathbf{e}_1 \wedge \mathbf{e}_2.$$

The dual extends to all elements of the $\mathbb{G}^3$ (not just to the exterior algebra $\wedge(\mathbb{R}^3) \subset \mathbb{G}^3$).

Remark 9.2 (cross product)**.** A well known example in $\mathbb{R}^3$ (whose definition does not extend to $\mathbb{R}^n$) is the cross product $\mathbf{u} \times \mathbf{v}$ of two vectors. In GA, it is defined as the (pseudo-vector) dual to the plane spanned by the bivector $\mathbf{u} \wedge \mathbf{v}$, that is

$$\mathbf{u} \times \mathbf{v} = (\mathbf{u} \wedge \mathbf{v})^* .$$

Indeed, in $\mathbb{G}^n$, the dual of a bivector is a $(n-2)$-vector (a scalar in $\mathbb{G}^2$, a vector in $\mathbb{G}^3$, a bivector in $\mathbb{G}^4$, etc).

17. Where J are multi-indexes, for example $J = (1, 2)$ denotes the basis element $\mathbf{e}_J = \mathbf{e}_1\mathbf{e}_2$.
18. Note that we can generalise to spaces of indefinite or mixed signature. The square, inner product, norm and signature $\mathbf{s}$ of a vector $\mathbf{u}$ are then linked by $\mathbf{u}^2 = \mathbf{u} \cdot \mathbf{u} = \mathbf{s}|\mathbf{u}|^2$ where $\mathbf{s} \in \{-\mathbf{1}, \mathbf{0}, \mathbf{1}\}$.
19. For example $|\mathbf{1} + 2\mathbf{e}_1 + 3\mathbf{e}_2 + 4\mathbf{e}_1\mathbf{e}_2|^2 = 1^2 + 2^2 + 3^2 + 4^2$.

9.1.1 Linear geometric transforms

Remark 9.3. One difficulty, when learning GA, comes from the necessity to *unlearn* the following implicit habits and expectations inherited from complex and linear algebra:

1. Linear transformations act on the left as in $y = \lambda x$ for complex numbers or $\mathbf{y} = \mathbf{A}\mathbf{x}$ for linear algebra.

2. Linear maps $L_\lambda : x \mapsto y = \lambda x$ and $L_{\mathbf{A}} : \mathbf{x} \mapsto \mathbf{y} = \mathbf{A}\mathbf{x}$ are usually *identified* with the complex number λ and the matrix $\mathbf{A}$ *using the same symbol.*

3. λ is an element of the complex algebra acting on complex numbers [a] however the matrix $\mathbf{A}$ is an element, from *outside* the set of vectors, acting on vectors $\mathbf{x}$.

At this point in GA, to avoid ambiguity, it is customary to introduce a notation to distinguish *transforms* from *elements* of the algebra used to *implement* the transform. For example, we have seen in (9.4) that we can implement a skew-symmetric map J acting on a vector $\mathbf{x}$ as a contraction with a bivector J

$$\mathsf{J}(\mathbf{x}) = \mathbf{x} \cdot J = \frac{1}{2}\left(\mathbf{x}J - J\mathbf{x}\right). \tag{9.9}$$

The adjoint map noted J^* is indeed skew-symmetric [b]

$$\mathsf{J}^*(\mathbf{x}) = J \cdot \mathbf{x} = \frac{1}{2}\left(J\mathbf{x} - \mathbf{x}J\right) = -\mathsf{J}(\mathbf{x}). \tag{9.10}$$

Distinguising notations J (map) and J (GA element), it is possible to unambiguously use the common notation AB to denote the composition of maps $\mathsf{A} \circ \mathsf{B}$.

a. In GA, this situation is generalized by the notion of even and odd *spinors*, i.e. elements of GA with even or odd grade, used to represent *transforms* on GA elements.

b. Note that, to avoid confusion between adjoint map and GA dual notations, an alternative notation in the GA literature uses $\underline{J}$ (linear map associated to a symbol J) and $\overline{J}$ (adjoint map) .

Projection

A vector can be decomposed into its projection and rejection $\mathbf{u} = \mathbf{u}_\parallel + \mathbf{u}_\perp$ with respect to a subspace. In GA, a subspace is represented by the blade formed by its spanning vectors (not necessarily orthonormals) $\mathbf{B} = \mathbf{e}_1 \wedge \ldots \wedge \mathbf{e}_k$.

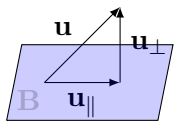


Figure 9.3 – Projection and rejection.

Theorem 9.2 (Projection-rejection [Mac10])**.** *Let* $\mathbf{u}$ *be a vector and* $\mathbf{B}$ *a blade. Then*

$$\mathbf{u}_\parallel = \mathsf{P}_{\mathbf{B}}(\mathbf{u}) := (\mathbf{u} \cdot \mathbf{B})/\mathbf{B}, \qquad \mathbf{u}_\perp = \mathsf{P}^\perp_{\mathbf{B}}(\mathbf{u}) := (\mathbf{u} \wedge \mathbf{B})/\mathbf{B}. \tag{9.11}$$

More generally, if $\mathbf{A}$ is a blade, then the projection of $\mathbf{A}$ on $\mathbf{B}$ is $\mathsf{P}_{\mathbf{B}}(\mathbf{A}) = (\mathbf{A} \cdot \mathbf{B})/\mathbf{B}$. This allows to compute the angle between the subspaces represented by blades $\mathbf{A}$ and $\mathbf{B}$ as ([Mac10], p.123)

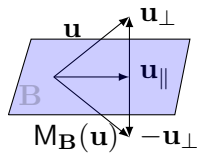
$$\cos\theta = \frac{\left|\mathsf{P}_{\mathbf{B}}(\mathbf{A})\right|}{|\mathbf{A}|} = \frac{|\mathbf{A} \cdot \mathbf{B}|}{|\mathbf{A}||\mathbf{B}|}.$$

Reflection

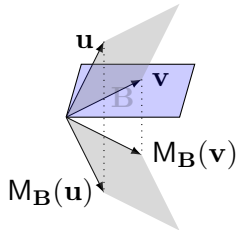
Geometrically, the reflection of a vector $\mathbf{u} = \mathbf{u}_\parallel + \mathbf{u}_\perp$ in a subspace $\mathbf{B}$ is $\mathbf{u}_\parallel - \mathbf{u}_\perp$. From theorem 9.2 and equation 9.5

$$\mathsf{M}_\mathbf{B}(\mathbf{u}) := \mathbf{u}_\parallel - \mathbf{u}_\perp = (\mathbf{u}\cdot\mathbf{B} - \mathbf{u}\wedge\mathbf{B})/\mathbf{B} = (-1)^{k+1}\mathbf{B}\mathbf{u}\mathbf{B}^{-1}.$$

More generally, the following results holds for blades



(a) Mirror of a vector



(b) Mirror of a blade

Theorem 9.3 (Reflection [Mac10])**.** *Let* $\mathbf{B}$ *be a* k*-blade. Then the* reflection *or* mirror *of a vector* $\mathbf{u}$ *into* $\mathbf{B}$ *is*

$$\mathsf{M}_\mathbf{B}(\mathbf{u}) = (-1)^{k+1}\mathbf{B}\mathbf{u}\mathbf{B}^{-1}. \tag{9.12}$$

By extension, the reflection of a ℓ*-blade* $\mathbf{U} = \mathbf{u}_1 \wedge \ldots \wedge \mathbf{u}_\ell$ *defined by* $\mathsf{M}_\mathbf{B}(\mathbf{u}_1) \wedge \ldots \wedge \mathsf{M}_\mathbf{B}(\mathbf{u}_\ell)$ *is*

$$\mathsf{M}_\mathbf{B}(\mathbf{U}) = (-1)^{\ell(k+1)}\mathbf{B}\mathbf{U}\mathbf{B}^{-1}. \tag{9.13}$$

Example 9.1 (Mirror in a line)**.** The reflection of a vector $\mathbf{u}$ in a line represented by $\mathbf{x}$ is

$$\mathbf{M}_\mathbf{x}(\mathbf{u}) = \mathbf{x}\mathbf{u}\mathbf{x}^{-1}. \tag{9.14}$$

When $\mathbf{x}$ has unit norm then $\mathbf{x}^{-1} = \mathbf{x}/|\mathbf{x}| = \mathbf{x}$ so that $\mathbf{M}_\mathbf{x}(\mathbf{u}) = \mathbf{x}\mathbf{u}\mathbf{x}$. One can show (using duality, see def. 9.2) that the reflection in the hyperplane $\mathbf{x}^*$ dual to a vector $\mathbf{x}$ is

$$\mathbf{M}_{\mathbf{x}^*}(\mathbf{u}) = -\mathbf{x}\mathbf{u}\mathbf{x}^{-1}. \tag{9.15}$$

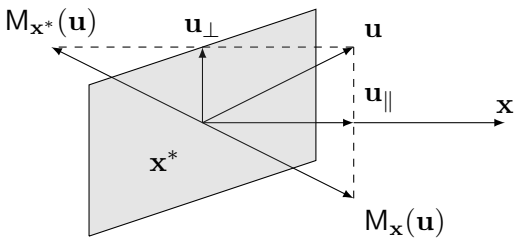


Figure 9.5 – Mirror in a line $\mathbf{x}$ and its dual hyperplane $\mathbf{x}^*$.

Rotation

Geometrically, Let $\mathbf{i}$ be a bivector, the rotation of a vector $\mathbf{u}_\parallel + \mathbf{u}_\perp$ in the plane $\mathbf{i}$ by the bivector angle $\mathbf{i}\theta$ is equal to the sum of its perpendicular component and of its rotated plane projection $\mathsf{R}(\mathbf{u}) = \mathsf{R}(\mathbf{u}_\parallel) + \mathbf{u}_\perp = \mathbf{u}_\parallel e^{\mathbf{i}\theta} + \mathbf{u}_\perp$. Similarly to reflections, one can show ([Mac10] p.89) that rotations can be canonically written using the "sandwich" product [20]

$$\mathsf{R}(\mathbf{u}) = R\mathbf{u}R^{-1},$$

where the *rotor* $R = e^{-\mathbf{i}\theta/2} := \sum_{n=0}^{\infty} \frac{(-\mathbf{i}\theta/2)^n}{n!}$ behaves like a "half-rotation" acting symmetrically on left and right. By extension, we have

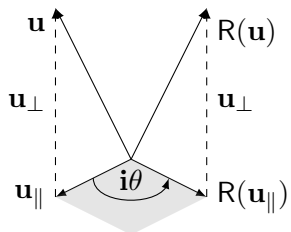


Figure 9.6 – Rotation.

20. In quaternion algebra, transforms are canonically represented using "sandwich products".

Theorem 9.4 (Rotation [Mac10]). *The rotation of a blade* $\mathbf{A} = \mathbf{a}_1 \wedge \ldots \wedge \mathbf{a}_\ell$ *defined by* $\mathsf{R}(\mathbf{A}) = \mathsf{R}(\mathbf{a}_1) \wedge \ldots \wedge \mathsf{R}(\mathbf{a}_\ell)$ *and rotor* R *is*

$$\mathsf{R}(\mathbf{A}) = R\mathbf{A}R^{-1}. \tag{9.16}$$

Unit norm rotors satisfy $R^{-1} = R^\dagger$ and $RR^{-1} = RR^\dagger = \mathbf{1}$ (see (9.7)). Furthermore, let $\mathsf{R}_1, \mathsf{R}_2$ be two rotations defined by rotors R_1, R_2, then composing rotations, we see that rotors form a group whose group composition is $R = R_2 R_1$

$$\mathsf{R}(\mathbf{u}) = (\mathsf{R}_2\mathsf{R}_1)(\mathbf{u}) = R_2\left(R_1\mathbf{u}R_1^{-1}\right)R_2^{-1} = R\mathbf{u}R^{-1}.$$

Rotations as compositions of reflections One can show (see [DGL+03] p.43) that the composition of two reflections in the hyperplanes perpendicular to unit vectors $\mathbf{x}$ and $\mathbf{y}$ yields a rotation by 2θ in the plane $\mathbf{B} = \mathbf{x} \wedge \mathbf{y}$ with $\cos\theta = \mathbf{x} \cdot \mathbf{y}$. It is given (eq. (9.15)) by

$$\mathsf{R}(\mathbf{u}) = (\mathsf{M}_{\mathbf{y}^*}\mathsf{M}_{\mathbf{x}^*})(\mathbf{u}) = (-\mathbf{y}(-\mathbf{x}\mathbf{u}\mathbf{x})\mathbf{y}) = (\mathbf{y}\mathbf{x})\mathbf{u}(\mathbf{x}\mathbf{y}) = R\mathbf{u}R^{-1} \qquad \text{with} \qquad R = \mathbf{y}\mathbf{x}.$$

Dually the composition of reflection in lines $\mathbf{x}, \mathbf{y}$ yields the same rotation $\mathsf{R}(\mathbf{u}) = (\mathsf{M}_{\mathbf{y}}\mathsf{M}_{\mathbf{x}})(\mathbf{u}) = (\mathbf{y}(\mathbf{x}\mathbf{u}\mathbf{x})\mathbf{y}) = (\mathbf{y}\mathbf{x})\mathbf{u}(\mathbf{x}\mathbf{y}) = R\mathbf{u}R^{-1}$.

9.1.2 Sub-algebras

Complex numbers A complex number in $\mathbb{G}^n$ is a multivector of the form $a + \mathbf{i}b$ with $a, b \in \mathbb{R}$ where $\mathbf{i} = \mathbf{a}\mathbf{b}$ is the unit pseudoscalar of *some plane* spanned by orthonormal vectors $\mathbf{a}, \mathbf{b}$. Since $\mathbf{i}^2 = -\mathbf{1}$, this means that every plane has its own complex number system [21] which is isomorphic to $\mathbb{C}$. For two vectors $\mathbf{u}, \mathbf{v} \in \operatorname{span}\{\mathbf{a}, \mathbf{b}\}$, we have the polar representation

$$\mathbf{u}\mathbf{v} = \mathbf{u} \cdot \mathbf{v} + \mathbf{u} \wedge \mathbf{v} = |\mathbf{u}||\mathbf{v}|\,(\cos\theta + \mathbf{i}\sin\theta) = re^{\mathbf{i}\theta},$$

where $r := |\mathbf{u}||\mathbf{v}|$ and $e^{\mathbf{i}\theta} := \cos\theta + \mathbf{i}\sin\theta$. The geometric interpretation of the quantity $|\mathbf{u}||\mathbf{v}|\sin\theta$ is shown in figure 9.7 (the red and blue oriented areas are equal to each other).

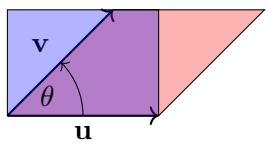


Figure 9.7 – GA identity $|\mathbf{u} \wedge \mathbf{v}| = |\mathbf{u}||\mathbf{v}|\sin\theta$.

Quaternions In $\mathbb{G}^3$, the bivectors squares to $-\mathbf{1}$. Defining $\mathbf{i} = -\mathbf{e}_2\mathbf{e}_3$, $\mathbf{j} = -\mathbf{e}_3\mathbf{e}_1$, $\mathbf{k} = -\mathbf{e}_1\mathbf{e}_2$, we obtain Hamilton's equation defining quaternions

$$\mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = \mathbf{i}\mathbf{j}\mathbf{k} = -\mathbf{1}.$$

We see by looking at table 9.2 that the bivectors $\mathbf{i}, \mathbf{j}, \mathbf{k}$ are duals of the vectors $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ (indeed multiplication of $\mathbf{i}, \mathbf{j}, \mathbf{k}$ by the pseudo-scalar $\mathbf{I}_3 = \mathbf{e}_1\mathbf{e}_2\mathbf{e}_3$ of $\mathbb{G}^3$ yields respectively $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$). Exactly like in $\mathbb{G}^2$ we can represent a vector either using the vectors basis $\{\mathbf{e}_1, \mathbf{e}_2\}$ (odd subalgebra) or the complex basis $\{\mathbf{1}, \mathbf{i} = \mathbf{e}_1\mathbf{e}_2\}$ (even subalgebra), in $\mathbb{G}^3$ we can represent vectors either using the vectors $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ or their duals $\{\mathbf{i}, \mathbf{j}, \mathbf{k}\}$.

21. We have $\mathbf{i}^2 = \mathbf{a}\mathbf{b}\mathbf{a}\mathbf{b} = -\mathbf{a}\mathbf{a}\mathbf{b}\mathbf{b} = -\mathbf{a}^2\mathbf{b}^2 = -\mathbf{1}$ (using anticommutation $\mathbf{a}\mathbf{b} = -\mathbf{b}\mathbf{a}$ of orthogonal vectors, and the metric signature $\mathbf{a}^2 = \mathbf{b}^2 = \mathbf{1}$). See table 9.2 to verify that unit bivectors squares to $-\mathbf{1}$ (in euclidean space).

Matrix isomorphisms

$\mathbb{G}^2$ There exists an *"accidental" isomorphism* between $\mathbb{R}^{2\times 2}$ and $\mathbb{G}^2$ given by

$$[\mathbf{1}] = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad [\mathbf{e}_1] = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \qquad [\mathbf{e}_2] = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \qquad [\mathbf{e}_1\mathbf{e}_2] = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}.$$

Indeed, identifying the geometric product with the matrix product, we can verify that

$$[\mathbf{e}_1]^2 = [\mathbf{e}_1^2] = [\mathbf{1}], \qquad [\mathbf{e}_2]^2 = [\mathbf{e}_2^2] = [\mathbf{1}], \qquad [\mathbf{e}_1][\mathbf{e}_2] = [\mathbf{e}_1\mathbf{e}_2].$$

so that matrices $\big\{[\mathbf{1}], [\mathbf{e}_1], [\mathbf{e}_2], [\mathbf{e}_1\mathbf{e}_2]\big\}$ satisfy the GA properties from definition 9.1 (see also the multiplication table 9.2a).

$\mathbb{G}^3$ There exists similar embeddings (see [Sob08, Sob20]) for $\mathbb{G}^3$, (which is of dimension $2^3 = 8$). However it requires a matrix embedding as a sub-algebra of either $\mathbb{R}^{4\times 4}$ or $\mathbb{C}^{2\times 2}$ (of dimension 16). The most famous one is the algebra generated by Pauli matrices

$$[\mathbf{e}_1] = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \qquad [\mathbf{e}_2] = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \qquad [\mathbf{e}_3] = \begin{bmatrix} 0 & i \\ -i & 0 \end{bmatrix}.$$

One can verify that $[\mathbf{e}_i]^2 = [\mathbf{1}]$, that we have the bivectors

$$[\mathbf{B}_1] = [\mathbf{e}_2\mathbf{e}_3] = \begin{bmatrix} 0 & i \\ i & 0 \end{bmatrix}, \qquad [\mathbf{B}_2] = [\mathbf{e}_3\mathbf{e}_1] = \begin{bmatrix} i & 0 \\ 0 & -i \end{bmatrix}, \qquad [\mathbf{B}_3] = [\mathbf{e}_1\mathbf{e}_2] = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix},$$

and that the pseudo scalar of the algebra is

$$[\mathbf{I}] = [\mathbf{e}_1\mathbf{e}_2\mathbf{e}_3] = \begin{bmatrix} i & 0 \\ 0 & i \end{bmatrix},$$

so that we have the duality relation between vectors and bivectors $[\mathbf{B}_i] = [\mathbf{I}]\mathbf{e}_i$ (see definition 9.2 and the multiplication table 9.2b).

9.2 Motivating examples and invariants

In this section, we review short motivating examples (the harmonic oscillator, a dissipative oscillator, and Maxwell equations). This shows the potential of GA for revealing hidden geometric structure, unifying and simplifying representations.

9.2.1 Harmonic oscillator

Consider a linear harmonic oscillator with unit mass and pulsation ω

$$\begin{bmatrix}\dot{x}\\ \dot{y}\end{bmatrix} = \mathbf{J}\begin{bmatrix}\partial_x H\\ \partial_y H\end{bmatrix}, \qquad \text{with} \qquad \mathbf{J} = \omega\begin{bmatrix}0 & 1\\ -1 & 0\end{bmatrix}, \qquad \text{and} \qquad H(x,y) = \frac{1}{2}\left(x^2+y^2\right). \tag{9.17}$$

Using the geometric algebra $\mathcal{G}(\mathbb{R}^2)$, in a basis $\{\mathbf{e}_1, \mathbf{e}_2\}$, it can be written as

$$\dot{\boldsymbol{x}} = \boldsymbol{J}\cdot\boldsymbol{\nabla}H(\boldsymbol{x}), \qquad \text{with} \qquad H(\mathbf{x}) = \frac{1}{2}\mathbf{x}^2, \tag{9.18}$$

where the vector $\mathbf{x}$, the gradient operator $\boldsymbol{\nabla}$ and the bivector $\boldsymbol{J}$ (see (9.10)) are represented as

$$\mathbf{x} = \mathbf{e}_1 x + \mathbf{e}_2 y, \qquad \boldsymbol{\nabla} = \mathbf{e}_1\frac{\partial}{\partial x} + \mathbf{e}_2\frac{\partial}{\partial y}, \qquad \boldsymbol{J} = \boldsymbol{i}\omega, \qquad \boldsymbol{i} := \mathbf{e}_1\mathbf{e}_2 = \mathbf{e}_1\wedge\mathbf{e}_2. \tag{9.19}$$

The system has two invariants: the energy E and the angular momentum $\boldsymbol{L}$ defined by

$$E(t) = H(\boldsymbol{x}(t)), \qquad \boldsymbol{L}(t) = \boldsymbol{x}(t)\wedge\dot{\boldsymbol{x}}(t). \tag{9.20}$$

Angular momentum To have a geometric interpretation of the angular momentum (see figure 9.8), denote $\mathrm{d}\boldsymbol{A}$ the infinitesimal oriented area swept from $\boldsymbol{x}(t)$ to $\boldsymbol{x}(t+\mathrm{d}t)$

$$\mathrm{d}\boldsymbol{A} := \boldsymbol{x}(t)\wedge\boldsymbol{x}(t+\mathrm{d}t). \tag{9.21}$$

Then, using (9.20) and (9.21), we see how the angular momentum $\boldsymbol{L}(t)$ quantifies the rate of change of $\boldsymbol{A}$ along the trajectory

$$\boldsymbol{L}(t) = \boldsymbol{x}(t)\wedge\dot{\boldsymbol{x}}(t) = \boldsymbol{x}(t)\wedge\left(\frac{\boldsymbol{x}(t+\mathrm{d}t)-\boldsymbol{x}(t)}{\mathrm{d}t}\right) = \frac{\boldsymbol{x}(t)\wedge\boldsymbol{x}(t+\mathrm{d}t)}{\mathrm{d}t} = \frac{\mathrm{d}\boldsymbol{A}}{\mathrm{d}t}. \tag{9.22}$$

Figure 9.8 shows the geometrical interpretations of $\boldsymbol{L}$ and $\mathrm{d}\boldsymbol{A}$.

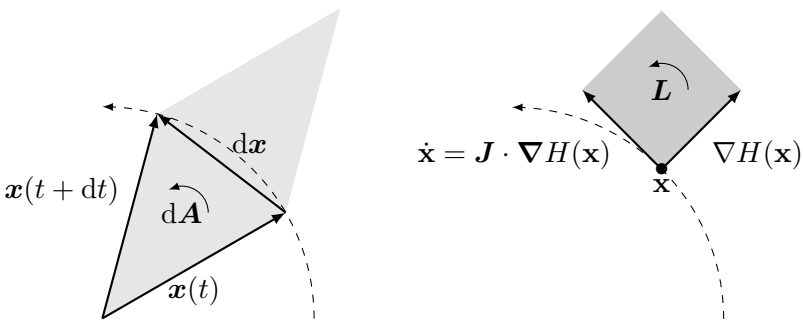


Figure 9.8 – Angular momentum of an harmonic oscillator. Note that according to (9.22), the angular momentum $\boldsymbol{L}$ and the infinitesimal area $\mathrm{d}\boldsymbol{A}$ are linked through $\mathrm{d}\boldsymbol{A} = \boldsymbol{L}\,\mathrm{d}t$.

Power To obtain a geometric insight on the power, we consider the time-derivative of the energy. Using the chain rule, we recover (as expected) the GA definition of the inner product

$$\frac{\mathrm{d}E}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\boldsymbol{x}^2}{2}\right) = \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\boldsymbol{x}\boldsymbol{x}}{2}\right) = \frac{1}{2}\left(\boldsymbol{x}\dot{\boldsymbol{x}} + \dot{\boldsymbol{x}}\boldsymbol{x}\right) = \boldsymbol{x}\cdot\dot{\boldsymbol{x}} = \boldsymbol{x}\cdot\boldsymbol{J}\cdot\boldsymbol{x} = 0, \tag{9.23}$$

which vanishes by orthogonality of $\boldsymbol{x}$ and $\dot{\boldsymbol{x}}$ (thanks to skew-symmetry of bivector $\boldsymbol{J}$).

Unification of power and angular momentum We remark that power involves the inner product, while momentum is linked to the exterior product. Using GA, we can unify (9.23) and (9.22) as the direct sum of a scalar and of a bivector using the the geometric product

$$\frac{\mathrm{d}E}{\mathrm{d}t} + \frac{\mathrm{d}\boldsymbol{A}}{\mathrm{d}t} = \boldsymbol{x}\cdot\dot{\boldsymbol{x}} + \boldsymbol{x}\wedge\dot{\boldsymbol{x}} = \boldsymbol{x}\dot{\boldsymbol{x}}. \tag{9.24}$$

> **Remark 9.4** (Multi-vector potential)**.** In equation (9.24), we notice the emergence of the time derivative of a *multivector*. This suggests that the *multivector functional*
>
> $$M(\boldsymbol{x}) = \int_0^t \boldsymbol{x}(t)\cdot\dot{\boldsymbol{x}}(t)\,\mathrm{d}t + \int_0^t \boldsymbol{x}(t)\wedge\dot{\boldsymbol{x}}(t)\,\mathrm{d}t = \int_{\boldsymbol{x}(0)}^{\boldsymbol{x}(t)} \mathbf{x}\cdot\mathrm{d}\mathbf{x} + \int_{\boldsymbol{x}(0)}^{\boldsymbol{x}(t)} \mathbf{x}\wedge\mathrm{d}\mathbf{x} = \int_{\boldsymbol{x}(0)}^{\boldsymbol{x}(t)} \mathbf{x}\,\mathrm{d}\mathbf{x},$$
>
> that is $M = E + \boldsymbol{A}$, plays an important role in the formulation of the dynamic.

In the conservative case, by orthogonality, the energy variation is zero ($\dot{E} = \boldsymbol{x}\cdot\dot{\boldsymbol{x}} = 0$), the energy $H(\mathbf{x})$ is thus constant and we have $\boldsymbol{x}\dot{\boldsymbol{x}} = \boldsymbol{x}\wedge\dot{\boldsymbol{x}}$. Furthermore, we can show that momentum and energy are proportional

$$\boldsymbol{L} = 2\boldsymbol{J}H(\mathbf{x}), \tag{9.25}$$

so that the momentum $\boldsymbol{L}$ is constant too.

Proof. Since $\boldsymbol{x} = \mathbf{e}_1 x + \mathbf{e}_2 y$, and $\dot{\boldsymbol{x}} = \boldsymbol{J}\cdot\boldsymbol{\nabla}H(\mathbf{x}) = \omega(-\mathbf{e}_1 y + \mathbf{e}_2 x)$, using the multiplication table of $\mathbb{G}^2$ from table 9.2 we show that

$$\begin{aligned}\boldsymbol{L} = \boldsymbol{x}\wedge\dot{\boldsymbol{x}} = \boldsymbol{x}\dot{\boldsymbol{x}} = (\mathbf{e}_1 x + \mathbf{e}_2 y)\omega(-\mathbf{e}_1 y + \mathbf{e}_2 x) &= \omega\left(\mathbf{e}_1\mathbf{e}_1 xy - \mathbf{e}_2\mathbf{e}_2 xy + \mathbf{e}_1\mathbf{e}_2 x^2 - \mathbf{e}_2\mathbf{e}_1 y^2\right)\\ &= \omega\left(\mathbf{1}(xy - xy) + \boldsymbol{i}(x^2 + y^2)\right) = 2\boldsymbol{J}H(\mathbf{x}).\end{aligned}$$

Alternatively (using the grade operator), $\boldsymbol{L} = \boldsymbol{x}(\boldsymbol{J}\cdot\boldsymbol{x}) = \langle\boldsymbol{x}\boldsymbol{J}\boldsymbol{x}\rangle_2 = \left\langle\boldsymbol{J}\boldsymbol{x}^2\right\rangle_2 = 2\boldsymbol{J}H(\mathbf{x})$. □

9.2.2 Dissipative oscillator

We introduce dissipation to obtain the following logarithmic spiral oscillator (see figure 9.9)

$$\begin{bmatrix}\dot{x}\\ \dot{y}\end{bmatrix} = \left(\omega\begin{bmatrix}0 & 1\\ -1 & 0\end{bmatrix} - \sigma\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}\right)\begin{bmatrix}x\\ y\end{bmatrix}. \tag{9.26}$$

An equivalent geometric algebra formulation is given by

$$\dot{\boldsymbol{x}} = (\boldsymbol{J} - \boldsymbol{R})\cdot\nabla H(\boldsymbol{x}) \qquad \text{with} \qquad \boldsymbol{J} = \boldsymbol{i}\omega, \qquad \boldsymbol{R} = \sigma\mathbf{1}. \tag{9.27}$$

The system is no longer conservative but it still has two constants of motion, which are the relative dissipation rate (prop. to σ) and the relative angular momentum (prop. to ω)

$$\frac{\dot{H}}{H} = -2\sigma, \qquad \frac{\boldsymbol{L}}{H} = -2\boldsymbol{i}\omega. \tag{9.28}$$

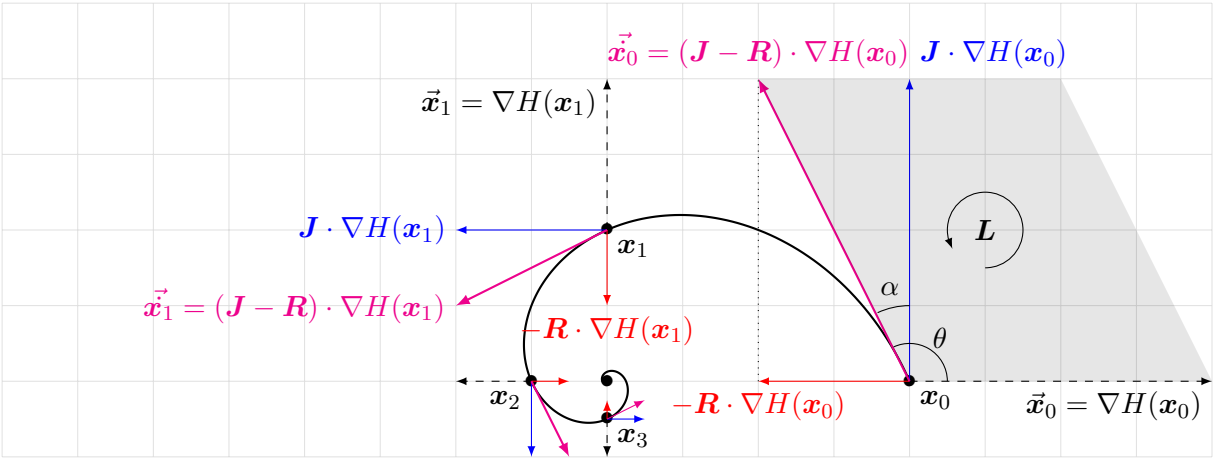


Figure 9.9 – Bernouilli's logarithmic spiral.

Proof. Recall that $\boldsymbol{i} = \mathbf{e}_1\mathbf{e}_2$ and $\boldsymbol{x}^2 = 2H(\boldsymbol{x}) = x^2 + y^2$. Then,

$$|\dot{\boldsymbol{x}}|^2 = |(-\sigma\mathbf{1} + \omega\boldsymbol{i})\mathbf{x}|^2 = |\sigma\boldsymbol{x}|^2 + |\omega\boldsymbol{i}\boldsymbol{x}|^2 = |\boldsymbol{x}|^2(\sigma^2 + \omega^2).$$

We also have the polar decomposition of the geometric product

$$\nabla H(\boldsymbol{x})\dot{\boldsymbol{x}} = \nabla H(\boldsymbol{x}) \cdot \dot{\boldsymbol{x}} + \nabla H(\boldsymbol{x}) \wedge \dot{\boldsymbol{x}} = |\boldsymbol{x}||\dot{\boldsymbol{x}}|(\mathbf{1}\cos\theta + \boldsymbol{i}\sin\theta)$$

where

$$\mathbf{1}\cos\theta := \frac{\boldsymbol{x} \cdot \dot{\boldsymbol{x}}}{|\boldsymbol{x}||\dot{\boldsymbol{x}}|} = \mathbf{1}\frac{-\sigma}{\sqrt{\sigma^2 + \omega^2}}, \qquad \boldsymbol{i}\sin\theta := \frac{\boldsymbol{x} \wedge \dot{\boldsymbol{x}}}{|\boldsymbol{x}||\dot{\boldsymbol{x}}|} = \boldsymbol{i}\frac{-\omega}{\sqrt{\sigma^2 + \omega^2}}.$$

Then, using the left identity for $\cos\theta$, the relative dissipation rate is

$$\frac{\mathrm{d}}{\mathrm{d}t}\ln H(\boldsymbol{x}) = \frac{\dot{H}(\boldsymbol{x})}{H(\boldsymbol{x})} = 2\frac{\boldsymbol{x} \cdot \dot{\boldsymbol{x}}}{\boldsymbol{x} \cdot \boldsymbol{x}} = 2\frac{|\boldsymbol{x}||\dot{\boldsymbol{x}}|}{|\boldsymbol{x}|^2}\mathbf{1}\cos\theta = 2\frac{|\dot{\boldsymbol{x}}|}{|\boldsymbol{x}|}\mathbf{1}\cos\theta = 2\sqrt{\sigma^2 + \omega^2}\mathbf{1}\cos\theta = -2\sigma.$$

Likewise, using the right identity for $\sin\theta$, the relative momentum is

$$\frac{\boldsymbol{L}(\boldsymbol{x})}{H(\boldsymbol{x})} = \frac{\dot{\boldsymbol{A}}}{H(\boldsymbol{x})} = 2\frac{\boldsymbol{x} \wedge \dot{\boldsymbol{x}}}{\boldsymbol{x} \cdot \boldsymbol{x}} = 2\frac{|\boldsymbol{x}||\dot{\boldsymbol{x}}|}{|\boldsymbol{x}|^2}\boldsymbol{i}\sin\theta = 2\frac{|\dot{\boldsymbol{x}}|}{|\boldsymbol{x}|}\boldsymbol{i}\sin\theta = 2\sqrt{\sigma^2 + \omega^2}\boldsymbol{i}\sin\theta = -2\boldsymbol{i}\omega.$$

Dividing both expressions, we obtain the *dissipation angle* θ given by

$$\frac{\dot{\boldsymbol{A}}}{\dot{E}} = \boldsymbol{i}\tan\theta = \boldsymbol{i}\left(\frac{\omega}{\sigma}\right). \tag{9.29}$$

□

9.2.3 Maxwell equations (in empty space)

As a last example, due to [Mac17, eq. (3.1)], one can show, using GA, that Maxwell equations can be elegantly unified as an instance of the wave equation $\partial_t^2 F = \boldsymbol{\nabla}^2 F$ over a *multivector field* $F(t, x, y, z) \in \mathbb{G}^3$. The derivation, not directly relevant to this thesis, is reproduced in appendix F.3 p.322. We mention this example to highlight the kind of paradigm shift that can be expected from adequate use of Geometric Algebra.

9.3 Port-Hamiltonian systems using Geometric Algebra

A direct translation in GA of an input-state-output port Hamiltonian system (see definition 1.22 p.33) is given by

$$\begin{cases} \dot{\mathbf{x}} = (\mathsf{J} - \mathsf{R})(\nabla H(\mathbf{x})) + \mathsf{G}(\mathbf{u}) \\ \mathbf{y} = \mathsf{G}^*(\nabla H(\mathbf{x})) \end{cases} \tag{9.30}$$

with *vectors* $\mathbf{u}, \mathbf{y} \in \mathbb{R}^p$, $\mathbf{x} \in \mathbb{R}^n$, and *linear maps* $\mathsf{J} \in L(\mathbb{R}^n, \mathbb{R}^n)$, $\mathsf{R} \in L(\mathbb{R}^n, \mathbb{R}^n)$, $\mathsf{G} \in L(\mathbb{R}^p, \mathbb{R}^n)$, satisfying skew-symmetry $\mathsf{J}^* = -\mathsf{J}$, and $\mathsf{R}^* = \mathsf{R} \succeq 0$. The Hamiltonian is $H \in \mathcal{C}^2(\mathbb{R}^n, \mathbb{R})$, the gradient is $\nabla = \sum_n \mathbf{e}_n \frac{\partial}{\partial x_n}$ expressed in the canonical basis $\{\mathbf{e}_k\}$ of $\mathbb{R}^n$ such that vectors are written as $\mathbf{x} = \sum_k \mathbf{e}_k x_k$.

At this point, nothing has changed, we only abstracted matrices by their linear maps. In this section we are interested in the *implementation* of the linear maps $\mathsf{J}, \mathsf{R}, \mathsf{G}, \mathsf{G}^*$ using elements of geometric algebra (exclusively) instead of their matrix representation.

Implementation of skew-symmetric maps

We have already seen during the GA introduction in (9.9) that skew-symmetric maps J can be implemented as a contraction with a bivector J so that (choosing right contraction) in GA [22]

$$\mathsf{J}(\nabla H(\mathbf{x})) = J \cdot \nabla H(\mathbf{x}) = \frac{1}{2}\left(J\nabla H(\mathbf{x}) - \nabla H(\mathbf{x})J\right). \tag{9.31}$$

Implementation of symmetric positive definite maps

Using a linear algebra argument, a possible strategy to implement a symmetric positive (semi-)definite map $\mathsf{R} = \mathsf{R}^* \succeq 0$ is to use its eigenvalue decomposition $\mathsf{R} = \mathsf{Q}\Lambda\mathsf{Q}^*$ with (real) orthonormal eigenvectors $\{\mathbf{q}_i\}_{i=1}^r$ $(r \leq n)$ and corresponding (positive real) eigenvalues $\{\lambda_i\}$. This way, for a vector $\mathbf{v}$, projecting on $\mathbf{q}_i$, scaling by λ_i and synthesising on $\mathbf{q}_i$ we obtain

$$\begin{aligned} \mathsf{R}(\mathbf{v}) = \sum_{i=1}^r \mathbf{q}_i\lambda_i(\mathbf{q}_i \cdot \mathbf{v}) &\overset{a}{=} \sum_{i=1}^r \mathbf{q}_i\lambda_i\left(\frac{\mathbf{q}_i\mathbf{v} + \mathbf{v}\mathbf{q}_i}{2}\right) \overset{b}{=} \sum_{i=1}^r \lambda_i\left(\frac{\mathbf{v} + \mathbf{q}_i\mathbf{v}\mathbf{q}_i}{2}\right) \overset{c}{=} \sum_{i=1}^r \lambda_i\left(\frac{\mathbf{v} + \mathsf{M}_{\mathbf{q}_i}(\mathbf{v})}{2}\right) \\ &\overset{d}{=} \sum_{i=1}^r \lambda_i \mathsf{P}_{\mathbf{q}_i}(\mathbf{v}). \end{aligned}$$

Where we used a) the definition of the inner product (9.1), b) the signature of euclidean vectors $\mathbf{q}_i^2 = 1$, c) the definition of reflection in a unit line (9.14) and d) the definition of projection on a unit vector (9.11). This representation explicitly emphasises that every SPD transforms determines a scaling in the direction of its eigenvectors (but requires that we know them).

Implementation of non square linear maps

To implement maps $\mathsf{G} \in L(\mathbb{R}^p, \mathbb{R}^n)$ (and their dual G^*) a similar approach is to use the singular value decomposition $\mathsf{G} = \mathsf{U}\Sigma\mathsf{V}^*$, with left and right eigenvector $\{\boldsymbol{\mu}_i\}$, $\{\boldsymbol{\nu}_i\}$ and corresponding singular values $\{\sigma_i\}_{i=1}^r$, with $r \leq \min(p, n)$ so that for a vector $\mathbf{x}$

$$\mathsf{G}(\mathbf{x}) = \sum_{i=1}^r \boldsymbol{\mu}_i\, \sigma_i(\boldsymbol{\nu}_i \cdot \mathbf{x})$$

22. Note the skew-symmetric similarity with the results from [Cel] and [MQR99, eq. (1.3)]: for a vector field $\boldsymbol{f}(\mathbf{x})$ with nonzero invariant $H(\mathbf{x})$, the structure matrix of the ODE is $\mathbf{J}(\mathbf{x}) = \frac{1}{2}\left(\boldsymbol{f}(\mathbf{x})\frac{\nabla H(\mathbf{x})^\mathsf{T}}{\|\nabla H\|^2} - \frac{\nabla H(\mathbf{x})}{\|\nabla H\|^2}\boldsymbol{f}(\mathbf{x})^\mathsf{T}\right)$.

Since the spaces $\mathbb{R}^n$ and $\mathbb{R}^p$ are distinct, we cannot say more about its geometric interpretation unless we embed both spaces in a larger space in which we can establish relations.

9.3.1 Going further: unifying transforms

So far we have managed to represent the geometric transform that we needed for PHS modelling and to obtain some geometric interpretation. However, compared to the simplicity of matrix linear algebra, this is still not sufficient: we had to use different patterns and strategies for each type of transform, we lack a unifying framework. An elegant solution to this problem has been proposed in [DHSVA93], which states that *every Lie algebra can be represented as a bivector algebra; hence every Lie group can be represented as a spin group.* The general idea is the following: one can represent general linear transforms

$$\mathsf{A} : \mathbb{R}^n \mapsto \mathbb{R}^n, \quad \mathbf{x} \mapsto \mathbf{y} \ \in \mathsf{GL}(n, \mathbb{R}),$$

by representing a vector $\mathbf{x} \in \mathbb{R}^n$ by its image $\vec{\mathbf{x}}$ in $\mathbb{R}^{n,n}$. This is obtained by using an embedding map $\phi : \mathbb{R}^n \to \mathbb{R}^{n,n}$ as

$$\vec{\mathbf{x}} = \phi(\mathbf{x}) \in \mathbb{R}^{n,n} \subset \mathcal{G}(\mathbb{R}^{n,n}).$$

The inverse operation is obtained through a projection $\pi : \mathbb{R}^{n,n} \to \mathbb{R}^n$ such that π is a left inverse of ϕ. In other words, $\pi \circ \phi = \mathbf{I}_{\mathbb{R}_n}$ so that $\pi = \phi^{-1}$.

The reason is that, in the space $\mathbb{R}^{n,n}$, we can represent the image $\vec{\mathsf{A}}$ of *any linear transform* A on $\mathbb{R}^n$ by an orthogonal transform implemented by a spinor [23] $R \in \mathsf{Spin}(n, n) \subset \mathcal{G}(\mathbb{R}^{n,n})$ as

$$\vec{\mathsf{A}} : \vec{\mathbf{x}} \mapsto \vec{\mathbf{y}} = \mathsf{R}(\vec{\mathbf{x}}) = R\,\vec{\mathbf{x}}\,R^{-1}, \tag{9.32}$$

(see (9.16) and (9.12) for the GA definition of orthogonal transforms (i.e. rotations and reflections) using spinors). Going back to $\mathbb{R}^n$, the transform A is realised by

$$\mathsf{A} : \mathbf{x} \mapsto \mathbf{y} = \phi^{-1}\left(\vec{\mathsf{A}}\left(\phi\left(\mathbf{x}\right)\right)\right). \tag{9.33}$$

It can be summarised by the commutation diagram

$$\begin{array}{ccc} \mathbf{x} \in \mathbb{R}^n & \xrightarrow{\ \mathsf{A}\ } & \mathbf{y} \in \mathbb{R}^n \\ \phi \Big\downarrow & & \Big\uparrow \pi = \phi^{-1} \\ \vec{\mathbf{x}} \in \mathbb{R}^{n,n} & \xrightarrow{\ \vec{\mathsf{A}}\ } & \vec{\mathbf{y}} \in \mathbb{R}^{n,n} \end{array} \tag{9.34}$$

The "beauty" in this approach is unification: every linear transform becomes an orthogonal transform, unifying for example rotations (bivectors squaring to $-\mathbf{1}$) and hyperbolic rotations (bivectors squaring to $\mathbf{1}$). We do not have enough space to develop this path further and point to the main reference [DHSVA93], see also the book [DGL+03, ch.11].

Coincidently, we note that in chapter 1, 1.3.1 p.18, on Dirac structures (and subsequently in 1.4 35 on wave variables), we also had to work using the indefinite metric of the space $\mathbb{R}^{n,n}$ to encode the duality of effort and flow spaces (with the consequence that incident and reflected wave spaces corresponds respectively to the positive and negative polarisations of $\mathbb{R}^{n,n}$).

Note that the space $\mathcal{G}(\mathbb{R}^{n,n})$ and its geometry will be explored further in the next sections, where we revisit flow-effort spaces, incident-reflected wave spaces and Dirac structures with the tools of geometric algebra.

23. An even (resp. odd) spinor is a GA element whose (multivector) components are of even (resp. odd) grade.

9.4 Representing Dirac structures with Geometric Algebra

In this section, we revisit the representation of Dirac structures, a cornerstone of PHS modelling (defined in subsection 1.3.1 p.18), using Geometric Algebra (see [Hes93] for Hamiltonian Mechanics). We have already seen in subsection 1.4.2 p.36 that the natural geometry of the bond space $\mathcal{B} = \mathcal{F} \oplus \mathcal{E}$ is that of an indefinite inner product space identifiable with $\mathbb{R}^{n,n}$ (i.e. its metric is not positive definite, see def. C.14 p.283). It can be separated either into euclidean and anti-euclidean wave subspaces $\mathcal{W}^+, \mathcal{W}^-$ or into dual flow and effort spaces $\mathcal{F}, \mathcal{E}$. Our goal here is to: 1) exhibit bases of these subspaces, 2) choose the respective metric signatures so that the geometric product regenerates the quadratic form Q from equation (1.22) p.19 and the associated bilinear form $\langle\langle \cdot, \cdot \rangle\rangle$ defined in (1.23) p.19, 3) extend the bond space $\mathcal{B} \sim \mathbb{R}^{n,n}$ to the geometric algebra $\mathbb{G}^{n,n} = \mathcal{G}(\mathbb{R}^{n,n})$ and formulate Dirac structures in $\mathbb{G}^{n,n}$ [24], 4) show that GA can simplify the results and definitions on Dirac structures from subsection 1.3.1 p.18 thanks to its ability to multiply vectors.

Wave spaces as pseudo-euclidean subspaces of $\mathbb{R}^{n,n}$ Following subsection 1.4.2, let $\mathcal{W}^+ = \mathbb{R}^{n,0}$ be an euclidean space of incident wave vectors with basis $\{\mathbf{a}_i\}_{i=1}^n$ and $\mathcal{W}^- = \mathbb{R}^{0,n}$ an anti-euclidean spaces of reflected wave vectors with basis $\{\mathbf{b}_i\}_{i=1}^n$ such that $\mathcal{B} = \mathcal{W}^+ \oplus \mathcal{W}^- = \mathbb{R}^{n,n}$. Any element of $\mathcal{B}$ can be represented as $\mathbf{x} = \mathbf{a} + \mathbf{b}$, with $\mathbf{a} \in \mathcal{W}^+$ and $\mathbf{b} \in \mathcal{W}^-$ where

$$\mathbf{a} = \sum_{i=1}^n \mathbf{a}_i a_i, \qquad \mathbf{b} = \sum_{i=1}^n \mathbf{b}_i b_i, \tag{9.35}$$

and where the basis vectors have the metric signature $\begin{bmatrix} \mathbf{I} & 0 \\ 0 & -\mathbf{I} \end{bmatrix}$ (see also [DHSVA93, eq.3.17]) i.e.

$$\mathbf{a}_i \cdot \mathbf{a}_j = \delta_{ij}, \qquad \mathbf{a}_i \cdot \mathbf{b}_j = 0, \qquad \mathbf{b}_i \cdot \mathbf{b}_j = -\delta_{ij}. \tag{9.36}$$

Using the language of geometric algebra, we see that $\mathcal{W}^+$ is a *positive Euclidean space* since its basis vectors square to one ($\mathbf{a}_i^2 = 1$), while $\mathcal{W}^-$ is a *negative Euclidean space* whose basis vector square to minus one ($\mathbf{b}_i^2 = -1$). Note that the metric encodes the sign of waves.

Dual flow and effort spaces as null spaces of $\mathbb{R}^{n,n}$ Define the flow space $\mathcal{F}$ with basis $\{\mathbf{f}_i\}_{i=1}^n$ and its dual, the effort space $\mathcal{E}$, with basis $\{\mathbf{e}_i\}_{i=1}^n$, through the change of basis

$$\mathbf{e}_i = \frac{\mathbf{a}_i + \mathbf{b}_i}{\sqrt{2}}, \qquad \mathbf{f}_i = \frac{\mathbf{a}_i - \mathbf{b}_i}{\sqrt{2}}, \tag{9.37}$$

so that any element of $\mathcal{B} = \mathcal{E} \times \mathcal{F}$ can be (alternatively) represented as $\mathbf{x} = \mathbf{f} + \mathbf{e}$ where

$$\mathbf{f} = \sum_{i=1}^n \mathbf{f}_i f_i, \qquad \mathbf{e} = \sum_{i=1}^n \mathbf{e}_i e_i. \tag{9.38}$$

In this basis, we have the following metric signature, encoding the duality of the subspaces $\mathcal{E}, \mathcal{F}$

$$\mathbf{e}_i \cdot \mathbf{e}_j = 0, \qquad \mathbf{e}_i \cdot \mathbf{f}_j = \delta_{ij}, \qquad \mathbf{f}_i \cdot \mathbf{f}_j = 0. \tag{9.39}$$

The subspaces $\mathcal{E}, \mathcal{F}$ are said to be *null spaces* [25] and vectors $\mathbf{e}, \mathbf{f}$ are said to be *null vectors* [PS02]. Indeed, using geometric algebra, one easily finds that their basis vectors all *square to zero* ($\mathbf{e}_i^2 = 0 = \mathbf{f}_i^2$). This is also called a *Witt basis* [PS02], [DHSVA93, p.8].

24. We note that $\mathbb{R}^{n,n}$, its null spaces and the geometric algebra $\mathcal{G}(\mathbb{R}^{n,n})$ also play important roles in [PS02] to represent matrix transforms and more generally to represent Lie groups as spin groups in [DHSVA93].

25. Not to be confused with the nullspace of an operator.

Proof. Using the metric signature (9.36) of $\mathcal{W}^+, \mathcal{W}^-$, in the variable change (9.37), we have

$$\begin{aligned}
\mathbf{e}_i \cdot \mathbf{e}_j &= \frac{1}{2}(\mathbf{a}_i + \mathbf{b}_i) \cdot (\mathbf{a}_i + \mathbf{b}_i) = \frac{1}{2}(\delta_{ij} + (-\delta_{ij})) = 0, \\
\mathbf{e}_i \cdot \mathbf{f}_j &= \frac{1}{2}(\mathbf{a}_i + \mathbf{b}_i) \cdot (\mathbf{a}_i - \mathbf{b}_i) = \frac{1}{2}(\delta_{ij} - (-\delta_{ij})) = \delta_{ij}, \\
\mathbf{f}_i \cdot \mathbf{f}_j &= \frac{1}{2}(\mathbf{a}_i - \mathbf{b}_i) \cdot (\mathbf{a}_i - \mathbf{b}_i) = \frac{1}{2}(\delta_{ij} + (-\delta_{ij})) = 0.
\end{aligned}$$

□

Quadratic form and power Let $\mathbf{x}$ be an element of the bondspace $\mathcal{B}$. To replicate the quadratic form defined in equation (1.22) p.19, we define

$$Q(\mathbf{x}) := \mathbf{x}^2. \tag{9.40}$$

a) *Flow and effort decomposition*: Consider an element $\mathbf{x} = \mathbf{e} + \mathbf{f}$ with $\mathbf{e} \in \mathcal{E}$ and $\mathbf{f} \in \mathcal{F}$. Then using the metric signature from equation (9.39) and the definition of the inner product of two vectors (9.1), we recover that the quadratic form represents power through the duality product of flow and efforts (see equation (1.22)).

$$Q(\mathbf{x}) = (\mathbf{e} + \mathbf{f})^2 = \underbrace{\mathbf{e}^2}_{=0} + \mathbf{ef} + \mathbf{fe} + \underbrace{\mathbf{f}^2}_{=0} = 2\,\mathbf{e} \cdot \mathbf{f}.$$

b) *Wave decomposition*: Consider an element $\mathbf{x} = \mathbf{a} + \mathbf{b}$, with $\mathbf{a} \in \mathcal{W}^+$, and $\mathbf{b} \in \mathcal{W}^-$. Then, from the metric (9.36), we recover that power is proportional to the difference between the squared (Euclidean) norms of incident and reflected waves (see equation (1.59) p.37)

$$Q(\mathbf{x}) = (\mathbf{a} + \mathbf{b})^2 = \mathbf{a}^2 + \underbrace{\mathbf{ab} + \mathbf{ba}}_{=0} + \mathbf{b}^2 = \mathbf{1}\left(|\mathbf{a}|^2 - |\mathbf{b}|^2\right).$$

Canonical bilinear form and inner product Let $\mathbf{x}, \mathbf{y}$ be two elements of $\mathcal{B}$, following remark 1.2 p.19, we introduce the canonically defined bilinear form B through the polarization identity

$$B(\mathbf{x}, \mathbf{y}) := \frac{1}{2}\left(Q(\mathbf{x} + \mathbf{y}) - Q(\mathbf{x}) - Q(\mathbf{y})\right). \tag{9.41}$$

Expanding (9.41), it is immediate that B is identical to the GA inner product (9.1) p.241.

$$B(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\left((\mathbf{x} + \mathbf{y})^2 - \mathbf{x}^2 - \mathbf{y}^2\right) = \frac{\mathbf{xy} + \mathbf{yx}}{2} = \mathbf{x} \cdot \mathbf{y}. \tag{9.42}$$

a) *flow and effort decomposition*: consider two elements $\mathbf{u} = \mathbf{e_u} + \mathbf{f_u}$, $\mathbf{v} = \mathbf{e_v} + \mathbf{f_v}$ with $\mathbf{e_u}, \mathbf{e_v} \in \mathcal{E}$ and $\mathbf{f_u}, \mathbf{f_v} \in \mathcal{F}$ then

$$B(\mathbf{u}, \mathbf{v}) = \mathbf{u} \cdot \mathbf{v} = (\mathbf{e_u} + \mathbf{f_u}) \cdot (\mathbf{e_v} + \mathbf{f_v}) = \mathbf{e_u} \cdot \mathbf{f_v} + \mathbf{f_u} \cdot \mathbf{e_v}$$

We recover the usual flow-effort representation of the symmetric bilinear form defined in definition 1.12 p.19. We note that equations (9.41), (9.42) do not rely on a particular choice of coordinates. This highlights the interest of GA to manipulate coordinate-free representations.

b) *wave decomposition*: consider two elements $\mathbf{u} = \mathbf{a_u} + \mathbf{b_u}$, $\mathbf{v} = \mathbf{a_v} + \mathbf{b_v}$, with $\mathbf{a_u}, \mathbf{a_v} \in \mathcal{W}^+$ and $\mathbf{b_u}, \mathbf{b_v} \in \mathcal{W}^-$ then the inner product between $\mathbf{u}$ and $\mathbf{v}$ is equal to the difference between the Euclidean inner products of their incident and reflected waves.

$$B(\mathbf{u}, \mathbf{v}) = \mathbf{u} \cdot \mathbf{v} = (\mathbf{a_u} + \mathbf{b_u}) \cdot (\mathbf{a_v} + \mathbf{b_v}) = \langle \mathbf{a_u}, \mathbf{a_v} \rangle_{\mathbb{R}^n} - \langle \mathbf{b_u}, \mathbf{b_v} \rangle_{\mathbb{R}^n}.$$

9.4.1 Dirac structures

Now that the geometric structure of the indefinite inner product space(see definition C.14 p.283) is setup, we can give the following GA definition of a Dirac structure

Definition 9.3 (Dirac structure (GA))**.** A *Dirac structure* $\mathcal{D}$ in $\mathcal{B} \sim \mathbb{R}^{n,n}$, is a self-orthogonal subspace of dimension n, i.e. a maximal subspace of vectors squaring to zero (for the GA product)

$$\mathcal{D} = \left\{ \mathbf{x} \in \mathcal{B} \quad | \quad \mathbf{x}^2 = 0 \right\}, \qquad \dim \mathcal{D} = n. \tag{9.43}$$

A Dirac structure $\mathcal{D}$ is said to be a *null space* (or a maximal isotropic space) and its elements are said to be *null vectors*.

Example 9.2. The following are examples of Dirac structures (see [Gua11])

- Let $\mathbf{x} \in \mathcal{D} = \mathcal{F}$, then $\mathbf{x}^2 = 0$, this corresponds to the constraint $\mathbf{e} = 0$ (short circuit),
- Let $\mathbf{x} \in \mathcal{D} = \mathcal{E}$, then $\mathbf{x}^2 = 0$, this corresponds to the constraint $\mathbf{f} = 0$ (open circuit),
- More generally (see [Gua11, ex 2.4]), let $V \subseteq \mathcal{F}$ be any subspace of $\mathcal{F}$ and define its annihilator space in $\mathcal{E}$ by $\mathrm{Ann}(V) := \left\{ e \in \mathcal{E} \mid \ \mathbf{e} \cdot \mathbf{f} = 0, \ \forall \mathbf{f} \in V \right\}$, then by construction $\mathcal{D} = V \oplus \mathrm{Ann}(V)$ is a Dirac structure since for $\mathbf{x} \in \mathcal{D}$ we have $\mathbf{x}^2 = (\mathbf{f} + \mathbf{e})^2 = 2\mathbf{e} \cdot \mathbf{f} = 0$.

Parametric representation of Dirac structures We revisit the representation of Dirac structures from a GA perspective. Let $\boldsymbol{\lambda} \in \mathbb{R}^n$ be a parametrisation of a Dirac structure so that

$$\mathcal{D} = \left\{ \mathbf{x} \in \mathcal{B} \mid \mathbf{x} = \mathsf{X}(\boldsymbol{\lambda}), \ \forall \boldsymbol{\lambda} \in \mathbb{R}^n \right\}, \tag{9.44}$$

with $\mathsf{X} : \mathbb{R}^n \to \mathcal{D} \subset \mathcal{B}$ structured as $\mathsf{X} = \mathsf{E} \oplus \mathsf{F}$, with an effort operator $\mathsf{E} : \mathbb{R}^n \to R(\mathsf{E}) \subset \mathcal{E}$, and a flow operator $\mathsf{F} : \mathbb{R}^n \to R(\mathsf{F}) \subset \mathcal{F}$. The Dirac structure constraint implies that $\mathbf{x}^2 = 0$. Since $\mathcal{E}$ and $\mathcal{F}$ are null spaces, only the cross-terms do not vanish so that we have

$$0 = \mathbf{x} \cdot \mathbf{x} = (\mathsf{E} \oplus \mathsf{F})(\boldsymbol{\lambda}) \cdot (\mathsf{E} \oplus \mathsf{F})(\boldsymbol{\lambda}) = \mathsf{F}(\boldsymbol{\lambda}) \cdot \mathsf{E}(\boldsymbol{\lambda}) + \mathsf{E}(\boldsymbol{\lambda}) \cdot \mathsf{F}(\boldsymbol{\lambda}) = \left\langle \boldsymbol{\lambda} \,\middle|\, \mathsf{F}^*\mathsf{E} + \mathsf{E}^*\mathsf{F} | \boldsymbol{\lambda} \right\rangle_{\mathbb{R}^n}.$$

we recover that Dirac structures should satisfy the constraint $\mathsf{F}^*\mathsf{E} + \mathsf{E}^*\mathsf{F} = 0$ and $\dim R(\mathcal{D}) = n$ from equation (1.27a) p.20.

Example 9.3. We want to represent (using GA) the Dirac structure induced by the hybrid skew-symmetric map

$$\begin{bmatrix} f_1 \\ f_2 \\ e_3 \end{bmatrix} = \begin{bmatrix} 0 & 0 & -1 \\ 0 & 0 & -1 \\ 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ f_3 \end{bmatrix}.$$

To do so, introduce the map $\mathsf{J}(\mathbf{x}) = J \cdot \mathbf{x} = \frac{1}{2}(J\mathbf{x} - \mathbf{x}J)$ defined by the bivector $J = -(\mathbf{f}_1 + \mathbf{f}_2) \wedge \mathbf{e}_3$. We verify using GA that the mapping is indeed

$$\mathsf{J}(\mathbf{f}_i) = -(\mathbf{f}_1 + \mathbf{f}_2)\delta_{i3}, \qquad \mathsf{J}(\mathbf{e}_i) = (\delta_{1i} + \delta_{2i})\mathbf{e}_3.$$

so that for $\mathbf{x} = \mathbf{e}_1 e_1 + \mathbf{e}_2 e_2 + \mathbf{f}_3 f_3$, the conjugated vector is $\mathsf{J}(\mathbf{x}) = \mathbf{f}_1 f_1 + \mathbf{f}_2 f_2 + \mathbf{e}_3 e_3$.

Proof. Using (a) the definition of the bivector contraction $J \cdot \mathbf{x}$, (b) the definition of bivector J, (c) commutation (with sign change) and factorisation of $(\mathbf{f}_1 + \mathbf{f}_2)$ on the left (respectively of $\mathbf{e}_3$ on the right), (d) the definition of the interior product and (e) the metric, we obtain

$$\begin{aligned}
\mathsf{J}(\mathbf{f}_i) &\overset{a}{=} \frac{J\mathbf{f}_i - \mathbf{f}_i J}{2} \overset{b}{=} -\frac{(\mathbf{f}_1 + \mathbf{f}_2)\mathbf{e}_3\mathbf{f}_i - \mathbf{f}_i(\mathbf{f}_1 + \mathbf{f}_2)\mathbf{e}_3}{2} \overset{c}{=} -(\mathbf{f}_1 + \mathbf{f}_2)\left(\frac{\mathbf{e}_3\mathbf{f}_i + \mathbf{f}_i\mathbf{e}_3}{2}\right) \\
&\overset{d}{=} -(\mathbf{f}_1 + \mathbf{f}_2)(\mathbf{e}_3 \cdot \mathbf{f}_i) \overset{e}{=} -(\mathbf{f}_1 + \mathbf{f}_2)\delta_{3i}, \\
\mathsf{J}(\mathbf{e}_i) &= \frac{J\mathbf{e}_i - \mathbf{e}_i J}{2} = -\frac{(\mathbf{f}_1 + \mathbf{f}_2)\mathbf{e}_3\mathbf{e}_i - \mathbf{e}_i(\mathbf{f}_1 + \mathbf{f}_2)\mathbf{e}_3}{2} = \left(\frac{(\mathbf{f}_1 + \mathbf{f}_2)\mathbf{e}_i + \mathbf{e}_i(\mathbf{f}_1 + \mathbf{f}_2)}{2}\right)\mathbf{e}_3 \\
&= \big((\mathbf{f}_1 + \mathbf{f}_2) \cdot \mathbf{e}_i\big)\,\mathbf{e}_3 = (\delta_{1i} + \delta_{2i})\mathbf{e}_3.
\end{aligned}$$

□

9.5 Exploring the geometry of $\mathbb{R}^{n,n}$ with Geometric Algebra

Indefinite inner product spaces are characterised by an involution. We follow the derivation of the main involution of $\mathbb{G}^{n,n}$ from [DHSVA93] and [DGL+03, p.413] to study its properties and their consequences. Introduce the linear duality map between $\mathcal{W}^+$ and $\mathcal{W}^-$ [26]

$$\mathsf{K}(\mathbf{x}) = \mathbf{x} \cdot K, \qquad \text{for the bivector} \qquad K = \sum_{i=1}^{n} \mathbf{a}_i \wedge \mathbf{b}_i. \tag{9.45}$$

Proposition 9.1. *The transform* K *is an involution. It satisfies* $(\mathsf{K}^2)(\mathbf{x}) = \mathbf{x}$ *and*

$$\mathsf{K}(\mathbf{a}_i) = \mathbf{b}_i, \qquad \mathsf{K}(\mathbf{b}_i) = \mathbf{a}_i. \tag{9.46}$$

It is a reflection in the subspace $\mathcal{E}$ *swapping spaces* $\mathcal{W}^+$ *and* $\mathcal{W}^-$ *(see figure 9.10).*

Proof. Using (a) associativity of the inner product, (b) the metric from (9.36) and (c) anti-commutativity of $\mathbf{a}_i\mathbf{b}_j = -\mathbf{b}_j\mathbf{a}_i$, we obtain.

$$\mathsf{K}(\mathbf{a}_i) = \mathbf{a}_i \cdot K \overset{a}{=} \sum_j (\mathbf{a}_i \cdot \mathbf{a}_j)\mathbf{b}_j \overset{b}{=} \mathbf{b}_i, \quad \mathsf{K}(\mathbf{b}_i) = \mathbf{b}_i \cdot K = \sum_j \mathbf{b}_i \cdot \mathbf{a}_j \mathbf{b}_j \overset{c}{=} \sum_j (-\mathbf{b}_i \cdot \mathbf{b}_j)\mathbf{a}_j \overset{b}{=} \mathbf{a}_i.$$

It follows that $\mathsf{K}^2(\mathbf{a}_i) = \mathbf{a}_i$ and $\mathsf{K}^2(\mathbf{b}_i) = \mathbf{b}_i$ so that $\mathsf{K}^2(\mathbf{x}) = \mathbf{x}$. □

Definition 9.4 (K-dual)**.** For any vector $\mathbf{x} \in \mathbb{R}^{n,n}$, we define its K-dual [a] by $\bar{\mathbf{x}} = \mathsf{K}(\mathbf{x})$.

a. This construct is analog to the complex-conjugate of a complex number.

Using (9.37) and (9.46), we find that the eigenvectors of K are given by (see figure 9.10)

$$\mathsf{K}(\mathbf{f}_i) = \mathsf{K}\left(\frac{\mathbf{a}_i - \mathbf{b}_i}{\sqrt{2}}\right) = \left(\frac{\mathsf{K}(\mathbf{a}_i) - \mathsf{K}(\mathbf{b}_i)}{\sqrt{2}}\right) = \left(\frac{\mathbf{b}_i - \mathbf{a}_i}{\sqrt{2}}\right) = -\mathbf{f}_i, \tag{9.47a}$$

$$\mathsf{K}(\mathbf{e}_i) = \mathsf{K}\left(\frac{\mathbf{a}_i + \mathbf{b}_i}{\sqrt{2}}\right) = \left(\frac{\mathsf{K}(\mathbf{a}_i) + \mathsf{K}(\mathbf{b}_i)}{\sqrt{2}}\right) = \left(\frac{\mathbf{b}_i + \mathbf{a}_i}{\sqrt{2}}\right) = \mathbf{e}_i. \tag{9.47b}$$

This induces a splitting $\mathbb{R}^{n,n} = \mathcal{F} \oplus \mathcal{E}$ according to positive and negative eigenvalues of K.

Proposition 9.2. *The projectors* [a] $\mathsf{P}_{\mathcal{F}} : \mathcal{F} \oplus \mathcal{E} \to \mathcal{F}$, *and* $\mathsf{P}_{\mathcal{E}} : \mathcal{F} \oplus \mathcal{E} \to \mathcal{E}$ *are defined by*

$$\mathsf{P}_{\mathcal{E}}(\mathbf{x}) = \frac{1}{2}(\mathbf{x} + \bar{\mathbf{x}}), \qquad \mathsf{P}_{\mathcal{F}}(\mathbf{x}) = \frac{1}{2}(\mathbf{x} - \bar{\mathbf{x}}). \tag{9.48}$$

a. similarly to the real and imaginary part of a complex number

Proof. Take the sum and differences of equations (9.47a) and (9.47b). □

The involution K is said to generate a *null structure* [27] because $\mathsf{K}(\mathbf{x})$ has opposite signature to $\mathbf{x}$ (use eq. (9.36))

$$\bar{\mathbf{a}}_i \cdot \bar{\mathbf{a}}_j = \mathbf{b}_i \cdot \mathbf{b}_j = -\delta_{ij}, \qquad \bar{\mathbf{a}}_i \cdot \bar{\mathbf{b}}_j = \mathbf{b}_i \cdot \mathbf{a}_j = 0, \qquad \bar{\mathbf{b}}_i \cdot \bar{\mathbf{b}}_j = \mathbf{a}_i \cdot \mathbf{a}_j = -\delta_{ij}, \tag{9.49}$$

so that one can decompose $\mathbf{x}$ as the sum of two *null vectors* $\mathbf{x} = \mathbf{x}^+ \oplus \mathbf{x}^-$ with $\mathbf{x}^- = \mathsf{P}_{\mathcal{F}}(\mathbf{x})$, and $\mathbf{x}^+ = \mathsf{P}_{\mathcal{E}}(\mathbf{x})$.

26. Note the identity of bivectors $\mathbf{a}_i \wedge \mathbf{b}_i = \mathbf{a}_i\mathbf{b}_i = \frac{1}{2}(\mathbf{e}_i + \mathbf{f}_i)(\mathbf{e}_i - \mathbf{f}_i) = \frac{1}{2}(\mathbf{e}_i^2 + \mathbf{f}_i\mathbf{e}_i - \mathbf{e}_i\mathbf{f}_i - \mathbf{f}_i^2) = \mathbf{f}_i \wedge \mathbf{e}_i$.
27. Instead of a complex structure.

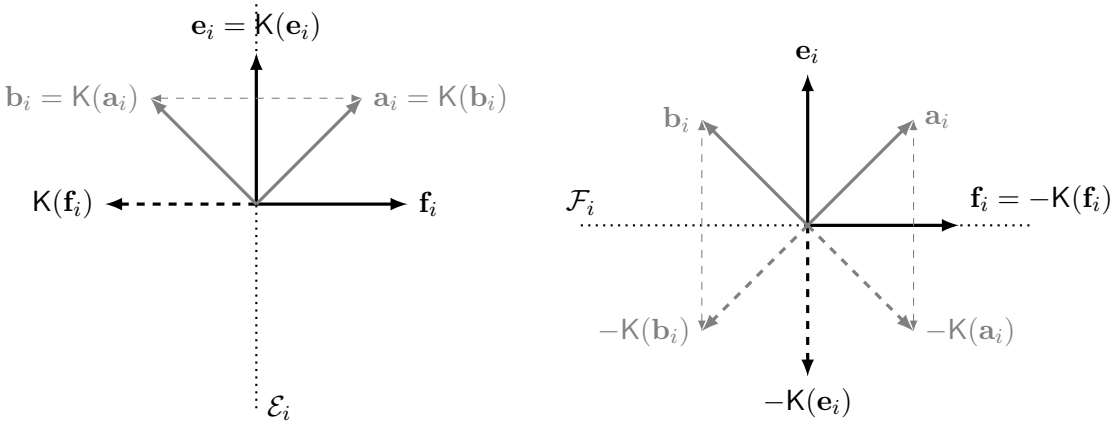


Figure 9.10 – Effects of the main involution K and $-\mathsf{K}$ on basis vectors. They correspond respectively to reflections in the spaces $\mathcal{E}_i$ and $\mathcal{F}_i$. Note that contrary to the Euclidean case, here contraction of vectors with bivector K yields a reflection instead of a 90 degrees rotation.

Subspaces

Basis functions $\{\mathbf{a}_i\}$ span the euclidean space $\mathcal{W}^+ \sim \mathbb{R}^{n,0}$ while the basis $\{\mathbf{b}_i\}$ span the anti-euclidean space $\mathcal{W}^- \sim \mathbb{R}^{0,n}$ so that $\mathbb{R}^{n,n}$ admits the decomposition

$$\mathbb{R}^{n,n} = \mathcal{W}^+ \oplus \mathcal{W}^-.$$

Following [DHSVA93, 3.19a-c], we can construct $(p+q)$-blades representing subspaces $\mathbb{R}^{p,q}$

$$W_{p,q} := A_p B_q^\dagger = A_p \wedge B_q^\dagger \tag{9.50}$$

where $A_p := \mathbf{a}_1 \dots \mathbf{a}_p$, $B_q := \mathbf{b}_1 \dots \mathbf{b}_q$. Each blade defines a projector $\mathsf{P}_{p,q} : \mathbb{R}^{n,n} \to \mathbb{R}^{p,q}$ defined by (9.11) as

$$\mathsf{P}_{p,q}(\mathbf{x}) = (\mathbf{x} \cdot W_{p,q}) W_{p,q}^{-1} = \frac{1}{2}\left(\mathbf{x} - (-1)^{p+q} W_{p,q} \mathbf{x} W_{p,q}^{-1}\right).$$

9.6 Rotor description of the flow-effort to wave variables change

We want a GA realisation of the flow-effort to wave variable change (see (1.56) p.35)

$$\begin{bmatrix} a \\ b \end{bmatrix} = \underbrace{\left(\frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \right)}_{\text{rotation } \mathsf{R}(\cdot)} \underbrace{\begin{bmatrix} Z^{1/2} & 0 \\ 0 & Z^{-1/2} \end{bmatrix}}_{\text{squeeze } \mathsf{S}(\cdot)} \begin{bmatrix} f \\ e \end{bmatrix}, \qquad \text{with port resistance} \qquad Z > 0. \tag{9.51}$$

To simplify, we only consider a single port. We see this variable change as a sequence of two inner-product preserving (and thus power-preserving) basis changes representing the *same vector*

$$\mathbf{x} = \vec{\mathbf{f}} f + \vec{\mathbf{e}} e \qquad \xrightarrow{\mathsf{S}} \qquad \mathbf{x} = \tilde{\mathbf{f}} \tilde{f} + \tilde{\mathbf{e}} \tilde{e} \qquad \xrightarrow{\mathsf{R}} \qquad \mathbf{x} = \mathbf{a} a + \mathbf{b} b.$$

9.6.1 Hyperbolic squeeze mapping

Since the metric of this space is indefinite, it is easier to start with hyperbolic rotations.

Proposition 9.3. *The bivector* $\mathbf{B} = \mathbf{a} \wedge \mathbf{b} = \mathbf{f} \wedge \mathbf{e}$ *is a generator of hyperbolic rotations so that the squeeze mapping* S *mapping* $(\vec{\mathbf{f}}, \vec{\mathbf{e}})$ *to* $(\tilde{\mathbf{f}}, \tilde{\mathbf{e}})$ *can be realised by a rotor* S *with hyperbolic angle* φ *as*

$$\mathsf{S}(\mathbf{x}) = S \mathbf{x} S^{-1}, \qquad \textit{with} \qquad S = \mathsf{e}^{\mathbf{B}\varphi/2} \qquad \textit{and} \qquad \varphi = -\ln(Z). \tag{9.52}$$

Proof. 1) since $\mathbf{a} \cdot \mathbf{b} = 0$ and $(\vec{\mathbf{e}})^2 = (\vec{\mathbf{f}})^2 = 0$, we have the identity $\mathbf{B} = \mathbf{a} \wedge \mathbf{b} = \mathbf{ab} = \mathbf{f} \wedge \mathbf{e}$:

$$\mathbf{a} \wedge \mathbf{b} = \mathbf{ab} = \frac{1}{2}(\vec{\mathbf{e}} + \vec{\mathbf{f}})(\mathbf{e} - \mathbf{f}) = \frac{1}{2}((\vec{\mathbf{e}})^2 + \vec{\mathbf{f}}\vec{\mathbf{e}} - \vec{\mathbf{e}}\vec{\mathbf{f}} + (\vec{\mathbf{f}})^2) = \frac{1}{2}(\vec{\mathbf{f}}\vec{\mathbf{e}} - \vec{\mathbf{e}}\vec{\mathbf{f}}) = \vec{\mathbf{f}} \wedge \vec{\mathbf{e}}.$$

2) We show using Taylor series expansion of exp and grouping terms that

$$\exp(\mathbf{B}\varphi) = \sum_n \frac{(\mathbf{B}\varphi)^n}{n!} = \mathbf{1} \sum_k \frac{\varphi^{2k}}{2k!} + \mathbf{B} \sum_k \frac{\varphi^{2k+1}}{(2k+1)!} = \mathbf{1} \cosh(\varphi) + \mathbf{B} \sinh(\varphi). \tag{9.53}$$

where[28] $\mathbf{B}^{2k} = \mathbf{1}$ and $\mathbf{B}^{2k+1} = \mathbf{B}$ since $\mathbf{B}^2 = (\mathbf{ab})(\mathbf{ab}) = (\mathbf{ab})(-\mathbf{ba}) = \mathbf{a}(-\mathbf{b}^2)\mathbf{a} = \mathbf{aa} = \mathbf{1}$.
3) It follows by substituting $\mathbf{Ba} = \mathbf{aba} = -\mathbf{b}$, $\mathbf{Bb} = \mathbf{abb} = -\mathbf{a}$ in the previous result that

$$\mathsf{e}^{\mathbf{B}\varphi}\mathbf{a} = \mathbf{a} \cosh(\varphi) - \mathbf{b} \sinh(\varphi), \qquad \mathsf{e}^{\mathbf{B}\varphi}\mathbf{b} = \mathbf{b} \cosh(\varphi) - \mathbf{a} \sinh(\varphi).$$

4) We finally show that $\vec{\mathbf{f}}, \vec{\mathbf{e}}$ are eigenvectors of left multiplication by $\mathsf{e}^{\mathbf{B}\varphi}$ with eigenvalues $\mathsf{e}^{\pm\varphi}$,

$$\tilde{\mathbf{f}} := \mathsf{e}^{\mathbf{B}\varphi}\vec{\mathbf{f}} = \mathsf{e}^{\mathbf{B}\varphi}\left(\frac{\mathbf{a} - \mathbf{b}}{\sqrt{2}}\right) = \left(\frac{\mathbf{a} - \mathbf{b}}{\sqrt{2}}\right)(\cosh\varphi + \sinh\varphi) = \mathsf{e}^{\varphi}\vec{\mathbf{f}}.$$
$$\tilde{\mathbf{e}} := \mathsf{e}^{\mathbf{B}\varphi}\vec{\mathbf{e}} = \mathsf{e}^{\mathbf{B}\varphi}\left(\frac{\mathbf{a} + \mathbf{b}}{\sqrt{2}}\right) = \left(\frac{\mathbf{a} + \mathbf{b}}{\sqrt{2}}\right)(\cosh\varphi - \sinh\varphi) = \mathsf{e}^{-\varphi}\vec{\mathbf{e}}.$$

5) From $\mathbf{aB} = \mathbf{aab} = \mathbf{b}$, $\mathbf{bB} = \mathbf{bab} = \mathbf{a}$, and 3), we have $\mathbf{Bx} = -\mathbf{Bx}$ for any vector $\mathbf{x}$. This yields the commutation rule $\mathsf{e}^{\mathbf{B}\varphi}\mathbf{x} = \mathbf{x}\mathsf{e}^{-\mathbf{B}\varphi}$ so that we have the symmetrised representation

$$\mathsf{S}(\mathbf{x}) = \mathsf{e}^{\mathbf{B}\varphi}\mathbf{x} = \left(\mathsf{e}^{\mathbf{B}\varphi/2}\right)^2 \mathbf{x} = \mathsf{e}^{\mathbf{B}\varphi/2}\mathbf{x}\mathsf{e}^{-\mathbf{B}\varphi/2}.$$

6) The constraint $\mathbf{x} = \vec{\mathbf{f}} f + \vec{\mathbf{e}} e = \tilde{\mathbf{f}}\tilde{f} + \tilde{\mathbf{e}}\tilde{e} = \mathsf{e}^{\varphi}\mathbf{f}\tilde{f} + \mathsf{e}^{-\varphi}\mathbf{e}\tilde{e}$ yields $\tilde{f} = \mathsf{e}^{-\varphi} f$ and $\tilde{e} = \mathsf{e}^{\varphi} e$ so that we must choose $\mathsf{e}^{-\varphi/2} = Z^{1/2} \implies \varphi = -\ln(Z)$. Choosing $S = \mathsf{e}^{\mathbf{B}\varphi/2}$ completes the proof. □

28. Note that contrary to the euclidean case, here, because of the indefinite metric, the bivector $\mathbf{B}$ squares to $\mathbf{1}$ instead of $-\mathbf{1}$. The rotation is thus an hyperbolic one.

9.6.2 Rotation by $\pi/4$

Geometrically, the transform R is a rotation of angle $\theta = \pi/4$ (see fig. 9.10). However, because of the indefinite metric, the geometric intuition of euclidean space is lost [29]. We need another strategy: instead of exponentiating a bivector to generate a rotation, we compose reflections.

From (9.51), the action of R, yields the following identity on basis vectors

$$\mathsf{R}(\mathbf{e}) = \frac{\mathbf{e}-\mathbf{f}}{\sqrt{2}} = \mathbf{b}, \qquad \mathsf{R}(\mathbf{e}) = \frac{\mathbf{e}+\mathbf{f}}{\sqrt{2}} = \mathbf{a}. \tag{9.54}$$

First, we introduce the duality map between $\mathbf{e}$ and $\mathbf{f}$ [30]

Proposition 9.4. *Let $\mathsf{T}(\mathbf{x}) = \mathbf{axa}$ denote reflection in vector $\mathbf{a}$. Then T is an involution acting as the duality map between $\mathbf{f}$ and $\mathbf{e}$ such that*

$$\mathsf{T}(\mathbf{e}) = \mathbf{f}, \qquad \mathsf{T}(\mathbf{f}) = \mathbf{e}. \tag{9.55}$$

Its eigenvectors are respectively $\mathbf{a}, \mathbf{b}$ with eigenvalues $+1, -1$.

Proof. Using the metric $\mathbf{a}^2 = \mathbf{1}$ and anti-commutation $\mathbf{ab} = -\mathbf{ba}$ we can show that

$$\mathsf{T}(\mathbf{e}) = \mathbf{aea} = \mathbf{a}\frac{\mathbf{a}+\mathbf{b}}{\sqrt{2}}\mathbf{a} = \frac{\mathbf{a}-\mathbf{b}}{\sqrt{2}} = \mathbf{f}, \qquad \mathsf{T}(\mathbf{f}) = \mathbf{afa} = \mathbf{a}\frac{\mathbf{a}-\mathbf{b}}{\sqrt{2}}\mathbf{a} = \frac{\mathbf{a}+\mathbf{b}}{\sqrt{2}} = \mathbf{e}.$$

and also that $\mathsf{T}(\mathbf{a}) = \mathbf{aaa} = \mathbf{a}$ and $\mathsf{T}(\mathbf{b}) = \mathbf{aba} = -\mathbf{aab} = -\mathbf{b}$. □

Then composing K and T, we have the following result.

Proposition 9.5. *Let K, T be the involutions defined by (9.45) and (9.55). Then $\mathsf{R} : \mathbb{R}^{n,n} \to \mathbb{R}^{n,n}$ satisfies (9.54). It can be written as*

$$\mathsf{R}(\mathbf{x}) = \frac{1}{\sqrt{2}}\left(\mathbf{x} + \mathsf{K}(\mathsf{T}(\mathbf{x}))\right). \tag{9.56}$$

Proof. using $\mathsf{K}(\mathbf{e}) = \mathbf{e}$ and $\mathsf{K}(\mathbf{f}) = -\mathbf{f}$ from (9.47a)-(9.47b), we prove that (9.54) is satisfied

$$\begin{aligned}
\mathsf{R}(\mathbf{e}) &= \frac{1}{\sqrt{2}}\left(\mathbf{e} + \mathsf{K}(\mathsf{T}(\mathbf{e}))\right) = \frac{1}{\sqrt{2}}\left(\mathbf{e} + \mathsf{K}(\mathbf{f})\right) = \frac{1}{\sqrt{2}}\left(\mathbf{e} - \mathbf{f}\right) = \mathbf{b}, \\
\mathsf{R}(\mathbf{f}) &= \frac{1}{\sqrt{2}}\left(\mathbf{f} + \mathsf{K}(\mathsf{T}(\mathbf{f}))\right) = \frac{1}{\sqrt{2}}\left(\mathbf{f} + \mathsf{K}(\mathbf{e})\right) = \frac{1}{\sqrt{2}}\left(\mathbf{f} + \mathbf{e}\right) = \mathbf{a},
\end{aligned}$$

See figure 9.10. □

29. We have seen in (9.53) that bivector $\mathbf{ab}$ squares to $\mathbf{1}$ instead of $-\mathbf{1}$ generating hyperbolic rotations.
30. This is analog to (9.55), the duality map K between $\mathbf{a}$ and $\mathbf{b}$.

Conclusion

In this chapter, an introduction to Geometric Algebra and its constructs has been presented in section 9.1, We have briefly presented some motivating examples in 9.2 to highlight both invariants and the unification power of GA. We note the emergence of a multi-vector field potential and the perspective of working with multivector ODE which seems like a promising direction of research in particular for the case of dissipative systems for which the energy is no longer an invariant.

In section 9.3 we have briefly considered the encoding of input-state-output PH-ODE using GA. We have proposed two strategies to encode skew-symmetric and symmetric semi-positive definite maps. A more promising perspective is to embrace indefinite inner product spaces so that every linear transform can be represented by an orthogonal transform (the canonical representation of orthogonal transforms in GA uses spinors).

In section 9.4, we continue our journey, this time using GA to encode Dirac structures. As in chapter 1, encoding the duality product of flows and efforts induces an indefinite metric so that the bond space is isomorphic to the pseudo-euclidean space $\mathbb{R}^{n,n}$. We show (see definition 9.3 p.256) that with this GA formulation, Dirac structures are simply subspaces of $\mathbb{R}^{n,n}$ whose elements square to 0 (sometimes called null-vectors in the litterature) independently of their internal representation (flow-effort or incident-reflected waves).

In section 9.5, we have a closer look at the geometry of $\mathbb{R}^{n,n}$. By contrast with the euclidean space $\mathbb{R}^n$, we note that in indefinite signature, contraction with a bivector yields a reflection instead of a rotation so that a significant part of the geometric intuition developed in section 9.1 (which relied on euclidean geometry) has to be abandoned.

Finally, in section 9.6, we consider a pure spinor representation of the flow-effort to power wave variables change, as a sequence of two power-preserving linear transforms. The power-preserving hyperbolic squeze mapping in the wave variable change is easily found to be an hyperbolic rotation. However, unintuitively (because of the indefinite metric), we had to rely on a sequence of two reflections to implement the linear combination of flows and efforts to wave variables (which looks like a simple $\pi/4$ rotation in euclidean space).

This chapter present some initial work that needs to be further developed and matured. It still lacks the elegance usually associated with GA. A main difficulty is that intuitions from euclidean geometry are no longer valid in non-euclidean spaces. Another difficulty is that transcoding concepts into a different mathematical language does not yields simplification by itself. This is only a preliminary condition [31]. I hope that this chapter motivates more people to adopt Geometric Algebra for PHS and find simplifying answers to these questions.

31. In this regards, the Maxwell equations example from appendix F.3 p.322 is telling: first we have to drop the cross product in favour of the exterior product, second we have to unify the four Maxwell equations into a single one using the geometric product, and third we have to embrace the concept of partial differential equations over a multi-vector field to finally reveal that Maxwell equations simplify to the wave equation (over a multivector field).

General Conclusion

This thesis considers the power-balanced modelling and simulation of nonlinear audio circuits using the port-Hamiltonian framework. We proposed "virtual analog" simulation methods for both PH-ODE and PH-DAE that a) operate in the continuous-time domain, b) can reproduce the regularity of physical trajectories c) can be of high-accuracy order, d) preserve the power-balance over time-frames (and thus energy or passivity).

Contributions

Continuous-time VA signal-processing framework and anti-aliasing In chapter 3, we propose a "virtual analog" signal processing chain. In order to address causality of computations, bandwidth expansion and compatibility with numerical schemes, this toolchain operates with non-bandlimited signals having instead a Finite Rate of Innovation. To this end, we use generalised sampling theory. We propose input-output reconstruction in B-spline spaces based on the literature on B-spline signal processing. We also propose an exact implementation of ARMA filtering for piecewise-defined input signals. This allows to benefit from all the literature on analog filter design for band-limiting and classical resampling (Butterworth, Chebyshev, elliptic filters, etc.).

(S)PAC methods In chapter 4, we propose a class of (symmetric) power-balanced adaptive collocation methods called (S)PAC of arbitrary regularity order which are (linearly) high-order accurate. They can be interpreted as a generalisation of Hermite–Obreshkov methods. Their analysis (restricted here to the linear case) shows that the power-balance cannot be unconditionally preserved. However, it is remarkable that the power-balanced orbits of PAC(1) are closer to the orbits of the exact solutions than the orbit of the mid-point method (see figure 4.2 p.111). While (S)PAC methods admit rather simple formulations, their implementation is difficult in general (implicit and nonlinear in its parameters). For this reason, this path is not explored further in this thesis.

Projected power-Balance condition In section 5.1.1, we propose continuous-time projected Dirac and resistive structures over time-frames (definition 5.1 p.119). Then, in theorem 5.1, we establish a sufficient condition on projectors so that the power-balance is satisfied. This implies energy conservation for Hamiltonian systems and passivity for PH-ODE and PH-DAE (see corrolaries 5.1, 5.2, 5.3 p.120). The power balance condition is quite permissive, which leaves room on the choices of bases and on the design of projectors to obtain additional properties. We outline several prospective scenarios in section 5.1.2, including partitionable systems. In particular, the power-balance condition is satisfied for scalar orthogonal (self-adjoint) L^2 projectors [32].

32. We show in 5.2.5 p.127 that (although the projection viewpoint is not always emphasised in the literature) energy-preserving CSRK methods (which includes the AVF and HBVM) rely on scalar orthogonal projection.

RPM Choosing the scalar orthogonal projection strategy, we introduce Regular Power-balanced Methods for both PH-ODE and PH-DAE. Regularity is achieved through supplementary multi-derivative boundary conditions. We show in section 5.2.7, that this induces *nested projections* in the Sobolev space H^k, for which Peano error kernels are detailed. We study existence, uniqueness and accuracy order for PH-ODE by reformulating RPM using the theory of CSRK methods. To this end, we rely on the reproducing kernel of the projector and its properties to perform the translation. We show that accuracy order $2p$ is automatically reached if the (orthogonal) projector reproduces polynomials of order p, relating CSRK simplified order conditions with Strang–Fix conditions. This results explains why choosing polynomial spaces is optimal (and the default choice) to construct general-purpose numerical methods (but this is not the only one, see the cosine basis example in section D.7 p.297). It is also shown in the examples of sections 5.5.1, 5.5.2 that higher-order projection yields higher frequency bandwidth and ultimately less aliasing (a consequence of the higher *rate of innovation*). To leverage this result, it is necessary to not only know the value of the trajectory on the boundaries of time-frames, but to continuously know the values of the trajectory in-between [33]. Generalised frequency bandwidth is revisited in depth in section 8.6 p.221. As a numerical challenge, we simulate an equaliser whose resonance frequency is beyond the Nyquist frequency. We formalise RPM projection for linear state-spaces as a mixed continuous/discrete Legendre filterbank whose Z-domain and Laplace transform are detailed. We show that the resonance can be simulated without aliasing and that errors in the audible bandwidth decrease faster with increasing order than with oversampling. This evidence supports our thesis.

Energy-preserving exponential integrators In chapter 6, we consider projection-based conservative/passive exponential integrators. To motivate the choice of exponential integrators, we show that they naturally arise when trying to minimise the L^2 norm of the vector field approximation error using functional Newton iteration. We introduce a new tool: the doubly-projected AVF discrete gradient which can be applied not only to piecewise affine trajectories but to trajectories in the Sobolev space H^1 (including exponential trajectories). Based on this, we provide an alternate proof (see theorem 6.1 p.163) that the Exponential AVF method is both unconditionally energy-preserving and dissipating and we provide a geometric interpretation. This resulted is extended to PHS by adding external ports. Finally, based on the results of chapter 5, we propose an extension strategy to higher projection orders. However, this extension is no longer exact for linear systems (one of the main advantages of exponential integrators).

Passive operational amplifier modelling In chapter 7, we propose passive operational amplifier models for PHS (with saturation and explicit power supply ports). Surprisingly, the passivity of OPA models seems to have been overlooked in the literature. First an idealised memoryless conservative model is proposed and used to simulate Sallen–Key filters. Its constitutive law is shown to be a nonlinear modulated Dirac structure whose modulation coefficient is linked to output current splitting and power supply saturation. Then, we consider the limit case of an infinite amplifier gain, which requires the use of *set-valued relations*. The linear branch of this relation corresponds to the so-called nullors in the litterature, while other branches corresponds to the OPA in saturation. Alternatively, we show (using *across* port variable changes) that this relation can be continuously parametrised by the sum of input and output voltages (see section 7.2.2 p.192). Finally, to model slew-rate and limited bandwidth, we sketch the structure of a passive three-stage grey-box OPA whose complete realisation is left for further research.

33. Indeed, increasing accuracy may lead to increased aliasing if the trajectory is simply sampled without being bandlimited by an antialiasing filter.

Perspectives

(S)PAC methods

- *Complex time (and collocation points).* On several occasions (suggested by mathematical equations)[34] we felt the need to give sense to complex-valued time. In particular, for (S)PAC methods, once the dissipation rate is too high, the power-balance is no longer solvable over the reals. However it remains solvable over complex numbers. Imaginary time has been popularised by Stephen Hawking [Haw01] and is sometimes used in special relativity and quantum mechanics. However, using complex-time (which we feel should not be motivated exclusively by mathematical intuition) has far-reaching consequences which are beyond the scope of this thesis, but it remains a fascinating subject of exploration.

- *(S)PAC implementation.* In thesis, we have favoured the functional projection approach of chapter 5 (which is more generally applicable and linear in the estimation of parameters). However, we have seen that, despite some implementation challenges, (S)PAC methods have interesting properties (improved orbits and dissipation rate, no quadrature involved, built-in smoothness, etc) which can motivate further work to address these issues.

- *Minimizing the power-balance error* In (S)PAC, when the power-balance cannot be satisfied, we could relax the power-balance constraint by minimising instead the power-balance error. Indeed, in RPM, the dissipation rate is no longer exactly satisfied (see p.157), but the residual error is such that energy is still unconditionally preserved (or dissipated).

Continuous-time projection methods and RPM

- *pH-DAE existence and uniqueness conditions* In section 5.3, p.135, we have considered existence and uniqueness conditions of pH-DAE. To this end, we proposed intermediate results to prove the invertibility of the Jacobian in Newton iteration. However, while convergence is observed in practice, further work is required to obtain theoretical result.

- *Joint power-balance and $\mathcal{C}^k$-smoothess* A long standing problem during this thesis has been the joint-preservation of both power-balance and smoothness. On one hand, (S)PAC methods show that joint-preservation of both power-balance and smoothness is possible and beneficial for accuracy (but the existence domain is bounded and the implementation difficult). On the other hand, for RPM, orthogonal L^2 projection is a powerful tool to address both power-balance and accuracy (but we had to rely on *nested projections*). To combine SPAC and RPM, we tried to explore the design of *doubly-orthogonal bases*[35] (see [Ber70, Sha79]) in both L^2 and H^k. However, we faced several issues that require further work: (i) double-orthogonality requires the exact resolution of functional eigenvalue problems, (ii) the Sobolev inner products we are interested yield not only differential operators, but also involve the boundary trace operator (see (5.9) p.122), (iii) there is no guarantee that generated bases have the polynomial reproduction property (see section 5.2.6 p.128).

- *Projector design.* An extension of (projected passivity) theorem 5.1 p.119 which is suggested in section 5.1.2, consists in substituting nonzero entries by projections and (possibly) zeros by rejections $(\mathcal{I} - \mathcal{P})$ in structure matrices $\mathbf{J}$ and $\mathbf{R}$, so that resulting matrix operators

34. Computing projections using complex contour integrals is another example.
35. For example, prolate-spheroidal wave functions [SP61] are doubly-orthogonal in both $L^2(-1,1)$ and the Paley–Wiener subspace of bandlimited function in $L^2(\mathbb{R})$.

$\boldsymbol{\mathcal{J}}, \boldsymbol{\mathcal{R}}$ are respectively skew-adjoint and self-adjoint. This setting is less constrained than scalar orthogonal projection (used in RPM). We may exploit the additional degrees of freedom to preserve additional properties (for example joint smoothness and passivity as mentioned above).

- *Fast computation of projections.* We have seen in section 5.4.1 p.140 that for affine trajectories, using anti-derivatives, we have closed-form formulas to compute projection coefficients (e.g. Legendre expansions). However, for arbitrary trajectories (and nonlinearities), we have to rely on quadratures with a number quadrature nodes sufficiently high to reach machine precision. This can make the implementation cost of high-order schemes prohibitive (specially for non-smooth nonlinearities). It is thus desirable to have either more general exact closed-form integration results or fast $\mathcal{O}(n\log(n))$ implementations of projections (as in the compuation of FFT, DCT, etc.). To this end, the following reference [Ise11], proposes fast $\mathcal{O}(n\log(n))$ Legendre expansion, which looks promising for the implementation of high-order power-balanced methods based on time-domain projection.

- *Generalised bandwidth and high-order* As we have seen (see figures 5.11 p.151, 5.15, p.155, 8.22 p.228 and 8.23 p.229), high-order methods converge faster than oversampling and have a larger generalised frequency bandwidth (or finite rate of innovation). This raises the following opportunity: to which point can we increase the step size h (i.e downsample) while increasing order without deteriorating audio quality (in particular aliasing)? Indeed, if we can trade step size, against order, then we can simulate several blocks of input-output samples at once and amortise the cost of iterative solvers. To this end, the V-system [36] seems like a promising basis to consider: its basis functions are orthogonal in L^2 (for power-balance), have the polynomial reproducing property (for time-stepping accuracy) and satisfy wavelet multi-scale similarity (for frequency resolution).

- *Implicit constraints and Lagrangian submanifolds* A theoretical perspective, is to generalise time-continuous projection methods to constrained PHS which are no longer described by Hamiltonians, but by Lagrangian submanifolds (see [VdSM18, GHVdSR20]). As a partial answer to this question, we proposed in [MH20] a fully-implicit generalisation of time-continuous projection for PH-DAE (reproduced in appendix) where PHS are no longer required to be in semi-explicit DAE form [37].

Exponential integrators

- *Existence, uniqueness and stiff order conditions.* In this thesis, we focused on the power-balance of exponential power-balanced methods. By analogy with RPM, higher projection orders are expected to yield higher time-stepping accuracy. However, this intuition, as well as existence and uniqueness conditions, remains to be established quantitatively.

- *Linearly-exact high-order extension.* To extend the EAVF method to higher projection order, we had to drop exact integration of linear systems. An obvious perspective is to consider alternate extension strategies that are linearly-exact.

- *Exponential splines* In chapter 3, we relied on B-spline signal processing theory [UAE93a, UAE93b]. Results from section 3.3.1 are based on exact exponential integration of linear

36. A basis inpired both by Haar wavelets and Legendre polynomials.
37. Both flow and effort laws can depend on hidden implicit control variables. And the method can directly address hidden constraints such as inductor loops and capacitor cutsets (i.e. causality conflicts).

ARMA filters. A natural perspective is to consider more closely the theory of cardinal exponential splines [UB05, Uns05] for both theoretical results and numerical implementation.

Operational amplifier The architecture and specifications of a passive three-stages grey-box operational amplifier model have been detailed in section 7.3 p.195. White-box modelling of the differential input and push-pull output have been considered (see appendix D.9.3), but efficient minimal black-box realisations of each submodule (simple enough for real-time simulation) remain to be derived.

Wave-domain PHS simulation In section 1.4, as a first step to bridge the gap between Wave Digital Filters and port-Hamiltonian Systems (pursuing the work of Falaize [Fal16]), we formally study the scattering representation of elementary PHS components (storage, Dirac and resistive structures). Note that the combination of waves with the continuous-time projection of chapter 5 allows the definition of *projected functional waves* over each time-step. While the linear case and its port-adaptation are well-understood (through the Cayley transform), local port-adaptation of nonlinear storage or resistive structures for efficient simulations remains difficult (see [BS16]).

Geometric Algebra In chapter 9, we started translating PHS formulations into the language of Geometric Algebra. In this formulation, Dirac structures can be elegantly encoded as *null spaces* (see definition 9.3 p.256) i.e. subspaces whose vectors square to zero (called *null vectors*). In this formulation flow and effort spaces are also found to be null spaces while incoming wave spaces are *positive spaces* and outgoing wave spaces are *negative spaces*. While skew-symmetric matrices are easily encoded by bivectors, semi-positive dissipation matrices require more work to be replaced by GA constructs. A drawback of the indefinite metric of the bondspace, is that the flow-effort to waves transformation (which looks like a simple $\pi/4$ rotation using matrix notation) could not be intuitively expressed as the exponential of a bivector. We had to use the composition of two reflections instead. Further work is necessary to fully benefit from the GA framework and its mathematical encoding(s) of PHS. In particular, we think that a direct translation of reference [Mak10] from Clifford to Geometric algebra notations could be a "rosetta stone" to emphasise the roles of spinors in PHS.

Time/frequency-warping and backward error analysis In this thesis, we have considered approximation of the vector field of varying orders. However, using backward-error analysis [HLW06], it is also possible to increase the approximation order of low order schemes: numerical dispersion is compensated by time/frequency warping. A well-known example in audio is the bilinear scheme: *frequency warping* is compensated by *time warping* the step size h so that the frequency of a simulated pole is exactly preserved. This approach has been extended to nonlinear systems (including discrete gradients methods) by Cieśliński in [CR10, CR11, Cie13, Cie14]. A perspective of this thesis is to incorporate similar mechanisms for error feedback within continuous-time projection methods [38].

38. Note that using time-warping may change the Lebesgue measure in the L^2 inner product and thus the mathematical expression of the power-balance.

Part V

Appendix

Appendix A

Relations: definitions and properties

We recall here some results from reference [RB16] regarding relations. See also [AC12].

Relation A *relation* (also called *operator* or *multi-valued function*) on $\mathbb{R}^n$ is a subset $\mathcal{R}$ of $\mathbb{R}^n \times \mathbb{R}^n$. It is frequent to overload function and matrix operator notation so that

$$\mathcal{R}(x) \equiv \mathcal{R}x := \left\{ y \mid (x, y) \in \mathcal{R} \right\}.$$

Function If $\mathcal{R}(x)$ is a singleton or empty, then $\mathcal{R}$ is a function with domain

$$\operatorname{dom} \mathcal{R} := \left\{ x \mid \mathcal{R}(x) \neq \emptyset \right\}.$$

By abuse of notation, we identify the singleton $\{y\}$ with its value y in

$$\mathcal{R}(x) = \{y\} \sim y.$$

Some trivial examples of relations are

$$\begin{aligned} 0 &= \left\{ (x, y) \in \mathbb{R}^n \times \mathbb{R}^n \mid y = 0 \right\}, && \text{Zero relation (a function)} \\ I &= \left\{ (x, y) \in \mathbb{R}^n \times \mathbb{R}^n \mid y = x \right\}, && \text{Identity relation (a function)} \\ X_0 &= \left\{ (x, y) \in \mathbb{R}^n \times \mathbb{R}^n \mid x = 0 \right\}. && \text{(multi-valued)} \end{aligned}$$

Composition Let R, S be two relations we define the composition

$$R \circ S := RS = \left\{ (x, z) \quad | \quad \exists y \quad \text{such that} \quad (x, y) \in R, \quad \text{and} \quad (y, z) \in S \right\}$$

Sum Let R, S be two relations we define their sum by

$$R + S := \left\{ (x, y) \quad | \quad (x, y_R) \in R,\ (x, y_S) \in S, \quad y = y_R + y_S \right\}.$$

Extensions to standard binary operators is done similarly.

Inverse relation The inverse relation is

$$\mathcal{R}^{-1} := \left\{ (x, y) \quad | \quad (y, x) \in \mathcal{R} \right\}$$

Lipschitz continuity A relation is *Lipschitz continuous* if there exists $M > 0$ such that

$$\|v - u\| \leq M\|y - x\|\,, \quad \forall v \in \mathcal{R}(y),\ u \in \mathcal{R}(x)$$

If $M < 1$ it is called a *contraction*, if $M = 1$ it is said to be *non-expansive*. Mapping a pair of points by a contraction reduces the distance between them; mapping them by a nonexpansive operator does not increase the distance between them.

Fixed point We say that x is a fixed point of F is $F(x) = x$. When F is non expansive, the set of fixed points of F,

$$\big\{x \in \operatorname{dom} F \mid x = F(x)\big\} = (I - F)^{-1}(\{0\}),$$

is *closed* and *convex*. Moreover if F is a contraction and $\operatorname{dom} F = \mathbb{R}^n$, the set of fixed-points of F is *a singleton* (uniqueness).

Monotone operator An operator $\mathcal{R}$ is said to be *monotone* (*incrementally passive*) when

$$\langle v - u, y - x\rangle \geq 0, \quad \forall v \in \mathcal{R}(y),\ u \in \mathcal{R}(x)$$

Maximal monotone operator A monotone set-valued map $\mathcal{R}$ is *maximal* if there is no other monotone set-valued map whose graph contains strictly the graph of $\mathcal{R}$.

Strongly monotonous (coercive) operator It is said to be *strongly monotonous* if $\exists m > 0$ such that

$$\langle v - u, y - x\rangle \geq m\|y - x\|^2\,, \quad \forall x, y \in \operatorname{dom} F.$$

Strongly monotonous and Lipschitz operator Consider and operator $\mathcal{R}$ and denote constants m the *maximal lower bound* and M the *minimal lower bound* $(0 < m \leq M)$ such that

$$m\|y - x\|^2 \leq \big\langle \mathcal{R}(y) - \mathcal{R}(x), y - x\big\rangle \leq M\|y - x\|^2\,, \quad \forall x, y \in \operatorname{dom} F.$$

we define the *condition number* of $\mathcal{R}$ by

$$\kappa = \operatorname{cond}(\mathcal{R}) := \frac{M}{m}.$$

This situation is closely related to the notion of norm equivalence.

Example For example consider the relation

$$\mathcal{R}(x) = \begin{cases} x + 1 & x > 0 \\ \emptyset & x = 0\,, \\ x - 1 & x < 0 \end{cases} \qquad \overline{\mathcal{R}}(x) = \begin{cases} x + 1 & x > 0 \\ [-1, 1] & x = 0\,. \\ x - 1 & x < 0 \end{cases}$$

Then $\mathcal{R}$ is monotone but not maximal, while $\overline{\mathcal{R}}$ is monotone and maximal. Furthermore $\mathcal{R}$ is strongly monotone with constant $m = 1$, but not Lipschitz. It is however one-sided Lipschitz with constant $L = 1$. We have $\operatorname{cond}(\kappa) = 1$.

Resolvent the *resolvent* $R_{A,\alpha}$ of a relation A is (dropping)

$$R_A = (I + \alpha A)^{-1}$$

Cayley operator The *Cayley operator* (or *reflected resolvent*) of a relation A is

$$C_A = 2R_A - I. \tag{A.1}$$

For $\alpha > 0$ we have

- if A is monotone, then operators R_A, C_A are non-expansive
- if A is maximal monotone, then $\operatorname{dom} R_A = \operatorname{dom} C_A = \mathbb{R}^n$
- $0 \in A(x) \iff x = R_A(x) = C_A(x)$

Proof. $0 \in A(x) \iff x \in (I + A)(x) \iff (I + A)^{-1}(x) \ni x \iff R_A(x) \ni x.$ □

Identities Let A be a (possibly multi-valued) operator. Then

a) if A is maximal monotone and single-valued and $\alpha \geq 0$, we have

$$C_A = (I - \alpha A)(I + \alpha A)^{-1},$$

b) otherwise if A is multi-valued and $\alpha > 0$, we only have the weaker identity

$$C_A(I + \alpha A) = (I - \alpha A).$$

Proof. To prove (a), if A is maximal monotone and single-valued, then it is invertible (bijective)

$$C_A = 2R_A - I = 2(I + \alpha A)^{-1} - I = (2I - (I + \alpha A))(I + \alpha A)^{-1} = (I - \alpha A)(I + \alpha A)^{-1}.$$

□

Appendix B

Reminder on ODEs

We consider ordinary differential equations of the form $\dot{\boldsymbol{x}}(t) = \boldsymbol{f}(t, \boldsymbol{x}(t))$.

B.1 Runge–Kutta methods

Definition B.1 (Runge–Kutta method [HLW06] p.29)**.** Let b_i, $a_{i,j}$ $(i, j = 1, \dots, s)$ be real numbers and let $c_i = \sum_{j=1}^{s} a_{ij}$. An *s-stage Runge–Kutta method is given by*

$$\begin{cases} \mathbf{k}_i = \boldsymbol{f}\left(t_0 + hc_i,\ \mathbf{x}_0 + h\sum_{j=1}^{s} a_{ij}\mathbf{k}_j\right), \quad i = 1, \dots, s \\ \mathbf{x}_1 = \mathbf{x}_0 + h\sum_{i=1}^{s} b_i\mathbf{k}_i. \end{cases} \tag{B.1}$$

The slopes $\mathbf{k}_i$ do not necessarily exists, however, the implicit function theorem assures that, for sufficiently small h, the nonlinear system for the values $\mathbf{k}_1, \dots, \mathbf{k}_s$ has a locally unique solution close to $\mathbf{k}_i \approx \boldsymbol{f}(t_0, \mathbf{x}_0)$. Since Butcher's work the coefficients are usually displayed as follows

$$\begin{array}{c|ccc} c_1 & a_{11} & \dots & a_{1s} \\ \vdots & \vdots & & \vdots \\ c_s & a_{s1} & \dots & a_{ss} \\ \hline & b_1 & \dots & b_s \end{array} \qquad \equiv \qquad \begin{array}{c|c} \mathbf{c} & \mathbf{A} \\ \hline & \mathbf{b} \end{array}.$$

A direct generalisation of B.1 to a continuum of stages is

Definition B.2 (Continuous-stage Runge–Kutta method)**.** Let $B(\tau)$, $A(\tau, \sigma)$ be real functions of $\tau, \sigma \in [0, 1]$ and let $C(\tau) = \int_0^1 A(\tau, \sigma)\,\mathrm{d}\sigma$. A *continuous-stage Runge–Kutta method*

is given by

$$\begin{cases} \boldsymbol{k}(\tau) = \boldsymbol{f}\left(t_0 + hC(\tau),\ \mathbf{x}_0 + h\displaystyle\int_0^1 A(\tau,\sigma)\boldsymbol{k}(\sigma)\,\mathrm{d}\sigma\right), \quad \tau \in [0,1] \\ \mathbf{x}_1 = \mathbf{x}_0 + h\displaystyle\int_0^1 B(\tau)\boldsymbol{k}(\tau)\,\mathrm{d}\tau. \end{cases} \tag{B.2}$$

Definition B.3 (Collocation methods [HLW06] p.30)**.** Let $c_1, \ldots, c_s$ be distinct real numbers (usually $0 \leq c_i \leq 1$). The collocation polynomial $\boldsymbol{X}(t)$ is a polynomial of degree s satisfying

$$\begin{cases} \boldsymbol{X}(t_0) & = \mathbf{x}_0, \\ \dot{\boldsymbol{X}}(t_0 + hc_i) & = \boldsymbol{f}(t_0 + hc_i, \boldsymbol{X}(t_0 + hc_i)), \quad i = 1, \ldots, s. \end{cases} \tag{B.3}$$

and the numerical solution of the *collocation method* is defined by $\mathbf{x}_1 = \boldsymbol{X}(t_0 + h)$.

B.2 Numerical Stabillity

Several notions of numerical stability exists, we recall here some important results and definitions.

Definition B.4 (Stability function [HW96] p.16)**.** Let $\Phi_h : x_0 \mapsto x_1$ be a time-stepping method whose application to the Dahlquist test equation $\dot{x} = \lambda x$ leads to

$$x_1 = R(z)x_0, \qquad z = h\lambda. \tag{B.4}$$

The function $R(z)$ is called the *stability function* of the method. The set

$$S = \left\{z \in \mathbb{C} \mid \left|R(z)\right| < 1\right\} \tag{B.5}$$

is called the *stability domain* of the method.

Proposition B.1 (Stability function of Runge–Kutta methods [BG08] p.243)**.** *The stability function of a Runge–Kutta method* (B.1) *is the rational function.*

$$R(z) = 1 + z\mathbf{b}^{\mathsf{T}}(\mathbf{I} - z\mathbf{A})^{-1}\mathbf{1}. \tag{B.6}$$

Definition B.5 (A-stability [BG08] p.243)**.** A method is *A-stable* if its stability function satisfies

$$\left|R(z)\right| \leq 1 \quad \text{whenever} \quad \Re(z) \leq 0. \tag{B.7}$$

Definition B.6 (L-stability ([Ehl69]))**.** A method is *L-stable* if it is A-stable and if in addition

$$\lim_{z\to\infty} R(z) = 0. \tag{B.8}$$

For nonlinear ODEs, B-stability characterize the fact that the distance between two solutions is a non increasing functions of time.

Definition B.7 (B-stability [HW96] p.181)**.** A Runge–Kutta method is called *B-stable* if the contractivity condition

$$\left\langle \boldsymbol{f}(t,\mathbf{x}_1) - \boldsymbol{f}(t,\mathbf{x}_2), \mathbf{x}_1 - \mathbf{x}_2 \right\rangle \leq 0, \tag{B.9}$$

implies for all $h \geq 0$.

$$\|\mathbf{x}_1 - \hat{\mathbf{x}}_1\| \leq \|\mathbf{x}_0 - \hat{\mathbf{x}}_1\| . \tag{B.10}$$

where $\mathbf{x}_1$ and $\hat{\mathbf{x}}_1$ are the numerical solutions after one step starting with initial values $\mathbf{x}_0$ and $\hat{\mathbf{x}}_0$, respectively.

Whereas B-stability relies on incremental dissipativity of the vector field $\boldsymbol{f}$, BN-stability only requires dissipativity.

Definition B.8 (BN-stability [BG08] p.263)**.** A Runge–Kutta method is called *BN-stable* if the condition

$$\left\langle \boldsymbol{f}(t,\mathbf{x}), \mathbf{x} \right\rangle \leq 0, \tag{B.11}$$

implies that the sequence of computed solutions satisfy

$$\|\mathbf{x}_n\| \leq \|\mathbf{x}_{n-1}\| . \tag{B.12}$$

We note that PHODEs satisfy the generalized passivity condition $\left\langle \boldsymbol{f}(\mathbf{x}), \nabla H(\mathbf{x}) \right\rangle \leq 0$.

A sufficient condition for B-stability is given by the algebraic conditions

Definition B.9 (Algebraic stability [BG08] p.263 and [HW96] p.182)**.** A Runge–Kutta method is *algebraically stable* if

- $b_i > 0$ for $i = 1, \ldots, s$,
- $M = (m_{ij} = (b_i a_{ij} + b_j a_{ji} - b_i b_j)_{i,j=1}^{s}$ is non negative definite.

Theorem B.1 ([BG08] p.263)**.** *If a Runge-Kutta method is algebraicallly stable then it is BN-stable.*

Theorem B.2 ([HW96] p.182)**.** *If a Runge-Kutta method is algebraicallly stable then it is B-stable.*

Theorem B.3 ([HW96] p.185))**.** *For a Runge–Kutta method it holds*

$$\textit{B-stable} \implies \textit{A-stable}. \tag{B.13}$$

Proposition B.2 ([CMM+09])**.** *À Runge–Kutta method with stability function $R(z)$ is energy-preserving for all quadratic Hamiltonians iff $R(z)R(-z) \equiv 1$.*

B.3 Elementary differentials and B-series

In the theory of B-series (see [HLW06, p.51] and [MMMKV17, CMOQ10]), and multi-derivatives Runge-Kutta methods, it is necessary to manipulate higher derivatives of the following systems

$$\dot{x}(t) = f(x(t)), \qquad \dot{x}(t) = f(t, x(t)), \qquad \begin{cases} \dot{x}(t) = f(x(t), u(t)), \\ y(t) = g(x(t), u(t)). \end{cases}$$

The Faa di Bruno formula is an important tool to manipulate derivatives of composed functions

$$\frac{\mathrm{d}^n}{\mathrm{d}t^n} f(x(t)) = \sum_S \frac{n!}{m_1! \dots m_n!} f^{(m_1+\dots+m_n)}(x(t)) \cdot \prod_{j=1} n \left(\frac{x^{(j)}(t)}{j!} \right)^{m_j} \tag{B.14}$$

where the sum is over the set $S = \left\{ (m_1, \dots, m_n) \in \mathbb{N}^n \mid 1 \cdot m_1 + 2 \cdot m_2 + \dots + n \cdot m_n = n \right\}$.

For ODEs, we have an additional piece of information $\dot{x} = f(x(t))$. Substituting this information and using the Faa di Bruno Formula recursively gives rise to B-series.

Autonomous case Let $x_0 = x(t_0)$, it is customary to note $f_k(t)(x_0) := \left(\frac{\mathrm{d}^k}{\mathrm{d}t^k} f(x(t_0 + t)) \right)(t)(x_0)$ such that the local derivatives f_k only depend on the evaluation point x_0. For compacity, it is also customary to use $f'[\cdot] = (Df)_{x_0}[\cdot], f''[\cdot,\cdot] = (D^2 f)_{x_0}[\cdot,\cdot], \dots$ where $D^n f$ denotes the Frechet derivatives (multi-linear operators) of f at x_0. It is also customary to omit parenthesis when possible i.e. $f'f = f'[f]$. As a final simplification, in B-series literature, to emphasize their combinatorial significance, elementary differentials are replaced by trees [1]. This yields [HLW06, p.51]

$$\begin{aligned}
\dot{x} &= f_0 := f && = \bullet, \\
\ddot{x} &= f_1 := f'f && = \bullet\!\!\bullet, \\
x^{(3)} &= f_2 := f''[f,f] + f'f'f, && = \bullet\!\!\bullet\!\!\bullet + \bullet\!\!\bullet\!\!\bullet, \\
x^{(4)} &= f_3 := f'''[f,f,f] + 3f''[f'f,f] + f'f''[f,f] + f'f'f'f && = \bullet\!\!\bullet\!\!\bullet\!\!\bullet + 3\,\bullet\!\!\bullet\!\!\bullet\!\!\bullet + \bullet\!\!\bullet\!\!\bullet\!\!\bullet + + \bullet\!\!\bullet\!\!\bullet\!\!\bullet.
\end{aligned}$$

Using this notation, the exact flow of the solution $x(t_0 + h) = \Phi_h(x_0)$ has the following series expansion (B-series when summing over rooted trees, Taylor series when summing for differentials)

$$\Phi_h(x_0) = \left(\mathcal{I} + h\,(\bullet) + \frac{h^2}{2!}\,(\bullet\!\!\bullet) + \frac{h^3}{3!}\left(\bullet\!\!\bullet\!\!\bullet + \bullet\!\!\bullet\!\!\bullet\right) + \frac{h^4}{4!}\left(\bullet\!\!\bullet\!\!\bullet\!\!\bullet + 3\,\bullet\!\!\bullet\!\!\bullet\!\!\bullet + \bullet\!\!\bullet\!\!\bullet\!\!\bullet + + \bullet\!\!\bullet\!\!\bullet\!\!\bullet\right) + \dots \right)(x_0).$$

Non-autonomous case The non autonomous case is slightly more complicated since f depends on variables t, x. We use the shorthand notation $f_t = \partial_t f$, $f_{tt} = \partial_t^2 f$, $f_{tx} = \partial_t \partial_x f$. Likewise by recursive substitution and application of the chain rule we obtain

$$\begin{aligned}
\dot{x} &= f_0 := f, \\
\ddot{x} &= f_1 := f_t + f_x f, \\
x^{(3)} &= f_2 := f_{tt} + f_{tx} f + f_x f_t + f_{xx}[f,f] + f_x f_x f.
\end{aligned}$$

1. The number of branches corresponds to the order of differentiation

State-space case For the state space systems $\dot{x}(t) = f(x(t), u(t))$, we obtain

$$\begin{aligned}
\dot{x} &= f_0 := f, \\
\ddot{x} &= f_1 := f_u[\dot{u}] + f_x[f], \\
x^{(3)} &= f_2 := f_{uu}[\dot{u}, \dot{u}] + f_u[\ddot{u}] + 2f_{xu}[f, \dot{u}] + f_x f_u[\dot{u}] + f_{xx}[f, f] + f_x f_x f,
\end{aligned}$$

and for $y(t) = g(x(t), u(t))$

$$\begin{aligned}
y &= g_0 := g, \\
\dot{y} &= g_1 := g_u[\dot{u}] + g_x[f], \\
\ddot{y} &= g_2 := g_{uu}[\dot{u}, \dot{u}] + g_u[\ddot{u}] + 2g_{ux}[\dot{u}, f] + g_x f_u[\dot{u}] + g_{xx}[f, f] + g_x f_x f.
\end{aligned}$$

PH-ODEs For the particular case of input-state-output PH-ODEs where

$$\begin{cases} \boldsymbol{f}(\mathbf{x}, \mathbf{u}) = \mathbf{A}\nabla H(\mathbf{x}) + \mathbf{G}\mathbf{u}, \\ \boldsymbol{g}(\mathbf{x}, \mathbf{u}) = \mathbf{G}^\mathsf{T}\nabla H(\mathbf{x}), \end{cases} \qquad \text{with} \quad \mathbf{A} = \mathbf{J} - \mathbf{R},$$

this yields the explicit expressions

$$\begin{aligned}
\boldsymbol{f}_0(\mathbf{x}, \mathbf{u}) &= \mathbf{A}\nabla H(\mathbf{x}) + \mathbf{G}\mathbf{u}, & \boldsymbol{f}_1(\mathbf{x}, \mathbf{u}, \dot{\mathbf{u}}) &= \mathbf{A}\nabla^2 H(\mathbf{x})[\mathbf{A}\nabla H(\mathbf{x}) + \mathbf{G}\mathbf{u}] + \mathbf{G}\dot{\mathbf{u}}, \\
\boldsymbol{g}_0(\mathbf{x}, \mathbf{u}) &= \mathbf{G}^\mathsf{T}\nabla H(\mathbf{x}), & \boldsymbol{g}_1(\mathbf{x}, \mathbf{u}, \dot{\mathbf{u}}) &= \mathbf{G}^\mathsf{T}\nabla^2 H(\mathbf{x})[\mathbf{A}\nabla H(\mathbf{x}) + \mathbf{G}\mathbf{u}].
\end{aligned}$$

Higher order derivatives can be obtained by following the same derivation process and substituting recursively but are not reproduced here.

> **Remark B.1** (Computer Algebra Software)**.** Note the existence of the symbolic calculus library [Sun15], written in Python, which can automate the manipulation of trees and B-series for the analysis of ODE (accuracy order, symplecticity, energy-preservation, modified equation, etc).

Appendix C

Functional Analysis

The functional results thereafter are gathered from references [BCL99, Aub11, CZ12, Chr16].

C.1 Definitions

Definition C.1 (Lipschitz continuity)**.** Let $\boldsymbol{f}$ be an operator on a normed space. If there exists constants $\ell_{\boldsymbol{f}}, L_{\boldsymbol{f}}$ such that

$$\ell_{\boldsymbol{f}}\|\boldsymbol{u}-\boldsymbol{v}\| \leq \left\|\boldsymbol{f}(\boldsymbol{u})-\boldsymbol{f}(\boldsymbol{v})\right\| \leq L_{\boldsymbol{f}}\|\boldsymbol{u}-\boldsymbol{v}\|\,, \tag{C.1}$$

then, $L_{\boldsymbol{f}}$ is called the *least upper bound Lipschitz constant* (or simply the *Lipschitz constant*) of $\boldsymbol{f}$ and $\ell_{\boldsymbol{f}}$ is called the *greatest lower bound Lipschitz constant* of $\boldsymbol{f}$.

Definition C.2 (One-sided Lipschitz continuity and logarithmic norm)**.** Let $\boldsymbol{f}$ be an operator on an inner product space, if there exists constants $m_{\boldsymbol{f}}, M_{\boldsymbol{f}}$ such that

$$m_{\boldsymbol{f}}\|\boldsymbol{u}-\boldsymbol{v}\|^2 \leq \left\langle \boldsymbol{f}(\boldsymbol{u})-\boldsymbol{f}(\boldsymbol{v}), \boldsymbol{u}-\boldsymbol{v}\right\rangle \leq M_{\boldsymbol{f}}\|\boldsymbol{u}-\boldsymbol{v}\|^2\,, \tag{C.2}$$

then, $M_{\boldsymbol{f}}$ is called the *least upper bound logarithmic Lipschitz constant* of $\boldsymbol{f}$ and $m_{\boldsymbol{f}}$ is the *greatest lower bound logarithmic Lipschitz constant* of $\boldsymbol{f}$.

Definition C.3. The logarithmic norm μ of a linear operator A is defined by

$$\mu(A) := \sup_{x\neq 0} \frac{\langle x, Ax\rangle}{\langle x, x\rangle}.$$

The logarithmic norm of a linear operator A is thus equivalent to its least upper bound logarithmic Lipschitz constant (i.e $\mu(A) = M_A$).

Definition C.4 (Contractivity)**.** An operator $\boldsymbol{f}$ satisfying $L_{\boldsymbol{f}} < 1$, is called *contractive*.

Definition C.5. An operator $\boldsymbol{f}$ satisfying $m_{\boldsymbol{f}} > 0$ is called *strongly convex*.

Definition C.6. An operator $\boldsymbol{f}$ satisfying $M_{\boldsymbol{f}} < 0$ is called *strongly concave.*

Definition C.7 (uniform monotonicity)**.** An operator $\boldsymbol{f}$ satisfying either $m_{\boldsymbol{f}} > 0$ or $M_{\boldsymbol{f}} < 0$, $\boldsymbol{f}$ is called *uniformly monotone.*

Definition C.8 (Fréchet derivative)**.** Let V and W be normed vector spaces, and $U \subset V$ be an open subset of V. A function $f : U \to W$ is called *Fréchet differentiable* at $x \in U$ if there exists a bounded linear operator $A : V \to W$ such that

$$\lim_{\|h\| \to 0} \frac{\left\| f(x+h) - f(x) - Ah \right\|_W}{\|h\|_V} = 0.$$

If there exists such an operator A, it is unique, so we write $Df(x) = A$ and call it the *Fréchet derivative* of f at x.

Alternative notations emphasizing the role of the operator A are $f'(x)(\cdot) \equiv f'_x(\cdot) \equiv A(\cdot)$.

Definition C.9 (Gateaux derivative)**.** A function $f : U \subset V \to W$ is called *Gateaux differentiable* at at $x \in U$ if f as a directional derivative along all directions at x. This means that there exists a function $g : V \to W$ such that

$$\mathrm{d}f(x; v) := \lim_{h \to 0} \frac{f(x + hv) - f(x)}{h} = g(v), \quad \forall v \in V.$$

and where h is from the scalar field associated with V (usually real).

If f is Frechet differentiable at x, it is also Gateaux differentiable there, and g is just the linear operator $A = Df(x)$. However, not every Gateaux differentiable function is Frechet differentiable.

Definition C.10 (Unilateral Laplace transform [CZ12])**.** Let V be a separable Hilbert space, let $u : \mathbb{R}^+ \to V$ have the property that $e^{-\beta t} u(t) \in L^1(\mathbb{R}^+, V)$ for some real β. We call these Laplace-transformable functions and we define their Laplace transform $\widehat{U}$ by

$$\mathcal{L} : u \mapsto \widehat{U}(s) = \int_0^\infty e^{-st} u(t) \, \mathrm{d}t, \tag{C.3}$$

for $s \in \overline{\mathbb{C}}_\beta^+ = \left\{ s \in \mathbb{C} \mid \operatorname{Re}(s) \geq \beta \right\}$.

Definition C.11 (Unilateral Z-transform)**.** Let $\mathbb{D} \subset \mathbb{C}$ denote the unit disc, and $H_2(\mathbb{D}, \mathbb{C})$ the Hardy space of square integrable holomorphic functions on the unit disk. The unilateral Z-transform is the operator $\mathcal{Z} : \ell^2(\mathbb{C}) \to H_2(\mathbb{D}, \mathbb{C})$ defined by

$$\mathcal{Z} : u \mapsto \widehat{U}(z) = \sum_{n=0}^{\infty} u_n z^n. \tag{C.4}$$

C.2 Banach, Hilbert and Sobolev spaces

Definition C.12 (Banach Space)**.** A Banach space $(X,\|\cdot\|)$ is a complete normed vector space.

Definition C.13 (Hilbert space)**.** A Hilbert space $(H, \langle\cdot,\cdot\rangle)$ is a real or complex inner product space that is also a complete metric space with respect to the norm $\|\cdot\|$ induced by the inner product $\langle\cdot,\cdot\rangle$.

Definition C.14 (indefinite inner product space)**.** An indefinite inner product space $(K, \langle\cdot,\cdot\rangle, J)$ is a vector space K equipped with both a positive semi-definite inner product $\langle\cdot,\cdot\rangle$ and an indefinite inner product $\langle u, v\rangle_J := \langle u, Jv\rangle$ where the metric operator J is an involution ($J^2 = I$).

The following subsets are defined in terms of the square norm induced by the indefinite inner product

$$K_0 := \left\{x \in K \mid \langle x, x\rangle_J = 0\right\}, \qquad \text{neutral space} \tag{C.5a}$$
$$K_+ := \left\{x \in K \mid \langle x, x\rangle_J > 0\right\}, \qquad \text{positive space} \tag{C.5b}$$
$$K_- := \left\{x \in K \mid \langle x, x\rangle_J < 0\right\} \qquad \text{negative space.} \tag{C.5c}$$

It is clear (see definition 1.3.1 p.18) that by definition Dirac structures are neutral spaces, while incident and reflected waves belong respectively to positive and negative spaces (see subsection 1.4.2 p.36). For more details on indefinite inner product spaces see [Bog12].

Definition C.15 (Base [Chr16])**.** Let X be a Banach space, A sequence of vectors $\{e_k\}_{k=1}^{\infty}$ of X is a *basis* for X if, for each $f \in X$, there exists unique scalar coefficients $\left\{c_k(f)\right\}_{k=1}^{\infty}$ such that

$$f = \sum_{k=1}^{\infty} e_k c_k(f). \tag{C.6}$$

Definition C.16 (Adjoint operator [Chr16])**.** Let U be a bounded operator from the Hilbert space $(K, \langle\cdot,\cdot\rangle_K)$ to the Hilbert space $(V, \langle\cdot,\cdot\rangle_V)$. The *adjoint* operator is defined as the unique operator $U^* : V \to K$ satisfying

$$\langle x, Uy\rangle_V = \left\langle U^* x, y\right\rangle_K, \qquad \forall x \in V, y \in K. \tag{C.7}$$

Definition C.17 (Lebesgue space)**.** The Hilbert space $L^2(\Omega, \mathbb{R})$ defined by

$$L^2(\Omega, \mathbb{R}) = \left\{u \;\middle|\; \int_\Omega \left|u(t)\right|^2 \mathrm{d}t < \infty\right\}$$

and equipped with the inner product

$$\langle u, v\rangle_{L^2} := \int_\Omega u(t) \cdot v(t)\, \mathrm{d}t,$$

is called the Lebesgue space of square-integrable real-valued functions.

Definition C.18 (Sobolev space [Aub11])**.** The subspace $H^m(\Omega)$ of $L^2(\Omega)$ defined by

$$H^m(\Omega) := \left\{ u \in L^2(\Omega) \mid \mathcal{D}^k u \in L^2(\Omega) \text{ for } k = 1 \dots m \right\} \tag{C.8}$$

is called the *Sobolev space* of order m, equipped with the scalar product

$$\langle u, v \rangle_{H^m(\Omega)} := \sum_{k=0}^{m} \left\langle \mathcal{D}^k u, \mathcal{D}^k v \right\rangle_{L^2(\Omega)} . \tag{C.9}$$

where $\mathcal{D}$ denote the derivative operator.

Definition C.19 (Dual space(s))**.** Let V be a Banach space, its *topological dual* V^* is the space of all linear functionals from V to a scalar field $\mathbb{F}$. Its *continuous dual* V' is the space all continuous (i.e. bounded) linear functionals on V.

Theorem C.1 (Riesz-Frechet representation theorem [BCL99])**.** *Let V be a Hilbert space with (continuous) dual V'. For all functionals $f^* \in V'$, there exists $f \in V$ such that*

$$f^*(g) = \langle f, g \rangle_V , \tag{C.10}$$

Definition C.20 (Volterra operator)**.** The *Volterra operator* $\mathcal{V}$ and its adjoint $\mathcal{V}^*$ are defined respectively, for any function u in $L^2([0,1])$, by

$$(\mathcal{V}u)(\tau) = \int_0^\tau u(\sigma)\,\mathrm{d}\sigma, \qquad (\mathcal{V}^*u)(\tau) = \int_\tau^1 u(\sigma)\,\mathrm{d}\sigma. \tag{C.11}$$

Property C.1. The Volterra operator $\mathcal{V}$ satisfies the following properties [Thi]

P1. The eigenvalues of $\mathcal{V}^*\mathcal{V}$ are $\sigma_n = \left(\frac{2}{\pi(2n+1)}\right)^2$, $n \in \mathbb{N}$.

P2. The operator norm of $\mathcal{V}$ is thus $\|\mathcal{V}\|_2 = 2/\pi = \sqrt{\sigma_0}$.

P3. The sum of $\mathcal{V}$ with its adjoing yields the self-adjoint averaging operator

$$\overline{\mathcal{V}} := \mathcal{V} + \mathcal{V}^* \qquad \text{with} \qquad (\overline{\mathcal{V}}u)(\tau) = \int_0^1 u(\sigma)\,\mathrm{d}\sigma. \tag{C.12}$$

P4. The difference $\mathcal{V} - \mathcal{V}^*$ is a skew-adjoint operator.

C.3 Strang–Fix conditions

Strang–Fix conditions, first formulated in [FS69, SF11] to analyse Finite Elements, are important in approximation theory, wavelets, and generalised sampling. They relates approximation order with polynomial reproduction, vanishing moments and spectral flatness of the approximation. Here we reproduce the following variant of Strang–Fix conditions from [Cha99].

Preparations For $h > 0$, the scaling operator $\mathcal{U}_h$ is defined by

$$(\mathcal{U}_h f)(x) = \frac{1}{\sqrt{h}} f\left(\frac{x}{h}\right).$$

Observe that it is norm preserving: $\|\mathcal{U}_h f\|_{L^2(\mathbb{R})} = \|f\|_{L^2(\mathbb{R})}$. More generally

$$\left\|(\mathcal{U}_h f)^{(n)}\right\|_{L^2(\mathbb{R})} = \frac{1}{h^n}\|f\|_{L^2(\mathbb{R})}.$$

Let $\mathcal{P}$ be an operator with localized shift-invariant[1] kernel $K(x,y)$ defined by $(\mathcal{P}f)(x) = \int_{\mathbb{R}} K(x,y)f(y)\,\mathrm{d}y$, and define the scaled operator $\mathcal{P}_h = \mathcal{U}_h \mathcal{P} \mathcal{U}_{\frac{1}{h}}$ so that $\|\mathcal{P}_h f\|_{L^2(\mathbb{R})} = \left\|\mathcal{P}\mathcal{U}_{\frac{1}{h}} f\right\|$.

Theorem C.2 (Strang–Fix conditions [Cha99])**.** *The following statements are equivalent:*

A1. For any $f \in H^k(\mathbb{R})$,

$$\frac{1}{h^k}\|\mathcal{P}_h f - f\|_{L^2(\mathbb{R})} \to 0 \text{ when } h \to 0, \tag{C.13}$$

A2. (Accuracy order) For any $f \in H^{k+1}(\mathbb{R})$ and $h \leq 1$,

$$\|\mathcal{P}_h f - f\|_{L^2} \leq C h^k \left\|f^{(k+1)}\right\|_{L^2(\mathbb{R})}, \tag{C.14}$$

A3. (Polynomial reproduction) For any integer $0 \leq p \leq k$

$$\int_{\mathbb{R}} K(x,y) y^p \,\mathrm{d}y = x^p. \tag{C.15}$$

for almost every x.

A different formulation is proposed in [JL93] with an emphasis on the spectral flatness

Definition C.21 ([JL93])**.** Let Φ be a finite collection of compactly supported functions in $L^1(\mathbb{R}^s)$. We denote by $\mathrm{span}(\Phi)$ the linear span of Φ and by $S(\Phi)$ the linear space spanned by the functions in Φ and all their shifts. Here by a shift we mean a multi-integer translate.

Given a positive integer k we say that the collection Φ satisfies the *Strang–Fix conditions of order k* if there is an element ψ of $S(\Phi)$ such that

$$\widehat{\psi}(0) = 1, \qquad \mathcal{D}^\lambda \widehat{\psi}(2\pi\alpha) = 0,$$

for all $\lambda \in \mathbb{N}^s$ with $|\lambda| < k$ and all $\alpha \in \mathbb{Z}^z \setminus \emptyset$, where $\widehat{\psi}$ denote the Fourier transform of ψ.

1. i.e. $K(x+1, y+1) = K(x,y)$ and $\exists M > 0$ such that $K(x,y) = 0$ for $|x-y| \geq M$.

C.4 Shifted orthonormal Legendre polynomials

Some properties of shifted [2] orthonormal Legendre polynomials on $\Omega = (0,1)$ are detailed below.

Rodrigues formula Legendre polynomials are defined explicitly by

$$P_n(\tau) := \frac{\sqrt{2n+1}}{n!} \frac{\mathrm{d}^n}{\mathrm{d}\tau^n} \left(\tau^n (1-\tau)^n\right), \quad \forall n \in \mathbb{N}. \tag{C.16}$$

Symmetry Shifted Legendre polynomials are symmetric (anti-symmetric) with respect to 1/2

$$P_n(1-\tau) = (-1)^n P_n(\tau).$$

Orthonormality They are orthonormal with respect to the L^2 inner product on $[0,1]$

$$\langle P_m, P_n \rangle = \int_0^1 P_m(\tau) P_n(\tau) \,\mathrm{d}\tau = \delta_{mn}, \quad \forall m, n \in \mathbb{N}.$$

Integration Their integral can be represented in the Legendre basis [TS12, BTI09] by .

$$\int_0^\tau P_n(s) \,\mathrm{d}s = \begin{cases} \xi_1 P_1(\tau) + \frac{1}{2} P_0(\tau) & n = 0, \\ \xi_{n+1} P_{n+1}(\tau) - \xi_{n-1} P_{n-1} \tau & n > 0, \end{cases}, \qquad \xi_n = \frac{1}{2\sqrt{4n^2-1}}. \tag{C.17}$$

Let $\mathcal{V}$ denote the Volterra operator (C.11) defined by $\mathcal{V}u = \int_0^\tau u(\sigma) \,\mathrm{d}\sigma$. Then, the Volterra operator is represented by the (almost skew-symmetric) tridiagonal operational matrix of integration

$$\mathbf{V} := \left[\langle P_m, \mathcal{V} P_n \rangle\right] = \begin{bmatrix} \frac{1}{2} & -\xi_1 & & \\ \xi_1 & 0 & \ddots & \\ & \ddots & \ddots & -\xi_n \\ & & \xi_n & 0 \end{bmatrix}. \tag{C.18}$$

Boundary values $P_n(\alpha) = (2\alpha - 1)^n \sqrt{2n+1}$ for $\alpha \in \{0,1\}$.

Vanishing integral on boundaries By orthogonality with $P_0(\tau) = 1$, integrals of shifted Legendre polynomials vanish on the boundary of the unit interval for $n > 0$

$$\int_0^\alpha P_n(s) \,\mathrm{d}s = 0, \quad \alpha \in \{0,1\}, \quad \forall n > 0. \tag{C.19}$$

Adjoint Volterra identity It follows by decomposing $\int_0^1 = \int_0^\tau + \int_\tau^1$ that we have the identity

$$\int_0^\tau P_n(s) \,\mathrm{d}s = -\int_\tau^1 P_n(s) \,\mathrm{d}s + \begin{cases} 1 & n = 0, \\ 0 & n > 0. \end{cases} \tag{C.20}$$

using operator notation, with the adjoint operator $\mathcal{V}^* u = \int_\tau^1 u(\sigma) \,\mathrm{d}\sigma$ from (C.11), this is equivalent to

$$\mathcal{V} P_n = -\mathcal{V}^* P_n + \delta_{0n}. \tag{C.21}$$

2. Legendre polynomials are defined on $(-1,1)$.

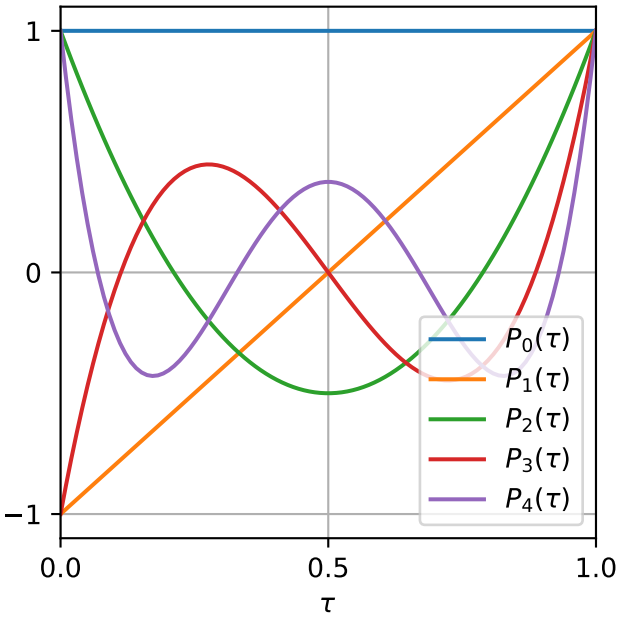


Figure C.1 – (Shifted orthonormal Legendre polynomials) $\{P_n(\tau)/P_n(1)\}$.

C.5 Hermite polynomial splines

Hermite splines of degree $d = 2k+1$, for $k > 0$ are the unique polynomials $h_{m,\alpha} \in \mathbb{P}^d([0,1])$ which satisfy the (bi-orthogonality) relations

$$\mathcal{B}_{\alpha'}^{m'}(h_{m,\alpha}) = \begin{cases} 1 & m = m' \text{ and } \alpha = \alpha' \\ 0 & \text{otherwise} \end{cases}, \qquad \text{where} \qquad \mathcal{B}_{\alpha}^{m}(f) = \frac{\mathrm{d}^m f}{\mathrm{d}\tau^m}(\alpha). \tag{C.22}$$

for all $\alpha \in \{0,1\}$ and $m \in \{0,\dots,k\}$.

Example C.1 (Cubic Hermite splines)**.** For $k = 1$, cubic Hermite splines are explicitly given by

$$h_{0,0}(\tau) = 2\tau^3 - 3\tau^2 + 1, \qquad h_{1,0}(\tau) = \tau^3 - 2\tau^2 + \tau,$$
$$h_{0,1}(\tau) = -2\tau^3 + 3\tau^2, \qquad h_{1,1}(\tau) = \tau^3 - \tau^2.$$

These functions are commonly used in Computer Assisted Design software and Computer Graphics to draw piecewise $\mathcal{C}^k$-continuous splines.

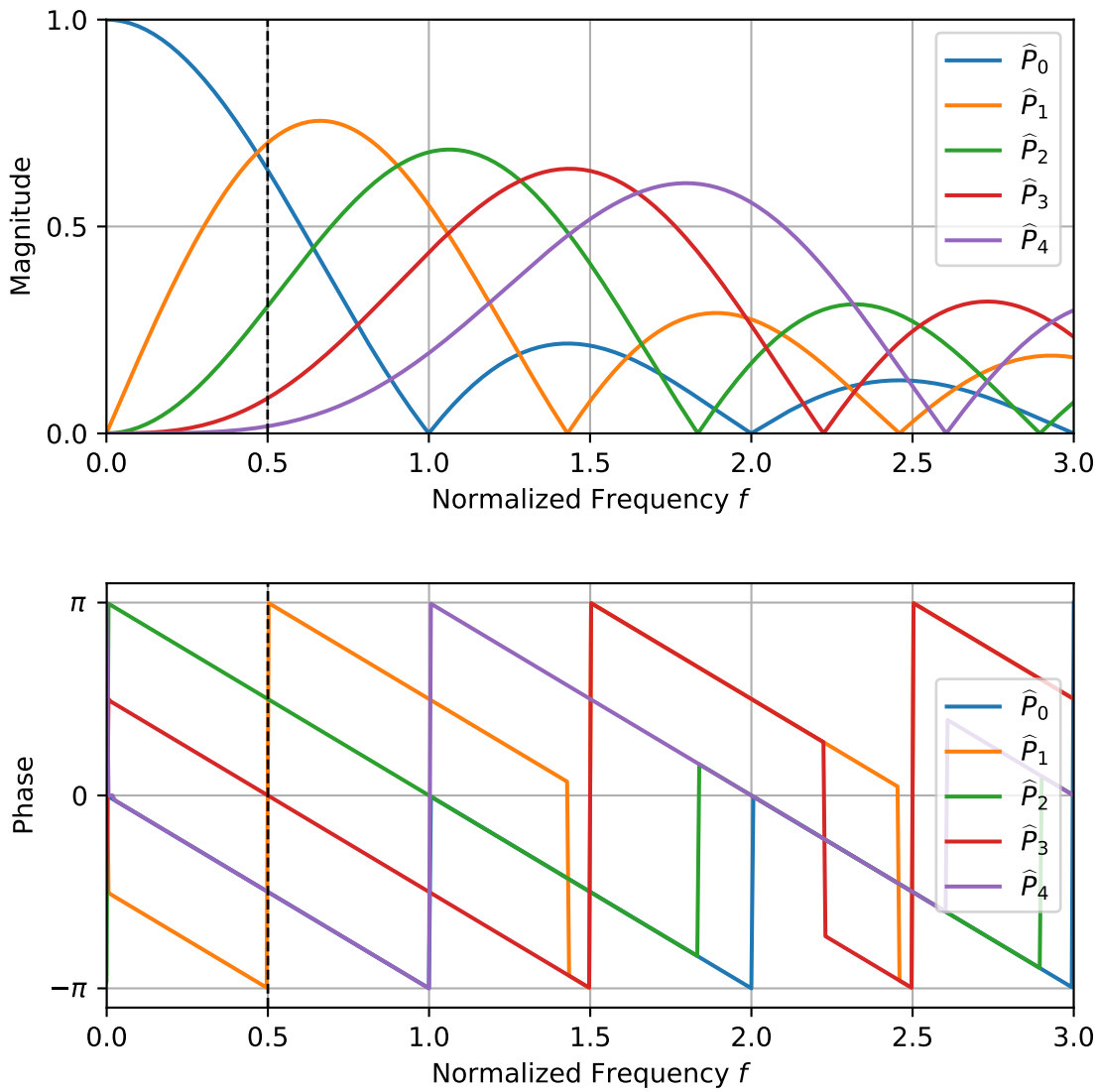


Figure C.2 – (Shifted orthonormal Legendre polynomials) Fourier spectrum $\widehat{P}_n(s = 2i\pi f)$ of $P_n(\tau)$ (restricted to $(0, 1)$). Note the phase linearity (constant phase slope), which is due to the time shift on the unit interval$(0, 1)$ and $\pm\pi$ discontinuitites at spectral zero-crossings.

Appendix D

Proofs

D.1 Exponential φ-functions: proofs and properties

The φ-functions, that appear when doing exact integration of an LTI system with polynomial input given in monomial form, are defined by the convolution integral

$$\varphi_k(\lambda, t) = \int_0^t e^{\lambda(t-\tau)} \frac{\tau^{k-1}}{(k-1)!} d\tau \qquad k \geq 1, \tag{D.1}$$

and by definition

$$\varphi_0(\lambda, t) := e^{\lambda t}. \tag{D.2}$$

For $\lambda = 0$ it is immediate that

$$\varphi_k(0, t) = \frac{t^k}{k!} \tag{D.3}$$

Recurrence relation We first prove that for $\lambda \neq 0$, they satisfy the recurrence formula

$$\varphi_{k+1}(\lambda, t) = \frac{\varphi_k(\lambda, t) - \varphi_k(0, t)}{\lambda}, \quad \lambda \neq 0. \tag{D.4}$$

Proof. Using integration by parts of (D.1) yields

$$\varphi_k(\lambda, t) = \left[e^{\lambda(t-\tau)} \frac{\tau^k}{k!} \right]_0^t + \lambda \int_0^t e^{\lambda(t-\tau)} \frac{\tau^k}{k!} d\tau = \frac{t^k}{k!} + \lambda \varphi_{k+1}(\lambda, t)$$

substituting (D.3) and collecting all term depending on $k+1$ on the left and k on the right, we obtain the recurrence relation

$$\varphi_{k+1}(\lambda, t) = \frac{\varphi_k(\lambda, t) - \varphi_k(0, t)}{\lambda}.$$

□

Explicit form Using (D.4) recursively for $\lambda \neq 0$, the first basis functions are given by

$$\varphi_0(\lambda, t) = e^{\lambda t}, \tag{D.5a}$$

$$\varphi_1(\lambda, t) = \frac{e^{\lambda t} - 1}{\lambda}, \tag{D.5b}$$

$$\varphi_2(\lambda, t) = \frac{e^{\lambda t} - (1 + \lambda t)}{\lambda^2}, \tag{D.5c}$$

$$\varphi_3(\lambda, t) = \frac{e^{\lambda t} - (1 + \lambda t + \frac{(\lambda t)^2}{2!})}{\lambda^3}, \tag{D.5d}$$

$$\varphi_4(\lambda, t) = \frac{e^{\lambda t} - (1 + \lambda t + \frac{(\lambda t)^2}{2!} + \frac{(\lambda t)^3}{3!})}{\lambda^4}. \tag{D.5e}$$

This suggests the following explicit form

$$\varphi_k(\lambda, t) = \frac{1}{\lambda^k} \left(e^{\lambda t} - \sum_{n=0}^{k-1} \frac{(\lambda t)^n}{n!} \right), \quad \lambda \neq 0 \tag{D.6}$$

Proof. 1) It is immediate to verify that (D.6) holds for $k = 0$. 2) Assuming that (D.6) is satisfied for some $k \in \mathbb{N}$ and using the recurrence relation (D.4), we prove by factoring the last term that (D.6) also holds for $k + 1$

$$\varphi_{k+1}(\lambda, t) = \frac{\varphi_k(\lambda, t) - \varphi_k(0, t)}{\lambda} = \frac{1}{\lambda^{k+1}} \left(e^{\lambda t} - \sum_{n=0}^{k-1} \frac{(\lambda t)^n}{n!} \right) - \frac{1}{\lambda} \frac{t^k}{k!} = \frac{1}{\lambda^{k+1}} \left(e^{\lambda t} - \sum_{n=0}^{k} \frac{(\lambda t)^n}{n!} \right).$$

Then by induction, equation (D.6) holds for all $k \in \mathbb{N}$. □

Taylor series form The φ-functions represent thus the tail of the truncated Taylor series expansion of $e^{\lambda t}$ up to a scaling factor. This is clear when rewriting (D.6) as

$$e^{\lambda t} = \sum_{n=0}^{k-1} \frac{(\lambda t)^n}{n!} + \lambda^k \varphi_k(\lambda, t). \tag{D.7}$$

By consequence, we may define φ-functions from the formal series

$$\varphi_k(\lambda, t) = \frac{1}{\lambda^k} \sum_{n=k}^{\infty} \frac{(\lambda t)^n}{n!} = \sum_{n=k}^{\infty} \lambda^{n-k} \frac{t^n}{n!}. \tag{D.8}$$

Note that, since no inversion is used, this definition can be extended to matrix-valued λ provided the series is convergent.

D.2 CSRK formulation of projected ODEs

We reformulate a projected ODE as a CSRK method (def. 5.3) and extract its parameters (A, B, C) which corresponds to equations (5.18a)-(5.18c) p.126.

Proof. Consider the projected ODE

$$\dot{\boldsymbol{X}} = h\mathcal{P}\boldsymbol{f}(\boldsymbol{X}, \boldsymbol{u}).$$

Rewrite the vector field equivalently using the substitution $\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{u}(\tau)) \to \boldsymbol{f}_{\boldsymbol{u}}(\tau, \boldsymbol{x})$ and drop the subscript $\boldsymbol{u}$. Let $K_{\mathcal{P}}$ be the reproducing kernel of the projector $\mathcal{P}$ (see eqs. (3.9) (3.10) p.84). By integration of $\dot{\boldsymbol{X}}$ and using Fubini's theorem, rewrite the projected ODE as a CSRK method (where functions $A(\tau, \sigma)$ and $B(\sigma)$ are extracted by identification)

$$\begin{aligned}
\boldsymbol{X}(\tau) &= \mathbf{x}_0 + h \int_0^\tau \left(\int_0^1 K_{\mathcal{P}}(\xi, \sigma) \boldsymbol{f}(t(\sigma), \boldsymbol{X}(\sigma)) \, \mathrm{d}\sigma \right) \mathrm{d}\xi \\
&= \mathbf{x}_0 + h \int_0^1 \underbrace{\left(\int_0^\tau K_{\mathcal{P}}(\xi, \sigma) \, \mathrm{d}\xi \right)}_{A(\tau,\sigma)} \boldsymbol{f}(t(\sigma), \boldsymbol{X}(\sigma)) \, \mathrm{d}\sigma, \\
\mathbf{x}_1 = \boldsymbol{X}(1) &= \mathbf{x}_0 + h \int_0^1 \underbrace{\left(\int_0^1 K_{\mathcal{P}}(\xi, \sigma) \, \mathrm{d}\xi \right)}_{B(\sigma)} \boldsymbol{f}(t(\sigma), \boldsymbol{X}(\sigma)) \, \mathrm{d}\sigma.
\end{aligned}$$

This proves (5.18a) p.126. Furthermore, by hypothesis, since $\mathcal{P}$ is self-adjoint ($\mathcal{P} = \mathcal{P}^*$), its reproducing kernel is symmetric ($K_{\mathcal{P}}(\tau, \sigma) = K_{\mathcal{P}}(\sigma, \tau)$) and since $\mathcal{P}$ reproduces contants, it follows that

$$B(\sigma) = \int_0^1 K_{\mathcal{P}}(\tau, \sigma) \, \mathrm{d}\tau = \int_0^1 K_{\mathcal{P}}(\tau, \sigma) \cdot 1 \, \mathrm{d}\tau = \mathcal{P}(1) = 1.$$

This proves (5.18b) p.126. Finally from the previous result, by symmetry of $K_{\mathcal{P}}$, it comes that

$$C(\tau) = \int_0^\tau \int_0^1 K_{\mathcal{P}}(\xi, \sigma) \, \mathrm{d}\sigma \, \mathrm{d}\xi = \int_0^\tau B(\sigma) \, \mathrm{d}\sigma = \tau.$$

This proves (5.18c) p.126. □

D.3 Proof of proposition 5.3 (CSRK order and Strang-Fix conditions)

We recall the following CSRK order conditions (5.21a), (5.21b), (5.21c) p. 128

$$\check{B}(\rho): \qquad \int_0^1 B(\tau)C(\tau)^{k-1}\,\mathrm{d}\tau = \frac{1}{k}, \qquad k=1,\dots,\rho,$$

$$\check{C}(\eta): \qquad \int_0^1 A(\tau,\sigma)C(\sigma)^{k-1}\,\mathrm{d}\sigma = \frac{C(\tau)^k}{k}, \qquad k=1,\dots,\eta,$$

$$\check{D}(\zeta): \qquad \int_0^1 B(\tau)C(\tau)^{k-1}A(\tau,\sigma)\,\mathrm{d}\tau = \frac{1}{k}B(\sigma)(1-C(\sigma)^k), \qquad k=1,\dots,\zeta.$$

Proof. Let $\mathcal{P}$ be a self-adjoint projector $(\mathcal{P}u)(\tau) = \int_0^1 K(\tau,\sigma)u(\sigma)\,\mathrm{d}\sigma$ that reproduces constants with reproducing kernel $K(\tau,\sigma)$. From (5.18a)-(5.18c) p.126, we recall that $B(\tau)=1$, $C(\tau)=\tau$, $A(\tau,\sigma)=\int_0^\tau K(\xi,\sigma)\,\mathrm{d}\xi$. Then we show that:

- condition $\check{B}(\rho=\infty)$ holds since for all $k\geq 1$

$$\int_0^1 B(\tau)C(\tau)^{k-1}\,\mathrm{d}\tau = \int_0^1 \tau^{k-1}\,\mathrm{d}\tau = \left[\frac{\tau^k}{k}\right]_0^1 = \frac{1}{k},$$

  where we used the definitions $B(\tau)=1$, $C(\tau)=\tau$.

- condition $\check{C}(\eta)$, is equivalent to projector $\mathcal{P}$ reproducing polynomials up to degree $\eta-1$:

$$\begin{aligned}
& \int_0^1 A(\tau,\sigma)C(\sigma)^{k-1}\,\mathrm{d}\sigma = \frac{C(\tau)^k}{k}, \\
\overset{a}{\Longleftrightarrow} \quad & \int_0^\tau\int_0^1 K(\xi,\sigma)\sigma^{k-1}\,\mathrm{d}\xi\,\mathrm{d}\sigma = \frac{\tau^k}{k}, \\
\overset{b}{\Longleftrightarrow} \quad & \int_0^1 K(\xi,\sigma)\sigma^{k-1}\,\mathrm{d}\sigma = \tau^{k-1}, \\
\overset{c}{\Longleftrightarrow} \quad & \mathcal{P}\tau^{k-1} = \tau^{k-1}.
\end{aligned}$$

  using (a) the definitions $C(\tau)=\tau$, $A(\tau,\sigma)=\int_0^\tau K(\xi,\sigma)\,\mathrm{d}\xi$ and Fubini's theorem, (b) differentiation with respect to τ, (c) the definition of the projector $\mathcal{P}$.

- condition $\check{D}(\zeta)$, is equivalent to the adjoint projector $\mathcal{P}^*$ reproducing polynomials from degree 1 to degree ζ

$$\begin{aligned}
& \int_0^1 B(\tau)C(\tau)^{k-1}A(\tau,\sigma)\,\mathrm{d}\tau = \frac{1}{k}B(\sigma)(1-C(\sigma)^k), \\
\overset{a}{\Longleftrightarrow} \quad & \int_0^1 \tau^{k-1}A(\tau,\sigma)\,\mathrm{d}\tau = \frac{1}{k}(1-\sigma^k), \\
\overset{b}{\Longleftrightarrow} \quad & \left[\frac{\tau^k}{k}A(\tau,\sigma)\right]_{\tau=0}^1 - \int_0^1 \frac{\tau^k}{k}K(\tau,\sigma)\,\mathrm{d}\tau = \frac{1}{k}(1-\sigma^k), \\
\overset{c}{\Longleftrightarrow} \quad & \int_0^1 \tau^k K(\tau,\sigma)\,\mathrm{d}\tau = \sigma^k, \\
\overset{d}{\Longleftrightarrow} \quad & \mathcal{P}^*\sigma^k = \sigma^k,
\end{aligned}$$

using (a) $B(\tau) = 1$, $C(\tau) = \tau$, (b) integration by parts with $\frac{\partial A}{\partial \tau} = K$, (c) $A(0, \sigma) = 0$, $A(1, \sigma) = B(\sigma) = 1$ and simplifying by $1/k$, (d) the definition of the adjoint projector $\mathcal{P}^*$.

□

D.4 Proof of theorem 5.2 (existence and uniqueness of CSRK solutions)

Proof. We formalize the solution of the CSRK as a fixed–point: find $\boldsymbol{X}_\star \in L^2(\Omega, \mathbb{R}^n)$ such that

$$\boldsymbol{X}_\star = G(\boldsymbol{X}_\star), \qquad \text{with} \qquad G(\boldsymbol{X}) = \mathbf{x}_0 + h \int_0^\tau \mathcal{P}\boldsymbol{f}(\boldsymbol{X}(s))\,\mathrm{d}s.$$

Let $\mathcal{V}$ be the Volterra operator defined by $(\mathcal{V}u)(\tau) = \int_0^\tau u(\sigma)\,\mathrm{d}\sigma$, we can rewrite G without ambiguity using operator notation as

$$G(\boldsymbol{X}) = \mathbf{x}_0 + h\mathcal{V}\mathcal{P}\boldsymbol{f}\boldsymbol{X} = \mathbf{x}_0 + h\mathcal{V} \circ \mathcal{P} \circ \boldsymbol{f} \circ \boldsymbol{X}.$$

Denote $\|\cdot\| := \|\cdot\|_{\mathbb{R}^n}$ and $\|\cdot\|_2 := \|\cdot\|_{L^2(\Omega,\mathbb{R}^n)}$ and let $\boldsymbol{X}_1, \boldsymbol{X}_2$ be two functions in $L^2(\Omega, \mathbb{R}^n)$. We prove the existence and uniqueness condition of the fixed-point in four steps.

step i) If $\boldsymbol{f}$ is $L_{\boldsymbol{f}}$-Lipschitz on $\mathbb{R}^n$, then it is also $L_{\boldsymbol{f}}$-Lipschitz on $L^2(\Omega, \mathbb{R}^n)$

$$\begin{aligned}\left\|\boldsymbol{f}(\boldsymbol{X}_1) - \boldsymbol{f}(\boldsymbol{X}_2)\right\|_2 &= \sqrt{\int_0^1 \left\|\boldsymbol{f}(\boldsymbol{X}_1(s)) - \boldsymbol{f}(\boldsymbol{X}_2(s))\right\|^2 \mathrm{d}s} \\ &\leq \sqrt{\int_0^1 L_{\boldsymbol{f}}^2 \left\|\boldsymbol{X}_1(s) - \boldsymbol{X}_2(s)\right\|^2 \mathrm{d}s} = L_{\boldsymbol{f}} \|\boldsymbol{X}_1 - \boldsymbol{X}_2\|_2 .\end{aligned}$$

step ii) The adjoint of $\mathcal{V}$ is $(\mathcal{V}^* u)(\tau) = \int_\tau^1 u(\sigma)\,\mathrm{d}\sigma$ and the eigenvalues of $\mathcal{V}^*\mathcal{V}$ are $\sigma_n = \left(\frac{2}{\pi(2n+1)}\right)^2$ so that the operator norm is $\|\mathcal{V}\|_2 = \sqrt{\|\mathcal{V}^*\mathcal{V}\|_2} = \sup_n \sqrt{\sigma_n} = 2/\pi$ (see reference [Thi]).

step iii) Using the operator norm of $\mathcal{V}$, $\mathcal{P}$ and the Lipschitz constant of $\boldsymbol{f}$ we obtain the bound

$$\left\|G(\boldsymbol{X}_1) - G(\boldsymbol{X}_1)\right\|_2 = \|h\mathcal{V}\mathcal{P}\boldsymbol{f}\boldsymbol{X}_1 - h\mathcal{V}\mathcal{P}\boldsymbol{f}\boldsymbol{X}_2\|_2 \leq h\|\mathcal{V}\|_2\|\mathcal{P}\|_2\, L_{\boldsymbol{f}}\, \|\boldsymbol{X}_1 - \boldsymbol{X}_2\|_2 .$$

step iv) Since the operator norm of an orthogonal projector is 1, then if $\alpha = 2hL_{\boldsymbol{f}}/\pi < 1$, the mapping G is contracting. By the Banach fixed–point theorem, this guarantees convergence of G to a unique fixed–point $\boldsymbol{X}_\star \in L^2$.

□

Remark D.1. Note that $\|\mathcal{V}\mathcal{P}\|_2 \leq \|\mathcal{V}\|_2\|\mathcal{P}\|_2$. In practice, the existence domain of fixed-point solutions is bigger than predicted above. For example, for the AVF projector $(\mathcal{P}u) = \int_0^1 u(s)\,\mathrm{d}s$, we have the majoration

$$\|\mathcal{V}\mathcal{P}\|_2 = \sup_{g \in L^2(\Omega)} \frac{\|\mathcal{V}\mathcal{P}g\|}{\|g\|} \leq \sup_{g \in L^2(\Omega)} \frac{\|\mathcal{V}\mathcal{P}g\|}{\|\mathcal{P}g\|} = \sup_{g \in \mathcal{P}(L^2(\Omega))} \frac{\|\mathcal{V}g\|}{\|f\|} = \left\|\mathcal{V}(1)\right\|_2 = \|\tau\|_2 = \frac{1}{\sqrt{3}}.$$

This leads to the improved bound $hL_{\boldsymbol{f}} < \sqrt{3}$. Note that this result is similar to the one obtained for SPAC methods in property 4.2 p.110. Convergence in other L^p normed spaces leads to different bounds.

D.5 Proof of proposition 5.5 p.129 (nested projectors)

Proof. From definition 5.1 p.122, we can reformulate (5.10) (5.11) in steps ii) and iii) as

$$\boldsymbol{\mathcal{Q}}\begin{bmatrix}\widetilde{\delta\boldsymbol{X}}\\ \widetilde{\boldsymbol{Y}}\end{bmatrix} = \boldsymbol{\mathcal{Q}}\left(\mathbf{J}-\mathbf{R}\right)\begin{bmatrix}\nabla H\left(\boldsymbol{X}\right)\\ \boldsymbol{u}\end{bmatrix}, \qquad \text{where} \qquad \begin{cases}\boldsymbol{X}(\tau) & := \mathbf{x}_0 + h\displaystyle\int_0^\tau \delta\boldsymbol{X}(\sigma)\,\mathrm{d}\sigma,\\ \widetilde{\boldsymbol{X}}(\tau) & := \mathbf{x}_0 + h\displaystyle\int_0^\tau \widetilde{\delta\boldsymbol{X}}(\sigma)\,\mathrm{d}\sigma,\\ \mathbf{x}_1 & := \boldsymbol{X}(1) = \widetilde{\boldsymbol{X}}(1).\end{cases}$$

where $\boldsymbol{\mathcal{Q}} = \mathcal{Q}\otimes\mathbf{I}_n$ ($\boldsymbol{\mathcal{Q}}$ commutes with $\mathbf{J}-\mathbf{R}$) and the operator $\mathcal{Q}$ is equivalently specified by the conditions

$$\mathcal{P}\mathcal{Q} \overset{a}{=} \mathcal{P}, \qquad \text{(orthogonal projector in } L^2\text{)} \tag{D.9a}$$

$$\mathcal{B}\mathcal{Q} \overset{b}{=} \mathcal{B}, \qquad \text{(multi-derivative interpolator in } H^k\text{)} \tag{D.9b}$$

$$\text{range}(\mathcal{B}^{-1}) = A_R \text{ where } A_R \perp A_{\mathcal{P}} \qquad \text{(orthogonality of } A_R \text{ and } A_{\mathcal{P}}\text{)} \tag{D.9c}$$

First we prove that $\mathcal{R} = \mathcal{B}^{-1}\mathcal{B}(\mathcal{I}-\mathcal{P})$ (equation (5.26) p.129): a) left multiply (D.9b) by $\mathcal{B}^{-1}$, b) from (D.9a), there exists an operator $\mathcal{R}$ such that $\mathcal{Q} = \mathcal{P}+\mathcal{R}$, c) finally use the relation $\mathcal{B}^{-1}\mathcal{B} = \mathcal{I}_{A_R}$ (prop. 5.4 p.129)

$$\begin{aligned} & \mathcal{B}^{-1}\mathcal{B}\mathcal{Q} \overset{a}{=} \mathcal{B}^{-1}\mathcal{B}\\ \iff\quad & \mathcal{B}^{-1}\mathcal{B}(\mathcal{P}+\mathcal{R}) \overset{b}{=} \mathcal{B}^{-1}\mathcal{B}\\ \iff\quad & \mathcal{B}^{-1}\mathcal{B}\mathcal{R} = \mathcal{B}^{-1}\mathcal{B}(\mathcal{I}-\mathcal{P})\\ \iff\quad & \mathcal{R} \overset{c}{=} \mathcal{B}^{-1}\mathcal{B}(\mathcal{I}-\mathcal{P}), \end{aligned} \tag{D.10}$$

Then we prove that $\mathcal{Q}$ (and $\mathcal{R}$) is a projector in four steps

i) We prove that $\mathcal{P}\mathcal{R} = 0$: using (D.9a) and idempotence of $\mathcal{P}$ we obtain

$$\mathcal{P}\mathcal{Q} = \mathcal{P} \iff \mathcal{P}(\mathcal{P}+\mathcal{R}) = \mathcal{P} \iff \mathcal{P}\mathcal{R} = \mathcal{P}-\mathcal{P}^2 \iff \mathcal{P}\mathcal{R} = 0.$$

ii) We prove that $\mathcal{R}\mathcal{P} = 0$: using (D.10) and the orthogonality relation $(\mathcal{I}-\mathcal{P})\mathcal{P} = 0$ we obtain

$$\mathcal{R}\mathcal{P} = \mathcal{B}^{-1}\,\mathcal{B}\underbrace{\mathcal{P}(\mathcal{I}-\mathcal{P})}_{=0} = 0.$$

iii) We show that $\mathcal{R}$ is a projector: a) we expand $\mathcal{R}$ using (D.10), b) from (D.9c) we deduce $(\mathcal{I}-\mathcal{P})\mathcal{B}^{-1} = \mathcal{B}^{-1}$, c) since $\mathcal{B}^{-1}\mathcal{B} = \mathcal{I}_{A_R}$ we have $(\mathcal{B}^{-1}\mathcal{B})^2 = \mathcal{B}^{-1}\mathcal{B}$, d) use equation (D.10)

$$\mathcal{R}^2 \overset{a}{=} \mathcal{B}^{-1}\mathcal{B}(\mathcal{I}-\mathcal{P})\mathcal{B}^{-1}\mathcal{B}(\mathcal{I}-\mathcal{P}) \overset{b}{=} (\mathcal{B}^{-1}\mathcal{B})^2(\mathcal{I}-\mathcal{P}) \overset{c}{=} \mathcal{B}^{-1}\mathcal{B}(\mathcal{I}-\mathcal{P}) \overset{d}{=} \mathcal{R},$$

iv) We finally show idempotence $\mathcal{Q}^2 = \mathcal{Q}$ so that $\mathcal{Q}$ is a projector (an oblique projector): using relations (i)-(iii), we obtain

$$\mathcal{Q}^2 = (\mathcal{P}+\mathcal{R})^2 = \mathcal{P}^2 + \mathcal{P}\mathcal{R} + \mathcal{R}\mathcal{P} + \mathcal{R}^2 = \mathcal{P}+\mathcal{R} = \mathcal{Q}.$$

This result is illustrate in figure D.1. □

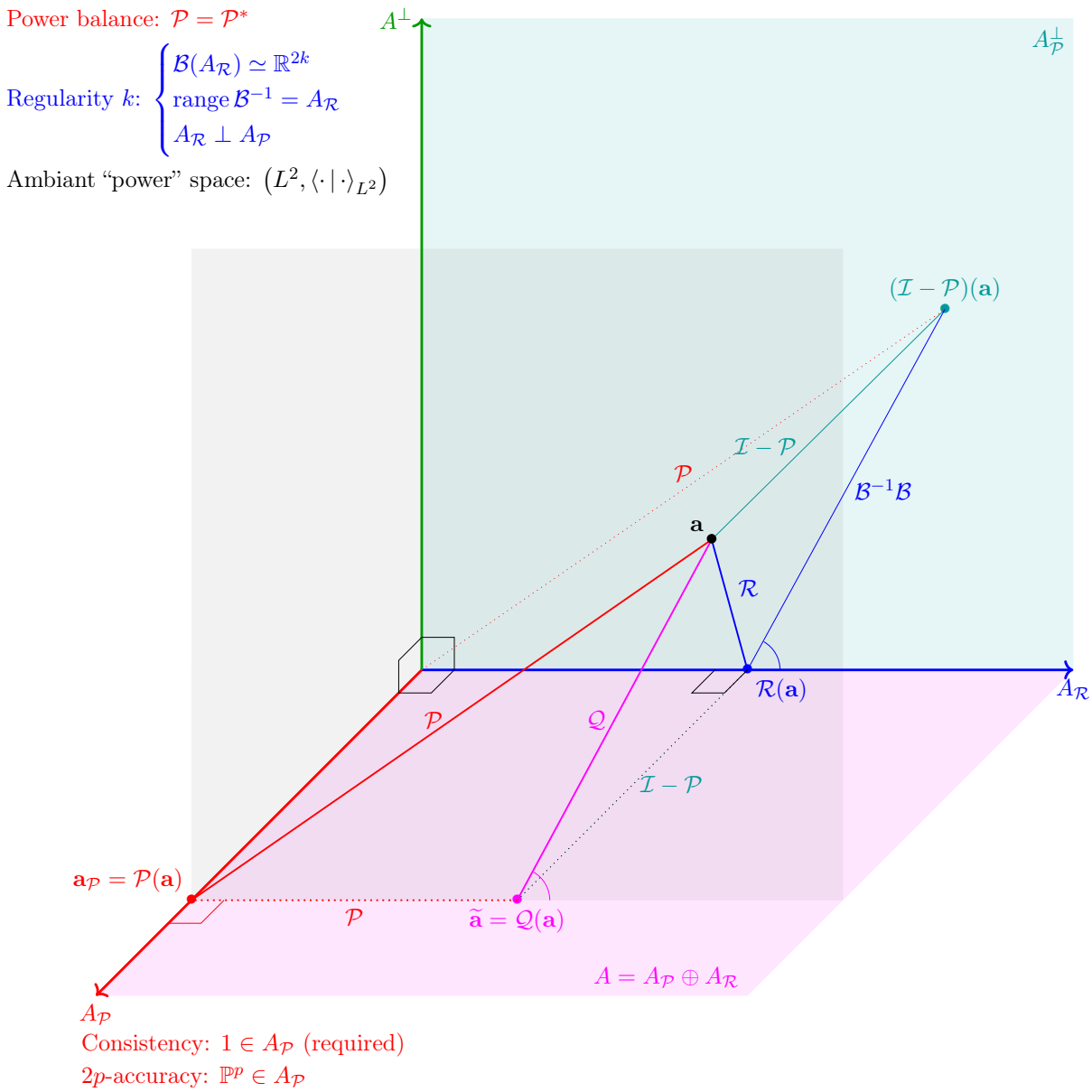


Figure D.1 – (RPM method) Illustration of orthogonal and oblique projectors $\mathcal{P}$, $\mathcal{Q}$, $\mathcal{R}$. Note that $\mathcal{P}$ and $\mathcal{Q}$ are nested projections ($\mathcal{P}\mathcal{Q} = \mathcal{P}$), the regularisation is $\mathcal{R} = \mathcal{B}^{-1}\mathcal{B}(\mathcal{I} - \mathcal{P})$ and $\mathcal{Q} = \mathcal{P} + \mathcal{R}$.

D.6 Proof of theorem 5.7 (Legendre expansion)

Proof. To simplify notations, here we use the shorthand $f_x^{[m]}(s) = f^{[(m)]}(x(s))$ for the antiderivatives of f evaluated at $x(s)$. We first prove the limit case $\delta x = 0$. Then, for $\delta x \neq 0$, we prove the general case by induction.

I) for $\delta x = 0$

- **Case** $n = 0$**,** $\widehat{\{f \circ x\}}_n = \int_0^1 f(x_0)\,\mathrm{d}s = f(x_0)$.
- **Case** $n \geq 1$**,** For $\delta x = 0$, $\forall n > 0$, by orthogonality of L_n with constants

$$\widehat{\{f \circ x\}}_n = \langle L_n, f(x_0)\rangle = 0.$$

II) for $\delta x \neq 0$

- **Case** $n = 0$**,** Since $\dot{x} = \delta x$, using the chain rule, $\frac{\mathrm{d}}{\mathrm{d}s}[f^{[1]}(x(s))] = f(x)\delta x$, we obtain

$$\widehat{\{f \circ x\}}_0 = \int_0^1 f(x(s))\,\mathrm{d}s = \frac{1}{\delta x}\int_0^1 \frac{\mathrm{d}}{\mathrm{d}s}\left[f^{[1]}(x(s))\right]\mathrm{d}s = \frac{f^{[1]}(x_0+\delta_x) - f^{[1]}(x_0)}{\delta x} \tag{D.11}$$

- **Case** $n = 1$**,** still using the chain rule, partial integration and $L_1' = \text{const}$, we obtain

$$\begin{aligned}\widehat{\{f \circ x\}}_1 = \int_0^1 L_1(s)f(x(s))\,\mathrm{d}s &= \frac{1}{\delta x}\left(\left[L_1(s)f^{[1]}(x(s))\right]_0^1 - \int_0^1 L_1'(s)f^{[1]}(x(s))\,\mathrm{d}s\right)\\ &= \frac{1}{\delta x}\left(\left[L_1(s)f^{[1]}(x(s))\right]_0^1 - \left\{L_1'(s)f^{[1]}(x(s))\right\}_0\right)\end{aligned}$$

  where the boundary terms are easily computable, and the inner product can be computed by substituting f by its antiderivative $f^{[1]}$ in (D.11).

- **Case** $n = 2$**,** using partial integration twice and $L_2'' = \text{const}$, we obtain

$$\begin{aligned}\widehat{\{f \circ x\}}_2 = \int_0^1 L_2(s)f(x(s))\,\mathrm{d}s &= \frac{1}{\delta x}\left[L_2 f_x^{[1]}\right]_0^1 - \frac{1}{\delta x}\int_0^1 L_2'(s)f^{[1]}(x(s))\,\mathrm{d}s\\ &= \frac{1}{\delta x}\left[L_2 f_x^{[1]}\right]_0^1 - \frac{1}{(\delta x)^2}\left[L_2' f_x^{[2]}\right] + \frac{1}{(\delta x)^2}\left\{L_2'' f_x^{[2]}\right\}_0\\ &= \frac{1}{\delta x}\left[L_2 f_x^{[1]}\right]_0^1 - \frac{1}{(\delta x)^2}\left[L_2' f_x^{[2]}\right]_0^1 + \frac{1}{(\delta x)^3}\left[L_2'' f_x^{[3]}\right]_0^1\end{aligned}$$

- **Case** $n = 3$**,** continuing partial integration, we obtain similarily

$$\begin{aligned}\widehat{\{f \circ x\}}_3 &= \int_0^1 L_3(s)f(x(s))\,\mathrm{d}s\\ &= \frac{1}{\delta x}\left[L_3 f_x^{[1]}\right]_0^1 - \frac{1}{(\delta x)^2}\left[L_3' f_x^{[2]}\right]_0^1 + \frac{1}{(\delta x)^3}\left[L_3'' f_x^{[3]}\right]_0^1 - \frac{1}{(\delta x)^4}\left[L_3''' f_x^{[4]}\right]_0^1\end{aligned}$$

- **General case** $n \geq 0$**,** by induction, we obtain the general solution

$$\widehat{\{f \circ x\}}_n = \sum_{k=0}^{n} \frac{(-1)^k}{(\delta x)^{k+1}}\left[L_n^{(k)}(s)f^{[k+1]}(x(s))\right]_0^1.$$

□

D.7 Stability function of L^2 projection methods

Proof. To prove proposition 5.2 p.127, we consider the Dahlquist test equation

$$\dot{x} = \lambda x, \quad x(0) = x_0, \quad \lambda \in \mathbb{C}.$$

For an orthonormal basis $\{\phi_n(\tau)\}_{n=0}^{p-1}$ over $\Omega = (0,1)$, orthogonal L^2 projection of the ODE yields

$$\begin{cases} \dot{x} = \lambda \left(1 \cdot x_0 + \int_0^\tau \dot{x}(s)\,\mathrm{d}s\right), \\ x_1 = x_0 + \int_0^1 1 \cdot \lambda \dot{x}(s)\,\mathrm{d}s, \end{cases} \quad \overset{\text{projection}}{\Longrightarrow} \quad \begin{cases} \dot{\vec{\mathbf{x}}} = \lambda \left(\mathbf{1}x_0 + \mathbf{V}\vec{\mathbf{x}}\right), \\ x_1 = x_0 + \lambda \mathbf{1}^\mathsf{T}\vec{\mathbf{x}}. \end{cases}$$

where $\mathbf{1} = [\langle \phi_m, 1\rangle]_{p\times 1}$, and the truncated operational matrix of integration is

$$\mathbf{V} = \left[\langle \phi_m, \mathcal{V}\phi_n\rangle\right]_{p\times p}, \qquad (\mathcal{V}u)(\tau) = \int_0^\tau u(s)\,\mathrm{d}s. \tag{D.12}$$

Solving for x_1, we obtain the time-stepping $x_0 \mapsto x_1 = R(\lambda)x_0$ with stability function

$$R(\lambda) = 1 + \lambda \mathbf{1}^\mathsf{T}(\mathbf{I} - \lambda\mathbf{V})^{-1}\mathbf{1}.$$

Using (a) the Sylvester determinant identity $\det(\mathbf{M})\det(1 + \mathbf{u}^\mathsf{T}\mathbf{M}^{-1}\mathbf{v}) = \det(\mathbf{M} + \mathbf{v}\mathbf{u}^\mathsf{T})$ with $\mathbf{u} = \mathbf{v} = \mathbf{1}$ and $\mathbf{M} = (\mathbf{I} - \lambda\mathbf{V})/\lambda$, and (b) identity $\mathbf{V}^\mathsf{T} + \mathbf{V} = \mathbf{1}\mathbf{1}^\mathsf{T}$ (eq. (D.13)), then

$$R(\lambda) \overset{a}{=} \frac{\det((\mathbf{I} - \lambda\mathbf{V}) + \lambda\mathbf{1}\mathbf{1}^\mathsf{T})}{\det(\mathbf{I} - \lambda\mathbf{V})} = \frac{\det(\mathbf{I} + \lambda(\mathbf{1}\mathbf{1}^\mathsf{T} - \mathbf{V}))}{\det(\mathbf{I} - \lambda\mathbf{V})} \overset{b}{=} \frac{\det(\mathbf{I} + \lambda\mathbf{V}^\mathsf{T})}{\det(\mathbf{I} - \lambda\mathbf{V})}.$$

□

A main difference with the stability function of Runge-Kutta methods (B.6) comes from the *explicit* construction (by orthogonal L^2 projection) of the operational matrix of integration $\mathbf{V}$ and representation of the constant function by $\mathbf{1}$ in the chosen basis $\{\phi_n\}$. (not necessarily using polynomials, see example below).

Legendre basis For the Legendre polynomials (an explicit formula for $\mathbf{V}$ is given in section C.4). As expected, we obtain the diagonal Padé approximations of the exponential

$$\begin{aligned}
R_{p=1}(\lambda) &= -\frac{\lambda+2}{\lambda-2} &&= \exp(\lambda) + \mathcal{O}(\lambda^3), \\
R_{p=2}(\lambda) &= \frac{\lambda^2+6\lambda+12}{\lambda^2-6\lambda+12} &&= \exp(\lambda) + \mathcal{O}(\lambda^5), \\
R_{p=3}(\lambda) &= -\frac{\lambda^3+12\lambda^2+60\lambda+120}{\lambda^3-12\lambda^2+60\lambda-120} &&= \exp(\lambda) + \mathcal{O}(\lambda^7), \\
R_{p=4}(\lambda) &= \frac{\lambda^4+20\lambda^3+180\lambda^2+840\lambda+1680}{\lambda^4-20\lambda^3+180\lambda^2-840\lambda+1680} &&= \exp(\lambda) + \mathcal{O}(\lambda^9).
\end{aligned}$$

Cosine basis For comparison, we consider the orthonormal cosine basis $\{1\} \cup \left\{\sqrt{2}\cos(n\pi\tau)\right\}$. Since it only reproduces constant functions, it only yields second order approximations (see proposition 5.3 p.128) but with diminishing error constants.

$$\begin{aligned}
R_{p=1}(\lambda) &= -\frac{\lambda+2}{\lambda-2} &&= \exp(\lambda) + \frac{\lambda^3+\lambda^4}{12} + \mathcal{O}(\lambda^5), \\
R_{p=2}(\lambda) &= \frac{\lambda^2 + \left(\frac{\pi}{2}\right)^4\lambda + 2\left(\frac{\pi}{2}\right)^4}{\lambda^2 - \left(\frac{\pi}{2}\right)^4\lambda + 2\left(\frac{\pi}{2}\right)^4} &&= \exp(\lambda) + \left(\frac{1}{12} - \frac{8}{\pi^4}\right)\left(\frac{\lambda^3+\lambda^4}{12}\right) + \mathcal{O}(\lambda^5), \text{ etc}
\end{aligned}$$

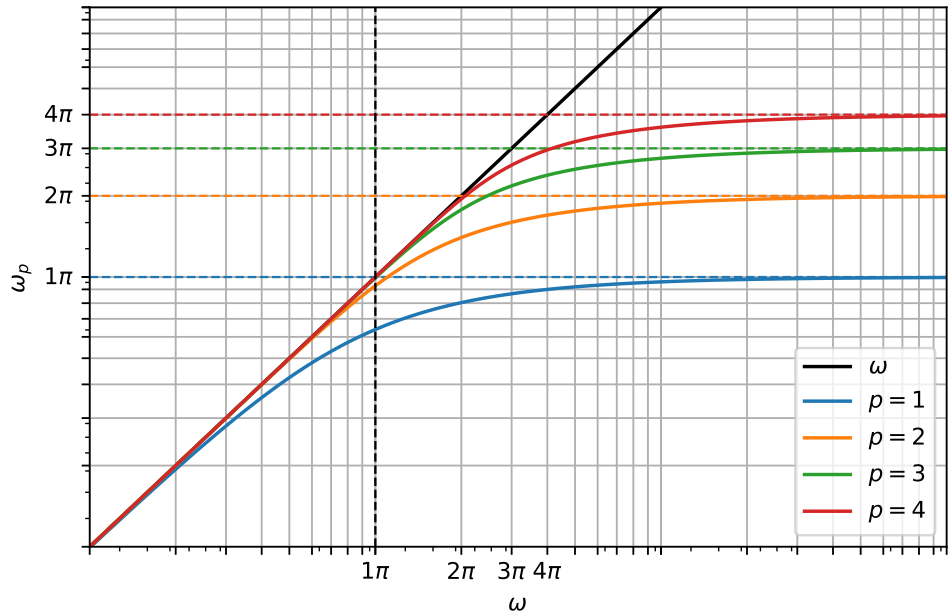


Figure D.2 – (Legendre projection) Frequency warping for a pole $\lambda = i\omega$ with pulsation $\omega_p := \text{unwrap}\left(\text{imag}\left(\ln R_p(i\omega)\right)\right)$

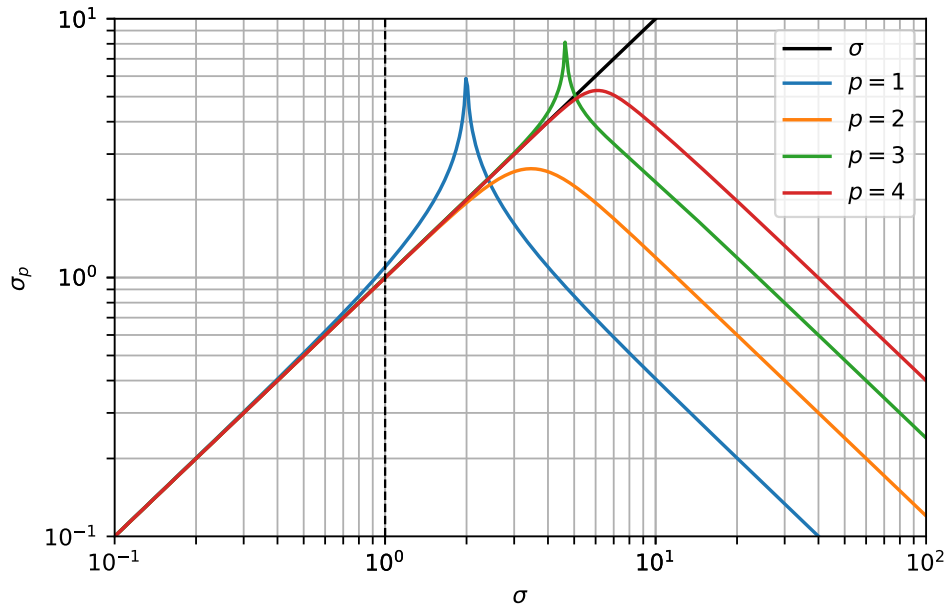


Figure D.3 – (Legendre projection) Dissipative warping for a pole $\lambda = -\sigma$ with dissipation rate $\sigma_p := -\ln\left|R_p(-\sigma)\right|$

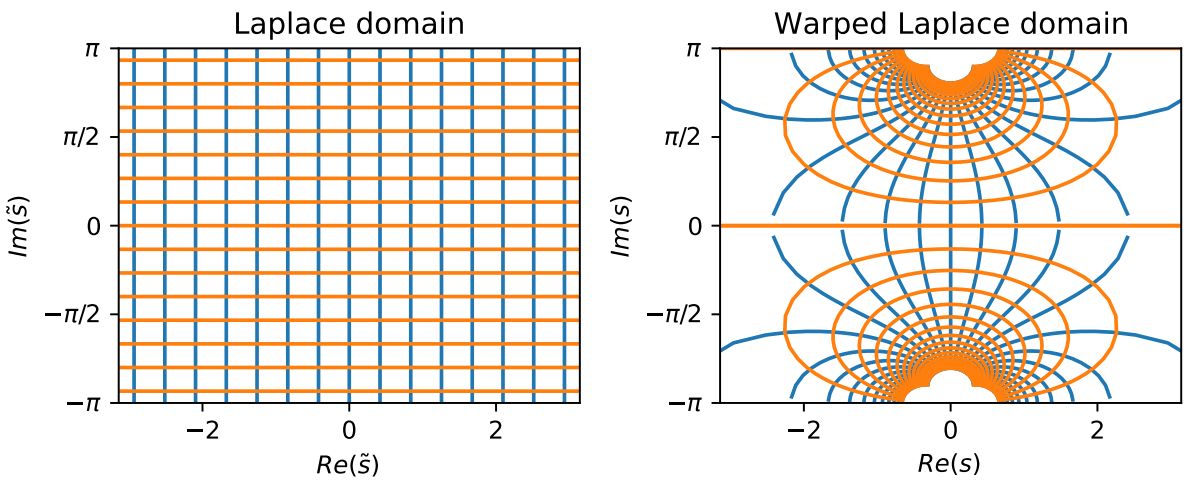


Figure D.4 – (Legendre projection) Laplace conformal map $s/2 = \mathrm{pv}\left(\mathrm{atanh}(\tilde{s}/2)\right)$ corresponding to bilinear, mid-point, AVF and RPM($p = 1$,$k = 0$) where pv denotes the complex principal value). Compression/warping of the Fourier axis $i\mathbb{R}$ to the interval $(-\pi, \pi)$ is noticeable. As a consequence, the rectilinear grid (left plot) is only accurately approximated (right plot) near the origin of the Laplace place ($s = 0$).

Proposition D.1. *Let* $\mathbf{1}$ *and* $\mathbf{V}$ *be respectively the matrix representations of the constant function and of the Volterra operator (as in* (D.12)*). Then, the following identity holds*

$$\mathbf{V}^\mathsf{T} + \mathbf{V} = \mathbf{1}\mathbf{1}^\mathsf{T}. \tag{D.13}$$

It is the finite dimensional equivalent of the functional identity $\mathcal{V} + \mathcal{V}^* = \overline{\mathcal{V}}$ *(eq.* (C.12) *p.284)*

Proof. For an orthonormal basis $\{\phi_i\}$ of $L^2([0,1])$, writing the averaging operator as $\overline{\mathcal{V}} = \int_0^1 = |\,1\rangle\,\langle 1\,|$, the coefficients of its operational matrix satisfy

$$\left[\overline{\mathcal{V}}\right]_{ij} = \left[\left\langle \phi_i \,\middle|\, \overline{\mathcal{V}}\phi_j \right\rangle\right] = \left[\langle \phi_i \,|\, 1\rangle \langle 1 \,|\, \phi_j\rangle\right] = \left[\langle \phi_i \,|\, 1\rangle\right] \left[\langle 1 \,|\, \phi_j\rangle\right] = \mathbf{1}\mathbf{1}^\mathsf{T}.$$

Likewise, using (a) $\overline{\mathcal{V}} = \mathcal{V} + \mathcal{V}^*$ (C.12), (b) linearity, (c) definition of the adjoint (C.7) , (d) definition of $\mathbf{V}$ (D.12), we get

$$\begin{aligned}\left[\overline{\mathcal{V}}\right]_{ij} = \left[\left\langle \phi_i \,\middle|\, \overline{\mathcal{V}}\phi_j \right\rangle\right] &\overset{a}{=} \left[\langle \phi_i \,|\, (\mathcal{V} + \mathcal{V}^*)\phi_j\rangle\right] \overset{b}{=} \left[\langle \phi_i \,|\, \mathcal{V}\phi_j\rangle\right] + \left[\langle \phi_i \,|\, \mathcal{V}^*\phi_j\rangle\right] \\ &\overset{c}{=} \left[\langle \phi_i \,|\, \mathcal{V}\phi_j\rangle\right] + \left[\langle \mathcal{V}\phi_i \,|\, \phi_j\rangle\right] \overset{d}{=} \mathbf{V} + \mathbf{V}^\mathsf{T}.\end{aligned}$$

□

D.8 Proof of Gauss-Legendre quadrature formula

For this thesis to be self-contained (and to highlight the role of the reproducing kernel), we prove that the Gauss-Legendre quadrature formula

$$\int_0^1 f(x)\,\mathrm{d}x = \sum_{k=1}^{n} w_k f(x_k),$$

is exact for all polynomials $f \in \mathbb{P}^{2n-1}$ and that the quadrature weights w_k are given by

$$w_k = \int_0^1 \ell_k(x)\,\mathrm{d}x, \qquad \text{where} \qquad \ell_k(x) = \prod_{i=1, i\neq k}^{n} \frac{x - x_i}{x_k - x_i},$$

are the Lagrange interpolation polynomials at Gauss-Legendre nodes x_k (the roots of $P_n(x)$).

Proof. We prove the result in four steps (a)-(d)

a) Let $n \geq 1$ and $f \in \mathbb{P}^{2n-1}$. Denote polynomials q and r the quotient and remainder of polynomial division of f by the Legendre polynomial P_n, so that we can write f as follows

$$f(x) = P_n(x)q(x) + r(x), \quad q, r \in \mathbb{P}^{n-1}.$$

b) Integrating f, by orthogonality of P_n with $\mathbb{P}^{n-1}$ we have

$$\int_0^1 f(x)\,\mathrm{d}x = \underbrace{\langle P_n, q\rangle}_{=0} + \langle 1, r\rangle = \int_0^1 r(x)\,\mathrm{d}x.$$

c) Likewise, using Gaussian quadrature nodes makes the first sum vanish

$$\sum_{k=1}^{n} w_k f(x_k) = \sum_{k=1}^{n} w_k \underbrace{P_n(x_k)}_{=0} q(x_k) + \sum_{k=1}^{n} w_k r(x_k) = \sum_{k=1}^{n} w_k r(x_k).$$

Note from (a)-(c), we only need to prove that the quadrature of the remainder is exact

$$\int_0^1 r(x)\,\mathrm{d}x = \sum_{k=1}^{n} w_k r(x_k), \quad \forall r \in \mathbb{P}^{n-1}.$$

d) Note that the frame $\{\ell_k\}_{k=1}^n$ is the dual frame to $\big\{K(x_k, \cdot)\big\}_{k=1}^n$ (where $K(x, y)$ is the reproducing kernel of $\mathbb{P}^{n-1}$) satisfying the biorthogonality conditions $\big\langle K(x_i, \cdot), \ell_j\big\rangle = \delta_{ij}$. Then for all $r \in \mathbb{P}^{n-1}$ we have the nodal representation

$$r(x) = \sum_{k=1}^{n} \ell_k(x) r(x_k), \qquad \text{with} \qquad r(x_k) = \big\langle K(x_k, \cdot), r\big\rangle.$$

Integrating r over $[0, 1]$ it comes that quadrature weights w_k are given by the average of the Lagrange interpolation polynomials ℓ_k

$$\int_0^1 r(x)\,\mathrm{d}x = \sum_{k=1}^{n} \underbrace{\left(\int_0^1 \ell_k(x)\,\mathrm{d}x\right)}_{w_k} r(x_k) = \sum_{k=1}^{n} w_k r(x_k).$$

Combining (a)(b)(c)(d) yields the result

$$\int_0^1 f(x)\,\mathrm{d}x = \sum_{k=1}^{n} w_k f(x_k), \quad \forall f \in \mathbb{P}^{2n-1}.$$

□

D.9 Proofs and appendix for section 7.1 (Minimal passive OPA)

D.9.1 Structure of the output equation

Proof. Using the passivity equation (7.8d) p.177, then introducing $V_{\mathsf{cm}}, V_{\mathsf{dm}}$ using (7.3) p.176, factoring $V_{\mathsf{cm}}, V_{\mathsf{dm}}$, finally, for $i_{\mathsf{out}} \neq 0$, dividing by i_{out} and using (7.8c) p.177 one gets the general form for the output equation (7.9) p.177.

$$\begin{aligned}
& i_{\mathsf{S}+} \cdot e_{\mathsf{S}+} + i_{\mathsf{S}-} \cdot e_{\mathsf{S}-} = -i_{\mathsf{out}} \cdot e_{\mathsf{out}} - P_{\mathsf{diss}} \\
\iff\ & i_{\mathsf{S}+}(V_{\mathsf{cm}} + V_{\mathsf{dm}}) + i_{\mathsf{S}-}(V_{\mathsf{cm}} - V_{\mathsf{dm}}) = -i_{\mathsf{out}} \cdot e_{\mathsf{out}} - P_{\mathsf{diss}} \\
\iff\ & V_{\mathsf{cm}}(i_{\mathsf{S}+} + i_{\mathsf{S}-}) + V_{\mathsf{dm}}(i_{\mathsf{S}+} - i_{\mathsf{S}-}) = -i_{\mathsf{out}} \cdot e_{\mathsf{out}} - P_{\mathsf{diss}} \\
\overset{i_{\mathsf{out}} \neq 0}{\iff}\ & V_{\mathsf{cm}} + V_{\mathsf{dm}} \left(\frac{i_{\mathsf{S}+} - i_{\mathsf{S}-}}{i_{\mathsf{S}+} + i_{\mathsf{S}-}} \right) = e_{\mathsf{out}} - \frac{P_{\mathsf{diss}}}{i_{\mathsf{out}}}.
\end{aligned}$$

□

D.9.2 Fixed-point Convergence

According to the Banach fixed-point theorem, existence and uniqueness of the solution are guaranteed if the fixed point (7.47) is contracting, i.e. there exists a Lipschitz constant $\alpha \in [0, 1)$ such that

$$\left\| \phi(\mathbf{x}_1) - \phi(\mathbf{x}_0) \right\| \leq \alpha \|\mathbf{x}_1 - \mathbf{x}_0\| \,. \tag{D.14}$$

A sufficient (but conservative) condition is given by

$$\alpha = 1.162\, G\omega_d < 1. \tag{D.15}$$

Proof. Using (7.46), then the derivative of the discrete gradient (7.42), (bounded by $G/2$), and using the matrix norm of $\mathbf{F}_d\mathbf{C}$, one gets

$$\begin{aligned}
\left\| \phi(\mathbf{x}_1) - \phi(\mathbf{x}_0) \right\|_2 &= \left\| \mathbf{F}_d \left(\overline{\nabla} N(\mathbf{C}\mathbf{x}_0, \mathbf{C}\mathbf{x}_1) - \nabla N(\mathbf{C}\mathbf{x}_0) \right) \right\|_2 \\
&\leq \left\| \mathbf{F}_d \frac{\partial \overline{\nabla} N}{\partial v_1} \mathbf{C} \right\|_2 \|\mathbf{x}_1 - \mathbf{x}_0\|_2 \\
&\leq \|\mathbf{F}_d\mathbf{C}\|_2 \sup_{v_1} \left| \frac{\partial \overline{\nabla} N}{\partial v_1} (v_0, v_1) \right| \|\mathbf{x}_1 - \mathbf{x}_0\|_2 \\
&\leq \frac{2\omega_d \sqrt{\omega_d^2 + 8\omega_d + 20}}{|\omega_d^2 + 2(3 - G)\omega_d + 4|} \frac{G}{2} \|\mathbf{x}_1 - \mathbf{x}_0\|_2 \\
&\leq 1.162\, G\omega_d \|\mathbf{x}_1 - \mathbf{x}_0\|_2
\end{aligned}$$

where the bound 1.162 is obtained numerically by majorizing over $G \in [0, 3]$ and $\omega_d \geq 0$. □

D.9.3 BJT Push-Pull

We detail a (tedious but systematic) derivation for an explicit algebraic (large-signals) relation for the simple BJT push-pull of figure D.5.

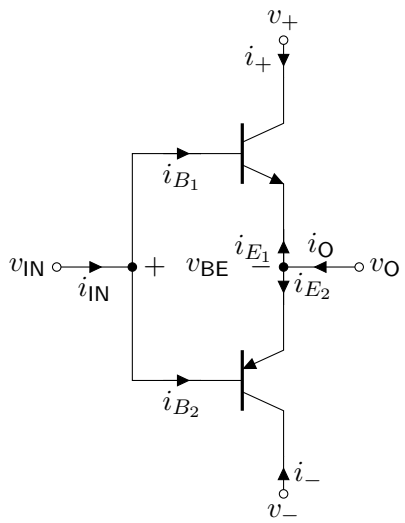


Figure D.5 – (Push-pull) class-B amplifier.

Input - Outputs ports We consider the algebraic relation $\mathbf{u} \mapsto \mathbf{y}$ for the following choice of input and output variables

$$\mathbf{u} = [v_{\mathsf{IN}}, v_+, v_-, i_{\mathsf{O}}]^{\mathsf{T}}, \qquad \mathbf{y} = [i_{\mathsf{IN}}, i_+, i_-, v_{\mathsf{O}}]^{\mathsf{T}}. \tag{D.16}$$

Kirchoff laws From Kirchhoff laws at input and output nodes we have

$$v_{\mathsf{IN}} = v_{B_1} = v_{B_2}, \qquad i_{\mathsf{IN}} = i_{B_1} + i_{B_2}, \qquad v_{\mathsf{O}} = v_{E_1} = v_{E_2}, \qquad i_{\mathsf{O}} = i_{E_1} + i_{E_2}, \tag{D.17a}$$

from which we obtain and the internal voltages

$$v_{\mathsf{BE}} = v_{\mathsf{IN}} - v_{\mathsf{O}}, \qquad v_{\mathsf{BC}_1} = v_{\mathsf{IN}} - v_+, \qquad v_{\mathsf{BC}_2} = v_{\mathsf{IN}} - v_-. \tag{D.17b}$$

Component equations For a given transistor model (here the Ebers-Moll BJT model from (1.47) p.32), there exists functions $\hat{i}_{B_1}, \hat{i}_{C_1}, \hat{i}_{E_1}, \hat{i}_{B_2}, \hat{i}_{C_2}, \hat{i}_{E_2}$ such that

$$i_{B1} = \hat{i}_{B_1}(v_{\mathsf{BC}_1}, v_{\mathsf{BE}}) \qquad i_{E1} = \hat{i}_{E_1}(v_{\mathsf{BC}_1}, v_{\mathsf{BE}}) \qquad i_{C1} = \hat{i}_{C_1}(v_{\mathsf{BC}_1}, v_{\mathsf{BE}}) \tag{D.18a}$$

$$i_{B2} = \hat{i}_{B_2}(v_{\mathsf{BC}_2}, v_{\mathsf{BE}}) \qquad i_{E2} = \hat{i}_{E_2}(v_{\mathsf{BC}_2}, v_{\mathsf{BE}}) \qquad i_{C2} = \hat{i}_{C_2}(v_{\mathsf{BC}_2}, v_{\mathsf{BE}}) \tag{D.18b}$$

Explicit Input-Output map We want to express everything as nonlinear map $\mathbf{y} = F(\mathbf{u})$, i.e. we look for functions $\hat{i}_{\mathsf{IN}}, \hat{i}_+, \hat{i}_-, \hat{v}_{\mathsf{O}}$ such that

$$i_{\mathsf{IN}} = \hat{i}_{\mathsf{IN}}(v_{\mathsf{IN}}, v_+, v_-, i_{\mathsf{O}}), \tag{D.19a}$$

$$i_+ = \hat{i}_+(v_{\mathsf{IN}}, v_+, v_-, i_{\mathsf{O}}), \tag{D.19b}$$

$$i_- = \hat{i}_-(v_{\mathsf{IN}}, v_+, v_-, i_{\mathsf{O}}), \tag{D.19c}$$

$$v_{\mathsf{O}} = \hat{v}_{\mathsf{O}}(v_{\mathsf{IN}}, v_+, v_-, i_{\mathsf{O}}), \tag{D.19d}$$

where according to Kirchhoff laws (D.17a)

$$\hat{i}_{\mathsf{IN}}(v_{\mathsf{IN}}, v_+, v_-, i_{\mathsf{O}}) := \hat{i}_{B_1}(v_{\mathsf{IN}} - v_+, v_{\mathsf{BE}}) + \hat{i}_{B_2}(v_{\mathsf{IN}} - v_-, v_{\mathsf{BE}}), \tag{D.20a}$$

$$\hat{i}_+(v_{\mathsf{IN}}, v_+, v_-, i_{\mathsf{O}}) := \hat{i}_{C_1}(v_{\mathsf{IN}} - v_+, v_{\mathsf{BE}}), \tag{D.20b}$$

$$\hat{i}_-(v_{\mathsf{IN}}, v_+, v_-, i_{\mathsf{O}}) := \hat{i}_{C_2}(v_{\mathsf{IN}} - v_-, v_{\mathsf{BE}}), \tag{D.20c}$$

$$\hat{v}_{\mathsf{O}}(v_{\mathsf{IN}}, v_+, v_-, i_{\mathsf{O}}) := v_{\mathsf{IN}} - v_{\mathsf{BE}}, \tag{D.20d}$$

$$\text{where} \quad v_{\mathsf{BE}} = \hat{v}_{\mathsf{BE}}[v_{\mathsf{IN}}, v_+, v_-](i_{\mathsf{O}}). \tag{D.20e}$$

To obtain an explicit relation, we need a formula for $\hat{v}_{\mathsf{BE}}$ which is defined as the inverse map

$$\hat{v}_{\mathsf{BE}}[v_{\mathsf{IN}}, v_+, v_-](i_{\mathsf{O}}) := \hat{i}_{\mathsf{O}}^{-1}[v_{\mathsf{IN}}, v_+, v_-](i_{\mathsf{O}}), \tag{D.21a}$$

$$\text{where} \quad \hat{i}_{\mathsf{O}}[v_{\mathsf{IN}}, v_+, v_-](v_{\mathsf{BE}}) := \hat{i}_{E_1}(v_{\mathsf{IN}} - v_+, v_{\mathsf{BE}}) + \hat{i}_{E_2}(v_{\mathsf{IN}} - v_+, v_{\mathsf{BE}}). \tag{D.21b}$$

Ebers–Moll Model We want to explicitly characterize $\hat{v}_{\mathsf{BE}}$ for the Ebers–Moll model from (1.47) p.32. Define the adimensioned variables $I_{B1} = i_{B1}/I_S \ldots$, $V_{\mathsf{BE}} = v_{\mathsf{BE}}/V_T$, etc and the adimensioned PN law

$$\mathsf{PN}(V) := \exp(V) - 1. \tag{D.22}$$

Assuming perfectly matched transistors (i.e. $\beta_F^1 = \beta_F^2$, $I_S^1 = I_S^2$), we have the adimensionned laws

$$\hat{I}_{B_1}(V_{\mathsf{BC}_1}, V_{\mathsf{BE}}) = \mathsf{PN}(V_{\mathsf{BE}})/\beta_F + \mathsf{PN}(v_{\mathsf{BC}_1})/\beta_R, \tag{D.23a}$$

$$\hat{I}_{E_1}(V_{\mathsf{BC}_1}, V_{\mathsf{BE}}) = \mathsf{PN}(V_{\mathsf{BC}_1}) - (1 + 1/\beta_F)\mathsf{PN}(V_{\mathsf{BE}}), \tag{D.23b}$$

$$\hat{I}_{C_1}(V_{\mathsf{BC}_1}, V_{\mathsf{BE}}) = \mathsf{PN}(V_{\mathsf{BE}}) - (1 + 1/\beta_R)\mathsf{PN}(V_{\mathsf{BC}_1}), \tag{D.23c}$$

$$\hat{I}_{B_2}(V_{\mathsf{BC}_2}, V_{\mathsf{BE}}) = -\mathsf{PN}(-V_{\mathsf{BE}})/\beta_F - \mathsf{PN}(-v_{\mathsf{BC}_2})/\beta_R, \tag{D.23d}$$

$$\hat{I}_{E_2}(V_{\mathsf{BC}_2}, V_{\mathsf{BE}}) = -\mathsf{PN}(-V_{\mathsf{BC}_2}) + (1 + 1/\beta_F)\mathsf{PN}(-V_{\mathsf{BE}}), \tag{D.23e}$$

$$\hat{I}_{C_2}(V_{\mathsf{BC}_2}, V_{\mathsf{BE}}) = -\mathsf{PN}(-V_{\mathsf{BE}}) + (1 + 1/\beta_R)\mathsf{PN}(-V_{\mathsf{BC}_2}). \tag{D.23f}$$

Substituting these relation in (D.21b), yields $\hat{I}_{\mathsf{O}}$ as a function of V_{BE}

$$\begin{aligned}\hat{I}_{\mathsf{O}}[V_{\mathsf{IN}}, V_+, V_-](V_{\mathsf{BE}}) &= \mathsf{PN}(V_{\mathsf{BC}_1}) - \mathsf{PN}(-V_{\mathsf{BC}_2}) - (1 + 1/\beta_F)(\mathsf{PN}(V_{\mathsf{BE}}) - \mathsf{PN}(-V_{\mathsf{BE}})) \\ &= \mathsf{PN}(V_{\mathsf{IN}} - V_+) - \mathsf{PN}(V_- - V_{\mathsf{IN}}) - (1 + 1/\beta_F)2\sinh(V_{\mathsf{BE}}).\end{aligned}$$

Inverting this function, we obtain the explicit form for $\hat{V}_{\mathsf{BE}}$ as a function of I_{O}.

$$\hat{V}_{\mathsf{BE}}[V_{\mathsf{IN}}, V_+, V_-](I_{\mathsf{O}}) = \operatorname{asinh}\left(\frac{\mathsf{PN}(V_{\mathsf{IN}} - V_+) - \mathsf{PN}(V_- - V_{\mathsf{IN}}) - I_{\mathsf{O}}}{2(1 + 1/\beta_F)}\right) \tag{D.24}$$

By consequence, we finally obtain the explicit output law as a function of input variables

$$V_{\mathsf{O}} = V_{\mathsf{IN}} - \operatorname{asinh}\left(\frac{\exp(V_{\mathsf{IN}} - V_+) - \exp(V_- - V_{\mathsf{IN}}) - I_{\mathsf{O}}}{2(1 + 1/\beta_F)}\right). \tag{D.25}$$

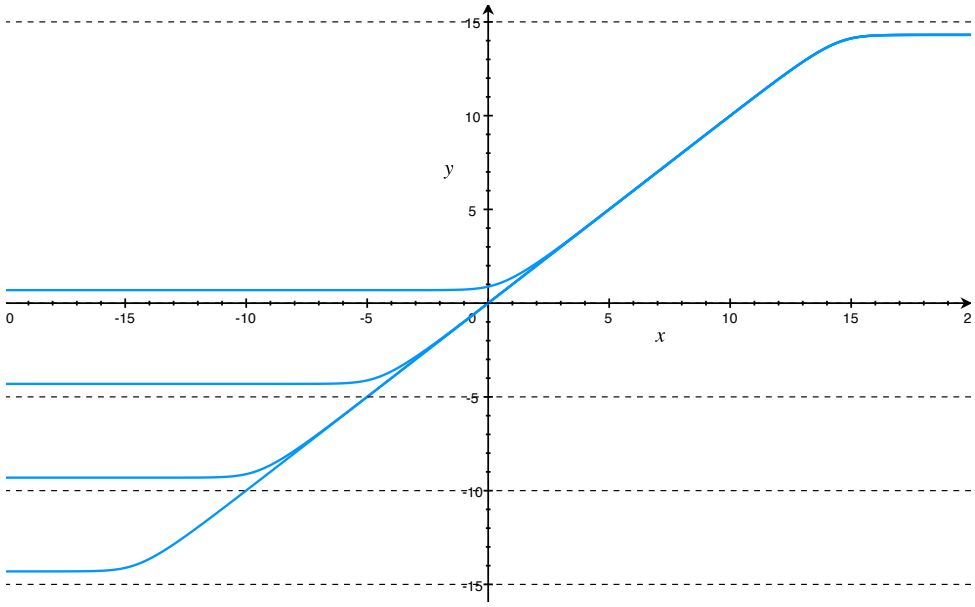
It can be factored as

$$V_{\mathsf{O}} = V_{\mathsf{IN}} - \operatorname{asinh}\left(\frac{1}{1 + 1/\beta_F}\left(\sinh\left(V_{\mathsf{IN}} - \frac{V^+ + V^-}{2}\right)\exp\left(-\frac{V^+ - V^-}{2}\right) - I_{\mathsf{O}}\right)\right). \tag{D.26}$$

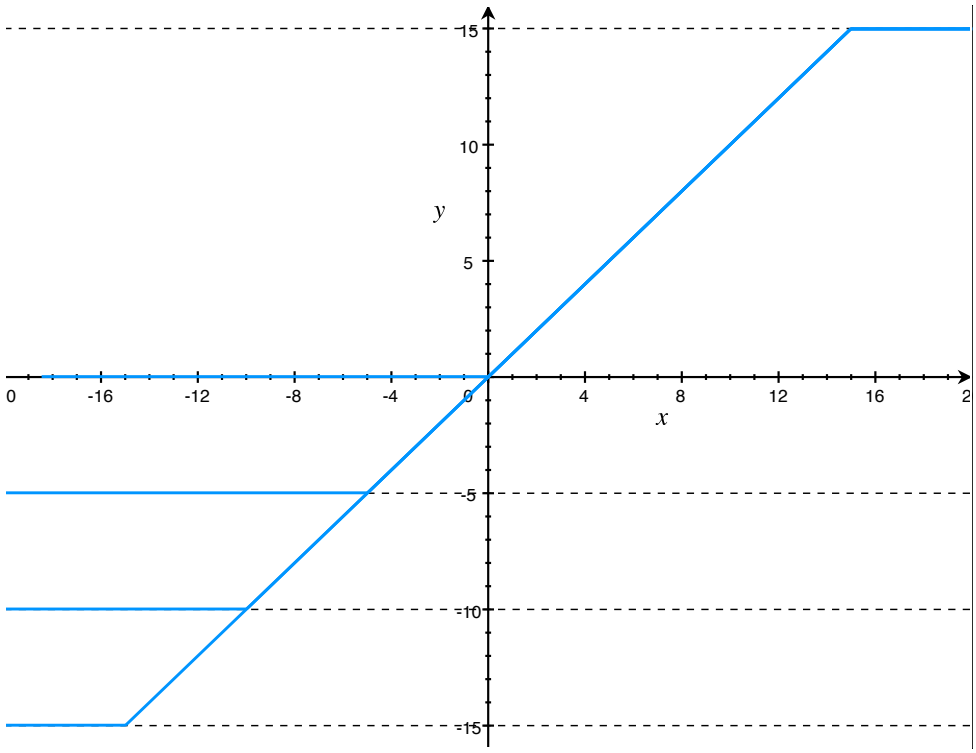
Assuming the symmetric power supply case $V^+ = -V^-$ and $I_{\mathsf{O}} \approx 0$, this simplifies to

$$V_{\mathsf{O}} = V_{\mathsf{IN}} - \operatorname{asinh}\left(\frac{\exp(-V^+)\sinh(V_{\mathsf{IN}})}{1 + 1/\beta_F}\right). \tag{D.27}$$

This is similar to (but different from) tanh as shown on figure D.6.



(a) adimensioned



(b) dimensioned

Figure D.6 – (Push-pull circuit) Output functions $\hat{v}_{\mathsf{O}}(v_{\mathsf{IN}})$ (dimensioned) and $\hat{V}_{\mathsf{O}}(V_{\mathsf{IN}})$ (adimensionned) for $V^+ = 15$, $V^- \in \{-15, -10, -5, 0\}$, for $I_{\mathsf{O}} = 0$.

D.10 Z-domain response of Legendre projection filterbank (linear state-space system)

D.10.1 Reminder on the Kronecker product

We recall the definition of the Kronecker product and its main properties.

Definition D.1 (Kronecker product)**.** Let $\mathbf{A}$ be an $m \times n$ matrix, and $\mathbf{B}$ a $p \times q$ matrix, then the Kronecker product $\mathbf{A} \otimes \mathbf{B}$ is the $pm \times qn$ block matrix

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & \dots & a_{1n}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{m1}\mathbf{B} & \dots & a_{mn}\mathbf{B} \end{bmatrix}. \tag{D.28}$$

Property D.1 (Kronecker product properties)**.** For suitable matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$ and scalar k, the Kronecker product is non-commutative and satisfies the following properties

- Bilinearity and associativity

$$\mathbf{A} \otimes (\mathbf{B} + \mathbf{C}) = \mathbf{A} \otimes \mathbf{B} + \mathbf{A} \otimes \mathbf{C}, \tag{D.29a}$$
$$(\mathbf{B} + \mathbf{C}) \otimes \mathbf{A} = \mathbf{B} \otimes \mathbf{A} + \mathbf{C} \otimes \mathbf{A}, \tag{D.29b}$$
$$(k\mathbf{A}) \otimes \mathbf{B} = \mathbf{A} \otimes (k\mathbf{B}) = k(\mathbf{A} \otimes \mathbf{B}), \tag{D.29c}$$
$$(\mathbf{A} \otimes \mathbf{B}) \otimes \mathbf{C} = \mathbf{A} \otimes (\mathbf{B} \otimes \mathbf{C}), \tag{D.29d}$$
$$\mathbf{A} \otimes \mathbf{0} = \mathbf{0} \otimes \mathbf{A} = \mathbf{0}. \tag{D.29e}$$

- Mixed product property: for suitable matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$

$$(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = (\mathbf{AC}) \otimes (\mathbf{BD}). \tag{D.30}$$

- Distributivity: $\mathbf{A} \otimes \mathbf{B}$ is invertible iff $\mathbf{A}$ and $\mathbf{B}$ are invertible, then the inverse, Moore-pseudo inverse, adjoint and transpose operators are distributive over the Kronecker product, i.e.

$$(\mathbf{A} \otimes \mathbf{B})^{-1} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1}, \tag{D.31a}$$
$$(\mathbf{A} \otimes \mathbf{B})^{\dagger} = \mathbf{A}^{\dagger} \otimes \mathbf{B}^{\dagger}, \tag{D.31b}$$
$$(\mathbf{A} \otimes \mathbf{B})^{*} = \mathbf{A}^{*} \otimes \mathbf{B}^{*}, \tag{D.31c}$$
$$(\mathbf{A} \otimes \mathbf{B})^{\mathsf{T}} = \mathbf{A}^{\mathsf{T}} \otimes \mathbf{B}^{\mathsf{T}}. \tag{D.31d}$$

D.10.2 proof of proposition 8.1

The proof of proposition 8.1 p.225 is detailed below.

Remark D.2 (Notations in this proof)**.** This proof uses Kronecker products whose properties are recalled in subsection D.10.1 below. Indeed we have to blend finite-dimensional representation of functional operators (matrices $\mathbf{I}_p, \mathbf{V}_p, \mathbf{e}_0$) with matrices from state-spaces systems $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$. To this end, we use matrix algebra (instead of multi-dimensional tensor

algebra) with large vectors and block matrices such as $\mathbf{I}_p \otimes \mathbf{A}$ built from Kronecker products.

Let $\boldsymbol{f}(t)$ be a q-dimensional vector-valued function of time with scalar components $\boldsymbol{f}^i(t)$. We denote by $\boldsymbol{f}_n(\tau)$ the sequence of functions $\boldsymbol{f}_n(\tau) = \boldsymbol{f}(n+\tau)|_{\tau\in(0,1)}$ at time frame n and by $\vec{\mathbf{f}}[n]$ its sequence of projection coefficients in the Legendre polynomial basis such that

$$\vec{\mathbf{f}}[n] = \begin{bmatrix} \langle P_0 \,|\, \boldsymbol{f}_n \rangle \\ \vdots \\ \langle P_{p-1} \,|\, \boldsymbol{f}_n \rangle \end{bmatrix} = \begin{bmatrix} \langle P_0 \,| \\ \vdots \\ \langle P_{p-1} \,| \end{bmatrix} \otimes |\, \boldsymbol{f}_n \rangle \,.$$

With the chosen convention, matrices and vectors corresponding to the Legendre representation of functional operators

$$\mathbf{I}_p = \left[\langle P_m \,|\, P_n \rangle\right], \qquad \mathbf{V}_p = \left[\left\langle P_m \,\middle|\, \int_0^\tau P_n \right\rangle\right], \qquad \mathbf{e}_0 = \left[\langle P_m \,|\, 1 \rangle\right],$$

are written to the left of the Kronecker product while state-space matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$ are written to the right. Matrix $\mathbf{I}_p$ is the identity of the Legendre space of order p, column vector $\mathbf{e}_0$ represents the synthesis operator $|\,1\rangle$. Its transpose $\mathbf{e}_0^\mathsf{T}$ represents the dual analysis operator $\langle 1\,|$. Matrix $\mathbf{V}_p$ is the Legendre operational matrix of integration defined in (C.18).

Proof. Consider a state-space system of dimension n_x defined by matrices $\mathbf{A} \in \mathbb{R}^{n_x \times n_x}, \mathbf{B} \in \mathbb{R}^{n_x \times 1}, \mathbf{C} \in \mathbb{R}^{1 \times n_x}, \mathbf{D} \in \mathbb{R}^{1\times 1}$. For each time index n and time frame $\Omega_n = (n, n+1)$, with initial condition $\mathbf{x}_n$, the local representation for normalized time $\tau \in (0,1)$ is given by the following equations

$$\begin{cases} \dot{\boldsymbol{X}}_n(\tau) &= \mathbf{A}\boldsymbol{X}_n(\tau) + \mathbf{B}\boldsymbol{u}_n(\tau), \\ \boldsymbol{y}_n(\tau) &= \mathbf{C}\boldsymbol{X}_n(\tau) + \mathbf{D}\boldsymbol{u}_n(\tau), \end{cases}, \quad \text{where} \quad \begin{cases} \boldsymbol{X}_n(\tau) &= \mathbf{x}_n + \int_0^\tau \dot{\boldsymbol{X}}_n(\sigma)\,\mathrm{d}\sigma, \\ \mathbf{x}_{n+1} &= \boldsymbol{X}_n(\tau = 1) \end{cases} \tag{D.32}$$

Step 1: Legendre projection. Denote $\vec{\mathbf{d}}[n], \vec{\mathbf{y}}[n], \vec{\mathbf{X}}[n], \vec{\mathbf{u}}[n]$ the Legendre projection coefficients of functions $\dot{\boldsymbol{X}}_n, \boldsymbol{y}_n, \boldsymbol{X}_n, \boldsymbol{u}_n$ and denote $\mathbf{x}[n] \equiv \mathbf{x}_n$. The four equations in (D.32), expressed in terms of Legendre coefficients, directly translate to the discrete system

$$\vec{\mathbf{d}}[n] = (\mathbf{I}_p \otimes \mathbf{A})\vec{\mathbf{X}}[n] + (\mathbf{I}_p \otimes \mathbf{B})\vec{\mathbf{u}}[n], \tag{D.33a}$$

$$\vec{\mathbf{y}}[n] = (\mathbf{I}_p \otimes \mathbf{C})\vec{\mathbf{X}}[n] + (\mathbf{I}_p \otimes \mathbf{D})\vec{\mathbf{u}}[n], \tag{D.33b}$$

$$\vec{\mathbf{X}}[n] = (\mathbf{e}_0 \otimes \mathbf{I})\mathbf{x}[n] + (\mathbf{V} \otimes \mathbf{I})\vec{\mathbf{d}}[n], \tag{D.33c}$$

$$\mathbf{x}[n+1] = \mathbf{x}[n] + (\mathbf{e}_0^\mathsf{T} \otimes \mathbf{I})\vec{\mathbf{d}}[n]. \tag{D.33d}$$

Equations (D.33a)-(D.33b) are just higher-dimensional embeddings. In equations (D.33c)-(D.33d), operators $|\,1\rangle$, $\int_0^\tau$, $\langle 1\,|$ are respectively replaced by matrices $\mathbf{e}_0$, $\mathbf{V}_p$, $\mathbf{e}_0^\mathsf{T}$ compared to (D.32).

Step 2: Z-domain representation: Denote by $\widehat{\boldsymbol{d}}, \widehat{\boldsymbol{y}}, \widehat{\boldsymbol{X}}, \widehat{\boldsymbol{u}}, \widehat{\boldsymbol{x}}$ the Z-transform of sequences $\vec{\mathbf{d}}, \vec{\mathbf{y}}, \vec{\mathbf{X}}, \vec{\mathbf{u}}, \mathbf{x}$. By linearity of the Z-transform, equations (D.33a)-(D.33d) become the Z-domain system

$$\widehat{\boldsymbol{d}}(z) = (\mathbf{I}_p \otimes \mathbf{A})\widehat{\boldsymbol{X}}(z) + (\mathbf{I}_p \otimes \mathbf{B})\widehat{\boldsymbol{u}}(z), \tag{D.34a}$$

$$\widehat{\boldsymbol{y}}(z) = (\mathbf{I}_p \otimes \mathbf{C})\widehat{\boldsymbol{X}}(z) + (\mathbf{I}_p \otimes \mathbf{D})\widehat{\boldsymbol{u}}(z), \tag{D.34b}$$

$$\widehat{\boldsymbol{X}}(z) = (\mathbf{e}_0 \otimes \mathbf{I})\widehat{\boldsymbol{x}}(z) + (\mathbf{V}_p \otimes \mathbf{I})\widehat{\boldsymbol{d}}(z), \tag{D.34c}$$

$$z\widehat{\boldsymbol{x}}(z) = \widehat{\boldsymbol{x}}(z) + (\mathbf{e}_0 \otimes \mathbf{I})^\mathsf{T}\widehat{\boldsymbol{d}}(z). \tag{D.34d}$$

Step 3: Z-domain transfer function We solve the linear system of equations (D.34a)-(D.34d) to obtain the matrix-valued transfer function $\widehat{\boldsymbol{H}}_p : \widehat{\boldsymbol{u}}(z) \mapsto \widehat{\boldsymbol{y}}(z)$ as follows.

Solving (D.34d) for the Z-transform of boundary values $\widehat{\boldsymbol{x}}(z)$ we obtain

$$\widehat{\boldsymbol{x}}(z) = \frac{1}{z-1}(\mathbf{e}_0 \otimes \mathbf{I})^\mathsf{T} \widehat{\boldsymbol{d}}(z). \tag{D.35}$$

Back subtitution of $\widehat{\boldsymbol{x}}$ in (D.34c) yields the Z-transform of the trajectory coefficients in term of its vector field coefficients $\widehat{\boldsymbol{d}}$

$$\widehat{\boldsymbol{X}}(z) = \left(\left(\frac{\mathbf{e}_0 \mathbf{e}_0^\mathsf{T}}{z-1} + \mathbf{V}_p \right) \otimes \mathbf{I} \right) \widehat{\boldsymbol{d}}(z). \tag{D.36}$$

Back substitution of $\widehat{\boldsymbol{X}}$ in (D.34a) leads to the implicit equation on $\widehat{\boldsymbol{d}}$

$$\widehat{\boldsymbol{d}}(z) = (\mathbf{I}_p \otimes \mathbf{A}) \left(\left(\frac{\mathbf{e}_0 \mathbf{e}_0^\mathsf{T}}{z-1} + \mathbf{V}_p \right) \otimes \mathbf{I} \right) \widehat{\boldsymbol{d}}(z) + (\mathbf{I}_p \otimes \mathbf{B}) \widehat{\boldsymbol{u}}(z).$$

Using the mixed Kronecker product property (D.30) and solving for $\widehat{\boldsymbol{d}}$ yields

$$\widehat{\boldsymbol{d}}(z) = \left(\mathbf{I}_p \otimes \mathbf{I} - \left(\frac{\mathbf{e}_0 \mathbf{e}_0^\mathsf{T}}{z-1} + \mathbf{V}_p \right) \otimes \mathbf{A} \right)^{-1} (\mathbf{I}_p \otimes \mathbf{B}) \widehat{\boldsymbol{u}}(z).$$

Back-subtitution of $\widehat{\boldsymbol{d}}$ in (D.36) yields the explicit expression of $\widehat{\boldsymbol{X}}$ in term of $\widehat{\boldsymbol{u}}$

$$\widehat{\boldsymbol{X}}(z) = \left(\left(\frac{\mathbf{e}_0 \mathbf{e}_0^\mathsf{T}}{z-1} + \mathbf{V}_p \right) \otimes \mathbf{I} \right) \left(\mathbf{I}_p \otimes \mathbf{I} - \left(\frac{\mathbf{e}_0 \mathbf{e}_0^\mathsf{T}}{z-1} + \mathbf{V}_p \right) \otimes \mathbf{A} \right)^{-1} (\mathbf{I}_p \otimes \mathbf{B}) \widehat{\boldsymbol{u}}(z).$$

Finally, back-substitution of $\widehat{\boldsymbol{X}}$ in the output equation (D.34b), yields the input-output mapping

$$\widehat{\boldsymbol{y}}(z) = \left(\left(\left(\frac{\mathbf{e}_0 \mathbf{e}_0^\mathsf{T}}{z-1} + \mathbf{V}_p \right) \otimes \mathbf{C} \right) \left(\mathbf{I}_p \otimes \mathbf{I} - \left(\frac{\mathbf{e}_0 \mathbf{e}_0^\mathsf{T}}{z-1} + \mathbf{V}_p \right) \otimes \mathbf{A} \right)^{-1} (\mathbf{I}_p \otimes \mathbf{B}) + (\mathbf{I}_p \otimes \mathbf{D}) \right) \widehat{\boldsymbol{u}}(z).$$

The Z-transform of the $p \times p$ Legendre filterbank representing the linear state-space system is thus

$$\widehat{\boldsymbol{H}}_p(z) = \left(\left(\frac{\mathbf{e}_0 \mathbf{e}_0^\mathsf{T}}{z-1} + \mathbf{V}_p \right) \otimes \mathbf{C} \right) \left(\mathbf{I}_p \otimes \mathbf{I} - \left(\frac{\mathbf{e}_0 \mathbf{e}_0^\mathsf{T}}{z-1} + \mathbf{V}_p \right) \otimes \mathbf{A} \right)^{-1} (\mathbf{I}_p \otimes \mathbf{B}) + (\mathbf{I}_p \otimes \mathbf{D}).$$

□

Remark D.3. Note that formula (D.35) is the Z-domain representation of the *cumulative sum of average vector fields* linking the initial condition at time n to the history of average vector fields over all previous time frames. Indeed $1/(z-1)$ is the Z-domain representation of the cumulative sum operator, and the term $(\mathbf{e}_0 \otimes \mathbf{I})^\mathsf{T} \widehat{\boldsymbol{d}}(z)$ is the Z-domain equivalent of the average vector field: $\mathbf{e}_0^\mathsf{T}$ selects the 0-th coefficient from the vector field coefficients $\widehat{\boldsymbol{d}}(z)$

Appendix E

Code listing (SPAC methods)

```
"""
    Plots
    - power balanced regions and
    - power balanced values of collocation parameter \alpha
      as a function of dissipation parameter \sigma
"""

from pylab import *
import sympy as sp

t,alpha = sp.symbols("t alpha", real=True)
a = sp.symbols("lambda")

def f(x):
    """ linear complex vector field """
    return a * x

def B(u, m, tau):
    """ Compute the Boundary operator """
    return sp.diff(u,t,m).subs(t, tau)

def get_R():
    """"
        find the stability function R(z) and
        its continuous extension R_t(z) such that

            X(t) = R_t(z) x_0 for t in [0,1]
    """
    d = 3
    c = sp.symbols("c0:%d" % (d+1))              # tuple of coefficients
    xt = sum([c[i] * t**i for i in range(d+1)]) # polynomial x(t)
    E = sp.diff(xt,t) - f(xt)                    # vector field error
    eqs = [
        sp.Eq(xt.subs(t,0), 1),                  # initial condition
        sp.Eq(B(E, 0, alpha), 0),                # adaptive collocation point
        sp.Eq(B(E, 0, 0), 0),                    # left boundary condition
        sp.Eq(B(E, 0, 1), 0),                    # right boundary condition
    ]
    sol = sp.solve(eqs, c)
    Rt = xt.subs(sol)
    R = Rt.subs(t,1).together().collect(a)
    return Rt, R

def inner_product(u,v):
```

```
    """ Compute the L^2 inner product """
    return sp.integrate(sp.conjugate(u) * v, (t,0,1))

def power_balance(x):
    """ Compute the power balance functional \rho(x) """
    dx = sp.diff(x,t); E = dx - f(x)
    rho = inner_product(x, E).together()
    return rho

def plot_power_balanced_dissipation(Rt):
    rho = power_balance(Rt).simplify()
    num,den = sp.fraction(rho)  #; sp.pprint(num)
    sol = sp.solve(sp.Eq(sp.re(num),0), alpha)
    sp.pprint(sol)
    figure(figsize=(5,4))
    Max = 6.7
    i = 0
    for s in sol:
        s = sp.lambdify(a,s)
        sigma = Max * arange(0, 200) / 200
        plot(sigma, s(sigma), 'C%d' % i)
        plot(-sigma, s(-sigma), 'C%d' % i)
        ylim([0,1])
        i += 1
    axhline(0.5, ls='--', c='k', lw=1)
    axvline(-Max, ls='--', c='k', lw=1)
    axvline(Max, ls='--', c='k', lw=1)
    xlabel("Dissipation rate $\\sigma$")
    ylabel("collocation point $\\alpha$")
    grid()
    tight_layout()
    savefig("alpha_sigma_PAC(1).pdf", bbox_inches="tight")
    show()

def plot_power_balanced_region(Rt):
    rho = power_balance(Rt).simplify()
    num,den = sp.fraction(rho)
    sp.pprint(num)
    sol = sp.solve(sp.Eq(sp.re(num),0), alpha)
    sp.pprint(sol)
    sol = [sp.lambdify(a,s) for s in sol]

    M = pi
    x = linspace(-M, M, 200)
    y = linspace(-M, M, 200)
    X,Y = meshgrid(x,y)
    Z = [zeros_like(X)] * len(sol)
    figure(figsize=(5,4))
    for i in range(X.shape[0]):
        for j in range(Y.shape[1]):
            pole = X[i,j] + 1j * Y[i,j]
            z = NaN
            for k in range(len(sol)):
                s = sol[k]
                zs = s(pole)
                if abs(imag(zs)) < 1e-20 and real(zs) >= 0 and real(zs) <= 1:
                    z = zs - 0.5
                Z[k][i,j] = z
    contourf(X,Y,Z[0], antialiased=True, alpha=0.9,cmap=plt.get_cmap("seismic"))
    contourf(X,Y,Z[1], antialiased=True, alpha=0.9,cmap=plt.get_cmap("seismic"))
    grid()
```

```
    xlabel("Dissipation rate $\\sigma$")
    ylabel("Pulsation $\\omega$")
    tight_layout()
    savefig("PB_region_PAC(1).pdf", bbox_inches="tight")
    show()

Rt,R = get_R()                                  # get stability function R(z)
error = sp.series(sp.exp(a) - R, a).simplify() # Taylor series expansion
sp.pprint(R)
sp.pprint(error)
plot_power_balanced_dissipation(Rt)
plot_power_balanced_region(Rt)
```

Listing E.1 – "Code for plotting SPAC(1) Power Balance region"

Appendix F

Geometric Algebra

Here we gather a collection of definition, theorem and properties related to Geometric Algebra and Geometric calculus. Our main references are [Mac10, Mac12b] and [DGL+03].

F.1 Algebra

F.1.1 Inner product spaces

Definition F.1 (Inner product)**.** If $\mathbf{u} = (u_1, \ldots, u_n)$, $\mathbf{v} = (v_1, \ldots, v_n)$ are vectors in $\mathbb{R}^n$, then their *inner product* is

$$\mathbf{u} \cdot \mathbf{v} := \sum_{i=1}^{n} u_i v_i. \tag{F.1}$$

Theorem F.1 (inner product properties)**.** *If* $\mathbf{u}, \mathbf{v}, \mathbf{w}$ *are vectors in* $\mathbb{R}^n$ *and* a *is a scalar in* $\mathbb{R}$*, then*

P1. $(a\mathbf{u}) \cdot \mathbf{v} = a(\mathbf{u} \cdot \mathbf{v})$

P2. $(\mathbf{u} + \mathbf{v}) \cdot \mathbf{w} = \mathbf{u} \cdot \mathbf{w} + \mathbf{v} \cdot \mathbf{w}$

P3. $\mathbf{u} \cdot \mathbf{v} = \mathbf{v} \cdot \mathbf{u}$

P4. *if* $\mathbf{v} \neq 0$, $\mathbf{v} \cdot \mathbf{v} > 0$

Definition F.2 (inner product space)**.** An *inner product space* is a vector space with a product called an inner product satisfying axioms P1-P4 of Theorem F.1.

Definition F.3 (Norm)**.** . The *norm* $|\mathbf{v}|$ of a vector $\mathbf{v}$ in an inner product space is given by $|\mathbf{v}|^2 = \mathbf{v} \cdot \mathbf{v}$.

Definition F.4 (angle)**.** The angle $\theta \in [0, \pi]$ between nonzero vectors $\mathbf{u}, \mathbf{v}$ in an inner

product space is defined by

$$\theta := \frac{\mathbf{u} \cdot \mathbf{v}}{|\mathbf{u}||\mathbf{v}|}. \tag{F.2}$$

Definition F.5 (Orthogonal vectors)**.** Let $\mathbf{u}$, $\mathbf{v}$ be vectors in an inner product space. Then $\mathbf{u}$ and $\mathbf{v}$ are *orthogonal* iff $\mathbf{u} \cdot \mathbf{v} = 0$.

Theorem F.2 (Pythagorean theorem)**.** *if* $\mathbf{u}$ *and* $\mathbf{v}$ *are orthogonal vectors then*

$$|\mathbf{u}^2 + \mathbf{v}^2| = |\mathbf{u}|^2 + |\mathbf{v}|^2. \tag{F.3}$$

Lemma F.1 (Cauchy-Schwartz inequality)**.** *if* $\mathbf{u}$ *and* $\mathbf{v}$ *are vectors in* $\mathbb{R}^n$ *then*

$$|\mathbf{u} \cdot \mathbf{v}| \leq |\mathbf{u}||\mathbf{v}|. \tag{F.4}$$

F.1.2 Geometric Algebra

Definition F.6 (Oriented length)**.** An *oriented length* $\mathbf{v}$ is and oriented segment of a line. The length of $\mathbf{v}$ is called its norm $|\mathbf{v}|$.

Definition F.7 (Oriented Area)**.** An *oriented area* $\mathbf{B}$ is and oriented segment of a plane (i.e. an area). The area of $\mathbf{B}$ is called its norm $|\mathbf{B}|$.

Theorem F.3. *Oriented areas in* $\mathbb{R}^n$ *form a vector space.*

Definition F.8 (Oriented Solid)**.** An *oriented solid* $\mathbf{T}$ is and oriented segment of a three dimensional space (i.e. a volume). The volume of $\mathbf{T}$ is called its norm $|\mathbf{T}|$.

Definition F.9 (Outer product)**.** The outer product denoted $\wedge$ is an operation satisfying the following properties

$$\mathbf{u} \wedge \mathbf{u} = 0, \tag{F.5}$$

$$\mathbf{u} \wedge \mathbf{v} = -\mathbf{v} \wedge \mathbf{u} \tag{F.6}$$

$$a(\mathbf{u} \wedge \mathbf{v}) = (a\mathbf{u}) \wedge \mathbf{v} \tag{F.7}$$

$$(\mathbf{u} + \mathbf{v}) \wedge \mathbf{w} := \mathbf{u} \wedge \mathbf{w} + \mathbf{v} \wedge \mathbf{w} \tag{F.8}$$

Theorem F.4. *Let* $\mathbf{e}_1, \mathbf{e}_2$ *be an orthonormal basis for a plane. Orient the plane with* $\mathbf{e}_1 \wedge \mathbf{e}_2$*. Let* $\mathbf{u}$ *and* $\mathbf{v}$ *be vectors in the plane. Let* $\theta \in (-\pi, \pi]$ *be the oriented angle from* $\mathbf{u}$ *to* $\mathbf{v}$*, then*

$$\mathbf{u} \wedge \mathbf{v} = |\mathbf{u}||\mathbf{v}| \sin\theta \, (\mathbf{e}_1 \wedge \mathbf{e}_2) \tag{F.9}$$

Theorem F.5 (Oriented area basis)**.** *Let $\mathbf{e}_1 \dots \mathbf{e}_n$ be an orthonormal basis of $\mathbb{R}^n$, then the oriented areas $\{\mathbf{e}_1 \wedge \mathbf{e}_2, \mathbf{e}_2 \wedge \mathbf{e}_3, \dots \mathbf{e}_n \wedge \mathbf{e}_1\}$ form a basis of the vector space of oriented areas.*

Definition F.10 (Geometric product)**.** The geometric of two vector $\mathbf{u}, \mathbf{v}$ is defined by

$$\mathbf{uv} := \mathbf{u} \cdot \mathbf{v} + \mathbf{u} \wedge \mathbf{v}. \tag{F.10}$$

Theorem F.6 (Geometric vector space $\mathbb{G}^n$)**.** *The inner product space $\mathbb{R}^n$ can be extended to the geometric algebra $\mathbb{G}^n$. Members of $\mathbb{G}^n$ are called multivectors. The geometric algebra is a vector space with a product called the* geometric product*.*

The geometric product of multivectors A and B is written AB. For all scalars a and multivectors A, B, C:

G0. $AB \in \mathbb{G}^n$.

G1. $A(B + C) = AB + AC$, $(B + C)A = BA + CA$.

G2. $(aAB) = A(aB) = a(AB)$.

G3. $A(BC) = (AB)C$.

G4. $1\,A = A\,1$.

G5. The geometric product of $\mathbb{G}^n$ is linked to the inner product of $\mathbb{R}^n$:

$$\mathbf{uu} = \mathbf{u} \cdot \mathbf{u} = |\mathbf{u}|^2, \quad \forall \mathbf{u} \in \mathbb{R}^n. \tag{F.11}$$

G6. Every orthonormal basis of $\mathbb{R}^n$ determines a canonical basis for the vector space $\mathbb{G}^n$.

G7. The k-vectors in a canonical basis for $\mathbb{G}^n$ form a basis for the subspace of k-vectors in $\mathbb{G}^n$, for $k \in \{0 \dots n\}$. Every multivector can be uniquely expressed as a sum of k-vectors.

Property F.1 (Symmetric inner product)**.** For all vectors $\mathbf{u}, \mathbf{v}$ in $\mathbb{R}^n$

$$\mathbf{u} \cdot \mathbf{v} = \frac{\mathbf{uv} + \mathbf{vu}}{2}. \tag{F.12}$$

Property F.2 (skew-symmetric exterior product)**.** For all vectors $\mathbf{u}, \mathbf{v}$ in $\mathbb{R}^n$

$$\mathbf{u} \wedge \mathbf{v} = \frac{\mathbf{uv} - \mathbf{vu}}{2}. \tag{F.13}$$

Definition F.11 (biorthogonal basis)**.** Let $\{\mathbf{b}_i\}$ be a basis of $\mathbb{R}^n$. There is a unique reciprocal basis $\left\{\tilde{\mathbf{b}}^i\right\}$ for $\mathbb{R}^n$ satisfying the biorthogonality conditon

$$\mathbf{b}_i \cdot \tilde{\mathbf{b}}^j = \delta_{ij}. \tag{F.14}$$

Definition F.12 (Unit pseudoscalar)**.** The unit pseudo-scalar of $\mathbb{G}^n$ is

$$\boldsymbol{I} := \mathbf{e}_1\mathbf{e}_2 \dots \mathbf{e}_n. \tag{F.15}$$

Definition F.13 (Unit pseudoscalar inverse)**.** The right inverse of the unit pseudoscalar of $\mathbb{G}^n$ is given by its retrograde symmetry

$$\boldsymbol{I}^{-1} = \mathbf{e}_n \dots \mathbf{e}_1. \tag{F.16}$$

Proof. $\boldsymbol{I}\,\boldsymbol{I}^{-1} = \mathbf{e}_1 \dots \mathbf{e}_n\mathbf{e}_n \dots \mathbf{e}_1 = \mathbf{e}_1 \dots \mathbf{e}_{n-1}\mathbf{e}_{n-1} \dots \mathbf{e}_1 = \dots = 1.$ □

Definition F.14 (Dual)**.** The dual of a multivector $M \in \mathbb{G}^n$ is

$$M^* := M\boldsymbol{I}^{-1}. \tag{F.17}$$

Theorem F.7 (Orthogonal complement)**.** *If a blade* $\mathbf{B}$ *represents a subpsace* S*, then* $\mathbf{B}^*$ *represents* $S^\perp$ *the orthogonal complement of* S

Remark F.1.

$$(M \wedge N)^* = M \cdot N^* \qquad (M \cdot N)^* = M \wedge N^* \qquad (\mathbf{u} \wedge \mathbf{v})^* = \mathbf{u} \times \mathbf{v} \tag{F.18}$$

Definition F.15 (Canonical Basis of $\mathbb{G}^2$)**.** Denote $\{\mathbf{e}_1, \mathbf{e}_2\}$ the canonical basis of $\mathbb{R}^2$. The canonical basis of $\mathbb{G}^2$ is

$$\begin{array}{ccc} \mathbf{1} & & \text{(scalar: grade 0)} \\ \mathbf{e}_1 \quad \mathbf{e}_2 & & \text{(vector: grade 1)} \\ \mathbf{e}_1\mathbf{e}_2 & & \text{(pseudoscalar, bivector: grade 2)} \end{array}$$

Definition F.16 (Canonical Basis of $\mathbb{G}^3$)**.** Denote $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ the canonical basis of $\mathbb{R}^2$. The canonical basis of $\mathbb{G}^2$ is

$$\begin{array}{ccc} \mathbf{1} & & \text{(scalar: grade 0)} \\ \mathbf{e}_1 \quad \mathbf{e}_2 \quad \mathbf{e}_3 & & \text{(vector: grade 1)} \\ \mathbf{e}_1\mathbf{e}_2 \quad \mathbf{e}_2\mathbf{e}_3 \quad \mathbf{e}_3\mathbf{e}_1 & & \text{(bivector: grade 2)} \\ \mathbf{e}_1\mathbf{e}_2\mathbf{e}_3 & & \text{(pseudoscalar, trivector: grade 3)} \end{array}$$

Remark F.2. A multivector $M \in \mathbb{G}^n$ separates uniquely into k-vector parts $\langle M \rangle_k$

$$M = \sum_{k=0}^{n} \langle M \rangle_k \tag{F.19}$$

Definition F.17 (Blade)**.** A *k-blade* $\mathbf{B}$ is a product of k nonzero orthogonal vectors

$$\mathbf{B} = \mathbf{b}_1 \mathbf{b}_2 \dots \mathbf{b}_k \tag{F.20}$$

Remark F.3. A k-blade $\mathbf{B} = \mathbf{b}_1 \mathbf{b}_2 \dots \mathbf{b}_k$ represents the subpsace of $\mathbb{R}^n$ with basis $\{\mathbf{b}_1, \dots, \mathbf{b}_n\}$

Definition F.18 (Norm of a blade)**.** The *norm* of a k-blade $\mathbf{B} = \mathbf{b}_1 \dots \mathbf{b}_k$ is

$$|\mathbf{B}| = |\mathbf{b}_1| \dots |\mathbf{b}_k|. \tag{F.21}$$

This is the volume of the parallelogram with edges $\mathbf{b}_1 \dots \mathbf{b}_k$.

Remark F.4. Algebraic operations on blades represent geometric operations on their subspaces

Definition F.19 (inner product (bis))**.** The *inner product* of a j vector A and k-vector B is

$$A \cdot B = \langle AB \rangle_{k-j} \,. \tag{F.22}$$

Definition F.20 (outer product (bis))**.** The *outer product* of a j vector A and k-vector B is

$$A \wedge B = \langle AB \rangle_{k+j} \,. \tag{F.23}$$

F.1.3 Generalized complex numbers

Definition F.21. Let $\{\mathbf{e}_1, \mathbf{e}_2\}$ be an orthonormal basis for a plane in $\mathbb{G}^n$. Then the unit bivector $\mathbf{i} = \mathbf{e}_1 \wedge \mathbf{e}_2 = \mathbf{e}_1 \mathbf{e}_2$ is the unit pseudoscalar of the oriented plane $\mathbf{e}_1 \wedge \mathbf{e}_2$.

Definition F.22 (bivector angle $\mathbf{i}\theta$)**.** Consider an angle θ in a plane $\mathbf{i} \in \mathbb{G}^n$. We call the bivector $\mathbf{i}\theta$ an *angle*. A bivector angle represents both the plane $\mathbf{i}$ and its size $|\mathbf{i}\theta| = \theta$.

Definition F.23 (exponential $e^{\mathbf{i}\theta}$)**.** Define the exponential

$$\exp(\mathbf{i}\theta) = \cos\theta + \mathbf{i}\sin\theta \tag{F.24}$$

Theorem F.8. *Let* $\mathbf{u}, \mathbf{v}$ *be vectors in* $\mathbb{G}^n$*. Let* $\mathbf{i}\theta$ *be the angle from* $\mathbf{u}$ *to* $\mathbf{v}$*. Set* $r = |\mathbf{u}||\mathbf{v}|$*,* $a = r\cos\theta$*,* $b = r\sin\theta$*, then*

$$\mathbf{uv} = r\mathbf{e}^{\mathbf{i}\theta} = a + \mathbf{i}b = \mathbf{u}\cdot\mathbf{v} + \mathbf{u}\wedge\mathbf{v}. \tag{F.25}$$

Theorem F.9 (complex conjugate)**.** *Let* $z = \mathbf{uv} = re^{\mathbf{i}\theta} = a + \mathbf{i}b$ *be a generalized complex number, then*

$$\bar{z} = \mathbf{vu} = re^{-\mathbf{i}\theta} = a - \mathbf{i}b = \mathbf{u}\cdot\mathbf{v} - \mathbf{u}\wedge\mathbf{v} \tag{F.26}$$

is called the complex conjugate of z*.*

Theorem F.10 (complex norm)**.** *Let* $z = \mathbf{uv} = re^{\mathbf{i}\theta} = a + \mathbf{i}b$ *be a generalized complex number, then*

$$|z| = |\mathbf{u}||\mathbf{v}| = r = \sqrt{a^2 + b^2} \tag{F.27}$$

is called the norm of z*.*

Theorem F.11 (complex inverse)**.** *Let* z *be a generalized complex number, then* $|z|^2 = z\bar{z}$*. Thus if* $z \neq 0$*, then* z^{-1} *exists and*

$$z^{-1} = \frac{\bar{z}}{|z|^2} \tag{F.28}$$

F.2 Calculus

Definition F.24 (Gradient)**.** The gradient is the vector operator

$$\boldsymbol{\nabla} := \sum_{i=1}^{n} \mathbf{e}_i \frac{\partial}{\partial x_i} . \tag{F.29}$$

Definition F.25 (Divergence)**.** For $F \in \mathbb{G}^n$

$$\operatorname{div} F := \boldsymbol{\nabla} \cdot F = \sum_{i=1}^{n} \mathbf{e}_i \cdot \frac{\partial}{\partial x_i} \tag{F.30}$$

Definition F.26 (Curl)**.** For $F \in \mathbb{G}^n$

$$\operatorname{curl} F := \boldsymbol{\nabla} \wedge F = \sum_{i=1}^{n} \mathbf{e}_i \wedge \frac{\partial}{\partial x_i} \tag{F.31}$$

Remark F.5.

$$\boldsymbol{\nabla} F = \boldsymbol{\nabla} \cdot F + \boldsymbol{\nabla} \wedge F = \operatorname{div} F + \operatorname{curl} F \tag{F.32}$$

Theorem F.12 (Tangent space basis)**.** *Let* $\boldsymbol{x} : \mathbf{q} = (u, v) \in \mathbb{R}^2 \to \boldsymbol{x}(\mathbf{q}) \in \mathcal{X} \subset \mathbb{R}^n$ *parametrize a surface, then*

$$\left\{ \boldsymbol{x}_u = \frac{\partial \boldsymbol{x}}{\partial u}(\mathbf{q}), \boldsymbol{x}_v = \frac{\partial \boldsymbol{x}}{\partial v}(\mathbf{q}) \right\} \tag{F.33}$$

is a basis of the tangent space $\mathsf{T}_{\mathbf{q}}\mathcal{X}$.

The vector derivative $\boldsymbol{\partial}$ on manifolds, generalizes the gradient on $\mathbb{R}^n$

Definition F.27 (Vector derivative $\boldsymbol{\partial}$)**.** Let $\boldsymbol{x} : (u, v) \in \mathbb{R}^2 \to \boldsymbol{x}(\mathbf{q}) \in \mathcal{X} \subset \mathbb{R}^n$ parametrize a surface with basis $\{\boldsymbol{x}_u, \boldsymbol{x}_v\}$ and reciprocal $\{\boldsymbol{x}^u, \boldsymbol{x}^v\}$. $F(\mathbf{x})$ a multi-vector valued function on $\mathcal{X}$. The vector derivative is

$$\boldsymbol{\partial} F := \boldsymbol{x}^u \frac{\partial F}{\partial u} + \boldsymbol{x}^v \frac{\partial F}{\partial v} . \tag{F.34}$$

Remark F.6. $\mathbf{x}^u, \mathbf{x}^v$ is not necessarily orthogonal, and is the *reciprocal basis* of $\mathbf{x}_u, \mathbf{x}_v$.

Definition F.28 (Line integral)**.** Let C be a curve in $\mathbb{R}^n$, $\boldsymbol{f} : C \to \mathbb{R}^n$ a vector valued function and $\mathrm{d}\mathbf{s}$ the infinitesimal vector tangent to C, the line integral is given by

$$I = \int_C \boldsymbol{f} \cdot \mathrm{d}\mathbf{s}. \tag{F.35}$$

Definition F.29 (Directed integral)**.** Let C be a curve in $\mathbb{R}^n$, $F : C \to \mathbb{G}^n$ a multi-vector valued function and $\mathrm{d}\mathbf{s}$ the infinitesimal vector tangent to C (i.e. the infinitesimal pseudoscalar in the tangent algebra $\mathbb{G}^1$ to C, or 1-form), the directed integral is given by

$$I = \int_C \mathrm{d}\mathbf{s} F = \int_C \mathrm{d}\mathbf{s} \cdot F + \int_C \mathrm{d}\mathbf{s} \wedge F. \tag{F.36}$$

Definition F.30 (Flux integral Vector calculus)**.** Let S be a surface in $\mathbb{R}^3$ (and only in $\mathbb{R}^3$), $\boldsymbol{f} : S \to \mathbb{R}^n$ a vector valued function and $\mathrm{d}\sigma$ the infinitesimal vector *normal* to S, the flux integral is given by

$$I = \iint_S \boldsymbol{f} \cdot \mathrm{d}\sigma. \tag{F.37}$$

Definition F.31 (Directed surface integral)**.** Let S be a surface in $\mathbb{R}^n$, $F : S \to \mathbb{R}^n$ a multivector valued function and $\mathrm{d}\mathbf{S}$ the infinitesimal bivector tangent to S (i.e the infinitesimal pseudo scalar in the tangent algebra $\mathbb{G}^2$ to S, or a 2-form), the flux integral is given by

$$I = \iint_S \mathrm{d}\mathbf{S} F. \tag{F.38}$$

Remark F.7. To generalize notations for the directed integral, let M be a k-dimensional manifold in $\mathbb{R}^n$, let $\mathrm{d}^k\mathbf{x}$ be the infinitesimal pseudoscalar of the algebra $\mathbb{G}^k$ tangent to M (a k-form), then the directed integral on M is noted

$$I = \int_M \mathrm{d}^k\mathbf{x}\, F. \tag{F.39}$$

Theorem F.13. *The boundary of an m-dimensional manifold M is an $m-1$ dimensional manifold, denoted ∂M.*

Remark F.8. If M is a ball (3D), ∂M is a sphere (2D), If M is an hemisphere (2D), ∂M is a circle (1D), if M is a curve (1D), ∂M is its boundary points $0D$.

Remark F.9. Recall the following vector calculus

$$\int_C \nabla H \cdot \mathrm{d}\mathbf{s} = H(\mathbf{x}_1) - H(\mathbf{x}_0) \tag{F.40}$$

$$\iint_S (\nabla \times \mathbf{f}) \cdot \mathrm{d}\sigma = \int_C \mathbf{f} \cdot \mathrm{d}\mathbf{s} \tag{F.41}$$

$$\iiint_V \nabla \cdot \mathbf{f}\, \mathrm{d}V = \iint_S \mathbf{f} \cdot \mathrm{d}\sigma \tag{F.42}$$

Theorem F.14 (Fundamental theorem of geometric calculus)**.** *If M is a m-dimensional manifold in $\mathbb{R}^n$, let $\mathrm{d}^m\mathbf{x}$ be the infinitesimal pseudoscalar of the algebra $\mathbb{G}^m$ tangent to M, then*

$$\int_M \mathrm{d}^m\mathbf{x}\boldsymbol{\partial}\, F = \int_{\partial M} \mathrm{d}^{m-1}\mathbf{x}\, F. \tag{F.43}$$

Theorem F.15 (Divergence theorem in $\mathbb{R}^3$)**.**

$$\iiint_V \nabla\cdot\mathbf{f}\,\mathrm{d}V = \oiint_S \mathbf{f}\cdot\mathrm{d}\sigma \tag{F.44}$$

Theorem F.16 (Generalized Divergence theorem)**.** *If M is a m-dimensional manifold in $\mathbb{R}^m$, and $\mathbf{n}$ is the unit normal to ∂M then,*

$$\int_M \boldsymbol{\nabla} F\,\mathrm{d}^m x = \int_{\partial M} \mathbf{n} F\,\mathrm{d}^{m-1}x \tag{F.45}$$

Remark F.10. These are not directed integrals, but standard multiple integrals. $\mathrm{d}^m x = \mathrm{d}x_1\,\mathrm{d}x_2\ldots\mathrm{d}x_m$.

Theorem F.17 (Curl theorem in $\mathbb{R}^3$)**.**

$$\iint_S (\boldsymbol{\nabla}\times\mathbf{f})\cdot\mathrm{d}\sigma = \oint_C \mathbf{f}\cdot\mathrm{d}\mathbf{s} \tag{F.46}$$

Theorem F.18 (Generalized Curl theorem)**.** *If M is a m-dimensional manifold in $\mathbb{R}^m$, and F is an $m-1$ vector field on M, then*

$$\int_M \mathrm{d}^m\mathbf{x}\cdot(\boldsymbol{\partial}\wedge F) = \int_{\partial M} \mathrm{d}^{m-1}\mathbf{x}\cdot F. \tag{F.47}$$

F.3 Maxwell equations (in empty space)

The following example is taken from [Mac17, eq. (3.1)]. Denote $\boldsymbol{b} \in \mathbb{R}^3$ the 3-dimensional magnetic field (dependence on time and space variables (t, x, y, z) is ommited). In geometric algebra it is represented through its dual: the *bivector* $\boldsymbol{B} \in \operatorname{span}\{\boldsymbol{B}_1 = \mathbf{e}_2\mathbf{e}_3, \boldsymbol{B}_2 = \mathbf{e}_3\mathbf{e}_1, \boldsymbol{B}_3 = \mathbf{e}_1\mathbf{e}_2\} \subset \mathbb{G}^3$. Denote $\boldsymbol{e} \in \mathbb{R}^3$ the electric field. The classical formulation of Maxwell equations using vector calculus is given by the four equations

$$\text{Gauss law} \qquad \operatorname{div} \boldsymbol{e} = \nabla \cdot \boldsymbol{e} = 0, \tag{F.48a}$$

$$\text{Gauss law for magnetism} \qquad \operatorname{div} \boldsymbol{b} = \nabla \cdot \boldsymbol{b} = 0, \tag{F.48b}$$

$$\text{Faraday's law of induction} \qquad \partial_t \boldsymbol{e} - \nabla \times \boldsymbol{b} = 0, \tag{F.48c}$$

$$\text{Ampere law} \qquad \partial_t \boldsymbol{b} + \nabla \times \boldsymbol{e} = 0 \tag{F.48d}$$

Using Geometric calculus, the exterior product $\wedge$ and bivectors are favoured over the cross product $\times$ (which doesn't generalize to n dimensions). Maxwell equations can be rewritten in term of the bivector $\boldsymbol{B}$ and the exterior product $\wedge$ as

$$\boldsymbol{\nabla} \cdot \boldsymbol{e} = 0, \qquad \boldsymbol{\nabla} \wedge \boldsymbol{B} = 0, \qquad \partial_t \boldsymbol{e} + \boldsymbol{\nabla} \cdot \boldsymbol{B} = 0, \qquad \partial_t \boldsymbol{B} + \boldsymbol{\nabla} \wedge \boldsymbol{e} = 0.$$

Using geometric algebra, it becomes possible to introduce the *multivector field* $F = \boldsymbol{e} + \boldsymbol{B}$ (the direct sum of a vector and a bivector) so that Maxwell equations becomes a single equation

$$(\partial_t + \boldsymbol{\nabla})F = 0. \tag{F.50}$$

Finally, multiplying on the left by $(\partial_t - \boldsymbol{\nabla})$ and expanding the differential operator reveals that Maxwell equations are simply an instance of the wave equation *but over a multi vector field* F

$$\partial_t^2 F - \boldsymbol{\nabla}^2 F = 0. \tag{F.51}$$

This is a significant reduction in complexity and a revelator of hidden structure.

> **Remark F.11** (Going further)**.** See [DGL+03, p.229] for a more detailed treatment of Maxwell equations using GA. See also [VLM12] for a port-Hamiltonian approach to Maxwell equations using k-forms applied to plasma dynamics in Tokamak reactors.

Appendix G
Articles

TRAJECTORY ANTI-ALIASING ON GUARANTEED-PASSIVE SIMULATION OF NONLINEAR PHYSICAL SYSTEMS

Rémy Muller, Thomas Hélie *

S3AM team, IRCAM - CNRS UMR 9912- UPMC
1 place Igor Stravinsky, 75004 Paris, France
remy.muller@ircam.fr

ABSTRACT

This article is concerned with the accurate simulation of passive nonlinear dynamical systems with a particular attention paid on aliasing reduction in the pass-band. The approach is based on the combination of Port-Hamiltonian Systems, continuous-time state-space trajectories reconstruction and exact continuous-time anti-aliasing filter realization. The proposed framework is applied on a nonlinear LC oscillator circuit to study the effectiveness of the method.

1. INTRODUCTION

The need for accurate and passive-guaranteed simulation of non-linear multi-physical systems is ubiquitous in the modelling of electronic circuits or mechanical systems.

Geometric numerical integration [1] is a very active research field that provides a theoretical framework for structure and invariant preserving integration of dynamical systems. Port-Hamiltonian Systems (PHS) [2] [3] that focus on the energy storage functions and power continuous component interconnections belong to this field and offer a well adapted framework to preserve the system energy (resp. passivity). In the context of nonlinear physical audio systems, it has been applied successfully to the modelling of the wah-wah pedal [4], Fender Rhodes [5], brass instruments [6] and the loudspeaker nonlinearities [7]. Automatic generation of the system equations from a graph of components has been investigated in [8]

However the presence of aliasing errors in the numerical simulation is annoying for three reasons. First it causes audible inharmonic audio artefacts. Second it deteriorates the accuracy of the numerical scheme leading to poor convergence rate. Third it requires the use of significant oversampling. This problem is even more pronounced in the case of systems such as sustained instruments that rely on nonlinearities to achieve auto-oscillation.

Aliasing errors in the context of finite elements simulation and some alternatives have been discussed in [9] (ch 11). Anti-aliased waveform generation without oversampling has been proposed in [10]. Static nonlinearity anti-aliasing has also been proposed in [11] [12] by combining exact anti-derivatives and finite-differences.

Continuous-time input reconstruction has been used in [13] to simulate the frequency response of LTI systems with higher accuracy. It is also central in collocation-based Runge-Kutta methods that rely on non-uniform polynomial interpolation of the vector field. Splines and in particular uniform B-splines [14] [15] [16], [17] also offer a particularly interesting framework to represent and manipulate piecewise continuous-time signals through their digital representations using the standard tools of linear algebra and digital signal processing.

In this article, we try to combine the geometric and the signal processing viewpoints: we choose a physically informed piecewise smooth polynomial reconstruction model based on a discrete sequence of points generated by a passive-guaranteed simulation method.

The paper is organized as follows. We first recall some results about Port-Hamiltonian systems in Section 3, then we consider passive numerical methods in section 4, we talk about piecewise-continuous trajectory reconstruction in section 5 and continuous-time filtering of piecewise polynomials in section 6. Finally we apply our method to a non linear LC oscillator circuit in section 7.

2. PROBLEM STATEMENT

2.1. Objective

The objective is to simulate nonlinear passive physical audio systems in such a way that:

(i) The nonlinear dynamics is accurately reproduced,

(ii) The power balance decomposed into its conservative, dissipative and source parts is satisfied,

(iii) The observation operator is designed to reduce the aliasing induced by the nonlinearities.

2.2. Approach

To address this problem, the following strategy is adopted.

First, trajectories are approximated in the continuous-time domain by smooth parametric piecewise-defined functions, such that the three following properties are fulfilled:

(P1) Regularity: functions and junctions are $\mathcal{C}^k$ with $k \in \mathbb{N}$,

(P2) Accuracy: the approximation has accuracy order p,

(P3) Passivity: the power balance is globally satisfied for each frame.

Second, the anti-aliased output is built *a posteriori* in three steps:

1. Observe the output from the approximated dynamics in the continuous-time domain,
2. Apply a continuous-time anti-aliasing filter in order to respect the Shannon-Nyquist sampling theorem,
3. Sample the filtered trajectories to convert them back to discrete-time.

* The contribution of this author has been done at laboratory STMS, Paris, within the context of the French National Research Agency sponsored project INFIDHEM. Further information is available at http://www.lagep.cpe.fr/wwwlagep7/anr-dfg-infidhem-fev-2017-jan-2020/

2.3. Methodology

In this article, we restrict ourselves to piece-wise continuous globally $\mathcal{C}^1$ *polynomial* trajectories of the form

$$\hat{\mathbf{x}}(t) = \sum_{n=-\infty}^{\infty} \hat{\mathbf{x}}_n \left(\frac{t-t_n}{h}\right) \mathrm{rect}_{]0,1]}\left(\frac{t-t_n}{h}\right), \quad t \in \mathbb{R} \quad (1)$$

with $\hat{\mathbf{x}} \in \mathbb{R}^N$, $\hat{\mathbf{x}}_n(\tau)$, $\tau \in [0,1]$ being a local polynomial model of order r, $t_n = hn$, $n \in \mathbb{Z}$ and h being the time step parameter. The continuity hypothesis (P1) is expressed mathematically by.

$$\hat{\mathbf{x}}_{n+1}^{(\ell)}(\tau) = \hat{\mathbf{x}}_n^{(\ell)}(\tau) \quad \forall n \in \mathbb{Z}, \ell \leq k \quad (2)$$

For property (P2) the local approximation error between the exact solution and its approximation is defined by

$$e(h) = \mathbf{x}(t_0+h) - \hat{\mathbf{x}}(t_0+h) \quad (3)$$

provided that $\mathbf{x}(t_0) = \hat{\mathbf{x}}(t_0)$ and it is required that for some p.

$$e(h) = \mathcal{O}(h^{p+1}) \quad (4)$$

Finally to express property (P3) we require the power-balance

$$E'(t) = -\mathcal{P}_d + \mathcal{P}_e \quad (5)$$

where $\mathcal{P}_d$ and $\mathcal{P}_e$ are respectively the dissipated and external power and $E'(t)$ is the instantaneous energy variation of the system.

3. PORT-HAMILTONIAN SYSTEMS

In this article, nonlinear passive physical audio systems are described under their Port-Hamiltonian formulation. The theory of Port-Hamiltonian Systems (PHS) [2] [3] extends the theory of Hamiltonian mechanics to non-autonomous and dissipative open systems. It provides a general framework where the dynamic state-space equations derives directly from an energy storage function and *power-conserving* interconnection of its subsystems.

3.1. Explicit differential form

Consider a system with input $\mathbf{u}(t) \in \mathbb{U} = \mathbb{R}^P$, with state $\mathbf{x}(t) \in \mathbb{X} = \mathbb{R}^N$ and output $\mathbf{y}(t) \in \mathbb{Y} = \mathbb{R}^P$ with the structured state-space equations [2]

$$\begin{cases} \mathbf{x}' &= (\mathbf{J}(\mathbf{x}) - \mathbf{R}(\mathbf{x}))\nabla\mathcal{H}(\mathbf{x}) + \mathbf{G}(\mathbf{x})\mathbf{u} = f(\mathbf{x},\mathbf{u}) \\ \mathbf{y} &= \mathbf{G}(\mathbf{x})^T\mathbf{u} \end{cases} \quad (6)$$

where $\mathcal{H}$ gives the stored energy of the system

$$E(t) = (\mathcal{H} \circ \mathbf{x})(t) \quad (7)$$

with $\mathcal{H} \in \mathcal{C}^1(\mathbb{X}, \mathbb{R}^+)$, ∇ being the gradient operator, $\mathbf{J} = -\mathbf{J}^T$ a skew-symmetric matrix and $\mathbf{R} = \mathbf{R}^T \succeq 0$ a positive-semidefinite matrix. The energy variation of this system satisfies the power-balance given by the derivative chain rule

$$E'(t) = \nabla\mathcal{H}(\mathbf{x})^T\mathbf{x}' \quad (8)$$

which can be decomposed as

$$E'(t) = \mathcal{P}_c - \mathcal{P}_d + \mathcal{P}_e \quad (9)$$

with.

$$\mathcal{P}_c = \nabla\mathcal{H}(\mathbf{x})^T\mathbf{J}(\mathbf{x})\nabla\mathcal{H}(\mathbf{x}) = 0 \quad (10)$$

$$\mathcal{P}_d = \nabla\mathcal{H}(\mathbf{x})^T\mathbf{R}(\mathbf{x})\nabla\mathcal{H}(\mathbf{x}) \geq 0 \quad (11)$$

$$\mathcal{P}_e = \nabla\mathcal{H}(\mathbf{x})^T\mathbf{G}(\mathbf{x})\mathbf{u} \quad (12)$$

The $\mathcal{P}_c$ term is null because $\mathbf{J}$ is skew-symmetric: it represents conservative power exchange between storage components in the system. The $\mathcal{P}_d$ term is positive because $\mathbf{R} \geq 0$: it represents the dissipated power. Finally the term $\mathcal{P}_e$ represents the power brought to the system by the external ports.

Equation (9) express the system's *passivity property*: with external inputs switched off ($\mathbf{u} = 0$) the energy can either be constant (conservative case $\mathcal{P}_d = 0$) or decaying (dissipative case $\mathcal{P}_d > 0$).

3.2. Component-based approach and semi-explicit DAE form

More generally, PHS can be expressed in Differential Algebraic Equation form. When we consider physical systems containing N energy-storage components, M dissipative components and P external interaction ports described by

- $\mathcal{P}_c$ the stored energy level e_n and its variation law defined by $e_n' = \nabla\mathcal{H}_n(x_n)x_n'$ for the state variable x_n.
- $\mathcal{P}_d$ the dissipated power $q_m(w) \geq 0$ with the component's flux and effort variables being in algebraic relation of a single variable w.
- $\mathcal{P}_e$ the external power $u_p y_p$ brought to the system through this port with u_p being the controllable input of the system and y_p being the observable output.

For a storage component, $e_n = \mathcal{H}_n(x_n)$ gives the physical *energy storage law*. If x_n' is a flux (resp. effort) variable then $\nabla\mathcal{H}_n(x_n)$ is the dual effort (resp. flux) variable.

Similarly, for a dissipative component, the power is $q_m = R_m(w_m)$ so that if w_m is a flux (resp. effort) variable then $z(w_m) = \frac{R_m(w_m)}{w_m}$ is the effort (resp. flux) and gives the *dissipation law*.

We then consider a passive system obtained by interconnection of these components given by

$$\underbrace{\begin{bmatrix} \mathbf{x}' \\ \mathbf{w} \\ -\mathbf{y} \end{bmatrix}}_{\mathbf{b}} = \mathbf{S}(\mathbf{x},\mathbf{w}) \underbrace{\begin{bmatrix} \nabla\mathcal{H}(\mathbf{x}) \\ z(\mathbf{w}) \\ \mathbf{u} \end{bmatrix}}_{\mathbf{a}} \quad (13)$$

with $\mathbf{S} = -\mathbf{S}^T$ being skew-symmetric, $\mathcal{H}(\mathbf{x}) = \sum_{i=1}^N \mathcal{H}_n(x_n)$ and $z(\mathbf{w}) = [z_1(w_1), \ldots, z_m(w_m)]^T$.

The $\mathbf{S}$ matrix represents the power exchange between components: since $\mathbf{S} = -\mathbf{S}^T$ we have $\mathbf{a} \cdot \mathbf{b} = \mathbf{a}^T\mathbf{S}\mathbf{a} = 0$ which again leads to the power balance[1].

$$\underbrace{\nabla\mathcal{H}(\mathbf{x}) \cdot \mathbf{x}'}_{\mathcal{P}_c = E'(t)} + \underbrace{z(\mathbf{w}) \cdot \mathbf{w}}_{\mathcal{P}_d} - \underbrace{\mathbf{u} \cdot \mathbf{y}}_{\mathcal{P}_e} = 0 \quad (14)$$

The explicit form (6) can be found by solving the second row of (13). The $\mathbf{S}$ matrix represents a *Dirac structure* [2] that expresses the power-balance and can be constructed from a component connection graph [8] [18].

[1] The minus sign in $-y$ in Eq. (13) is used to restore the receiver convention used for internal components.

4. PASSIVE NUMERICAL INTEGRATION

Whereas most numerical schemes concentrate their efforts on the the temporal derivative or the numerical integration quadrature, discrete gradient methods preserve the energy (resp. passivity) given by the power-balance (9), (14) in discrete-time by providing a discrete equivalent of the chain rule derivation property $E'(t) = \nabla\mathcal{H}(\mathbf{x})^T\mathbf{x}'$. A discrete gradient [19] $\overline{\nabla}H$ is required to satisfy the following conditions.

$$\mathcal{H}(\mathbf{x}+\delta\mathbf{x}) - \mathcal{H}(\mathbf{x}) = \overline{\nabla}\mathcal{H}(\mathbf{x},\delta\mathbf{x})^T\delta\mathbf{x} \tag{15}$$

$$\overline{\nabla}\mathcal{H}(\mathbf{x},0) = \nabla\mathcal{H}(\mathbf{x}) \tag{16}$$

In this article, we will focus on the average vector field [20].

4.1. Average Vector Field

In the general case, the AVF method is defined by.

$$\frac{\delta\mathbf{x}_n}{\delta t} = \int_0^1 f(\mathbf{x}_n + \tau\delta\mathbf{x}_n)d\tau, \quad \mathbf{x}_{n+1} = \mathbf{x}_n + \delta\mathbf{x}_n \tag{17}$$

When the matrices $\mathbf{J}(\mathbf{x}), \mathbf{R}(\mathbf{x}), \mathbf{G}(\mathbf{x})$ are approximated by constant matrices $\bar{\mathbf{J}}, \bar{\mathbf{R}}, \bar{\mathbf{G}}$, we obtain the separable structure-preserving approximation of (17)

$$\frac{\delta\mathbf{x}_n}{\delta t} = (\overline{\mathbf{J}} - \overline{\mathbf{R}})\overline{\nabla}H(\mathbf{x}_n,\delta\mathbf{x}_n) + \overline{\mathbf{G}}\bar{\mathbf{u}}_n \tag{18}$$

with the discrete gradient being defined by

$$\overline{\nabla}\mathcal{H}(\mathbf{x},\delta\mathbf{x}) = \int_0^1 \nabla\mathcal{H}(\mathbf{x}+\tau\delta\mathbf{x})d\tau \tag{19}$$

and it satisfies the *discrete power balance*

$$\begin{aligned}\delta E = \overline{\nabla}\mathcal{H}^T\frac{\delta\mathbf{x}}{\delta t} &= \overline{\nabla}\mathcal{H}^T(\overline{\mathbf{J}} - \overline{\mathbf{R}})\overline{\nabla}\mathcal{H} + \overline{\nabla}\mathcal{H}^T\overline{\mathbf{G}}\overline{\mathbf{u}} \\ &= 0 - \mathcal{P}_d + \mathcal{P}_e\end{aligned}$$

Then, by the fundamental theorem of calculus, for mono-variant components, i.e. separable Hamiltonians of the form $\mathcal{H}(\mathbf{x}) = \sum_{i=1}^N \mathcal{H}_i(x_i)$, we have for each coordinate:

$$\overline{\nabla}\mathcal{H}_i(x_i,\delta x_i) = \begin{cases} \frac{\mathcal{H}_i(x_i+\delta x_i)-\mathcal{H}_i(x_i)}{\delta x_i} & \delta x_i \neq 0 \\ \nabla\mathcal{H}(x_i) & \delta x_i = 0 \end{cases} \tag{20}$$

which satisfies the discrete gradient conditions (15)-(16). For non-separable Hamiltonians, a discrete-gradient can also be uniquely defined, see [21] for more details.

To summarize, this method relies on two complimentary approximations: the differential operator $\frac{d\mathbf{x}}{dt} \to \frac{\delta\mathbf{x}}{\delta t}$ and the vector field $f \to \overline{f}$ to achieve energy (resp. passivity) conservation. The discrete PHS equivalent of (6) is given by the numerical scheme.

$$\begin{cases} \frac{\delta\mathbf{x}_n}{\delta t} &= \left(\overline{\mathbf{J}} - \overline{\mathbf{R}}\right)\overline{\nabla}\mathcal{H}(\mathbf{x}_n,\delta\mathbf{x}_n) + \overline{\mathbf{G}}\overline{\mathbf{u}}_n \\ \mathbf{y}_n &= \overline{\mathbf{G}}^T\overline{\nabla}\mathcal{H}(\mathbf{x}_n,\delta\mathbf{x}_n) \\ \mathbf{x}_{n+1} &= \mathbf{x}_n + \delta\mathbf{x}_n \end{cases} \tag{21}$$

4.2. Accuracy order

As shown in [22], the AVF has accuracy order $p = 2$, it is a B-series method, is affine-covariant and self-adjoint. When approximated as in Eq (19) by evaluating matrices $\mathbf{J}, \mathbf{R}, \mathbf{G}$ for $\mathbf{x}^* = \mathbf{x}_n$ the accuracy is only of order 1. Order 2 is achieved when either $\mathbf{J}, \mathbf{R}, \mathbf{G}$ are independent of $\mathbf{x}$ or when evaluated at the mid-point $\mathbf{x}^* = \mathbf{x}_n + \frac{\delta\mathbf{x}_n}{2}$ in the conservative case. It is also possible to restore the accuracy order $p = 2$ in the general case using a Runge-Kutta refinement [21].

4.3. Implicit resolution

The discrete system is implicit on $\delta\mathbf{x}_n$ and admits a unique solution when $\mathcal{H}$ is convex. In the general case, an iterative solver is required (typically a fixed-point or Newton iteration), but when the Hamiltonian is quadratic we can avoid the need for an iterative resolution. Furthermore, when the Hamiltonian is convex the method can also be made non-iterative by quadratization of the Hamiltonian [21].

Proof. When the Hamiltonian is quadratic of the form $\mathcal{H}(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x}$, the discrete gradient reduces to the mid-point rule

$$\overline{\nabla}\mathcal{H}(\mathbf{x},\delta\mathbf{x}) = \int_0^1 \mathbf{Q}(\mathbf{x}_n + \delta\mathbf{x}_n\tau)d\tau = \mathbf{Q}\left(\mathbf{x}_n + \frac{1}{2}\delta\mathbf{x}_n\right)$$

the implicit dependency on $\delta\mathbf{x}$ can thus be solved by matrix inversion

$$\delta\mathbf{x}_n = \delta t\left(I - \frac{\delta t}{2}\mathbf{A}\right)^{-1}\left(\mathbf{A}\mathbf{x}_n + \overline{\mathbf{G}}\overline{\mathbf{u}}_n\right) \tag{22}$$

with $\mathbf{A} = (\overline{\mathbf{J}} - \overline{\mathbf{R}})\mathbf{Q}$ □

5. PIECEWISE-CONTINUOUS TRAJECTORIES

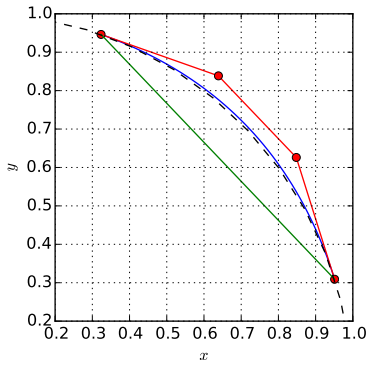


Figure 1: Example of a cubic trajectory with conservative end-points. The affine trajectory used to compute the average vector field is shown (in green), the associated cubic interpolated approximation (in blue), its control polygon (in red), and the exact manifold (in dashed black).

Given the sequence of points $\{\mathbf{x}_n\}$ obtained by a passive-guaranteed method, we would like to reconstruct piece-wise $\mathcal{C}^k$-continuous polynomial trajectories informed by the system dynamics.

The idea is to exploit the dynamic equation at each junction point $\mathbf{x}_n$ where the approximation is known to be $\mathcal{O}(h^{p+1})$.

Indeed, if we had the samples of the exact trajectory, by the Weierstrass approximation theorem, arbitrarily close polynomial approximations converging uniformly to the exact solution could be obtained by computing its derivatives to any desired order.

Since we only have an approximation of order $p = 2$, we restrict ourselves to a regularity $k = 1$. This gives four constraints

$$\hat{\mathbf{x}}(0) = \mathbf{x}_n, \quad \hat{\mathbf{x}}(1) = \mathbf{x}_{n+1}, \quad \hat{\mathbf{x}}'(0) = f(\mathbf{x}_n), \quad \hat{\mathbf{x}}'(1) = f(\mathbf{x}_{n+1})$$

that can be satisfied by a cubic polynomial ($r = 3$). We choose to represent it using the Bézier form,

$$\hat{\mathbf{x}}(\tau) = \sum_{i=0}^{3} \mathbf{X}_i B_i^3(\tau), \quad B_i^n(t) = \binom{n}{i}(1-t)^{n-i}t^i \tag{23}$$

with $\{\mathbf{X}_i\}$ being its control polygon and $B_i^n(t)$ being the Bernstein polynomial basis functions, because they have important geometric and finite differences interpretations [23].

This choice immediately leads to the following equations,

$$\mathbf{X}_0 = \mathbf{x}_n \qquad \mathbf{X}_1 = \mathbf{x}_n + \frac{1}{3}f(\mathbf{x}_n) \tag{24}$$

$$\mathbf{X}_3 = \mathbf{x}_{n+1} \qquad \mathbf{X}_2 = \mathbf{x}_{n+1} - \frac{1}{3}f(\mathbf{x}_{n+1}) \tag{25}$$

where the internal control points $\mathbf{X}_1, \mathbf{X}_2$ are computed from the end points $\mathbf{x}_n, \mathbf{x}_{n+1}$ by first order forward / backward prediction using the derivative rule.

$$\hat{\mathbf{x}}'(t) = \sum_{i=0}^{n-1} \mathbf{D}_i B_i^{n-1}(t), \quad \mathbf{D}_i = n(\mathbf{X}_{i+1} - \mathbf{X}_i) \tag{26}$$

An example trajectory is shown in Figure 1.

6. ANTI-ALIASED OBSERVATION

Given an observed signal $\tilde{\mathbf{u}}(t) = \mathbf{y}(t)$ belonging to the class of piecewise polynomials, in order to reject the non-band-limited part of the spectrum, we would like to apply an antialiasing filter operator given by its continuous-time ARMA transfer function $H(s)$, then sample its output $\tilde{y}(t)$ to get back to the digital domain.

Since our anti-aliasing filter will be LTI, we will make use of *exact exponential integration* and decompose its output on a *custom basis* of exponential polynomial functions.

Without loss of generality we only consider single-input single-output filters (SISO) since we can always filter each observed output independently.

6.1. State-space ARMA filtering of polynomial input

We want to filter the trajectory by an ARMA filter given by its Laplace transfer function

$$H(s) = \frac{Y(s)}{U(s)} = \frac{b_0 s^N + b_1 s^{N-1} + \ldots + b_N}{s^N + a_1 s^{N-1} + \ldots + a_N} \tag{27}$$

This filter can be realized in state-space form as

$$\tilde{\mathbf{x}}' = \mathbf{A}\tilde{\mathbf{x}} + \mathbf{B}\tilde{u} \tag{28}$$

$$\tilde{y} = \mathbf{C}\tilde{\mathbf{x}} + \mathbf{D}\tilde{u} \tag{29}$$

Common choices are the observable and controllable state-space forms.

Furthermore when the denominator can be factored with distinct roots, it is possible to rewrite the transfer function using partial fraction expansion as.

$$H(s) = c_0 + \frac{c_1}{s - \lambda_1} + \ldots + \frac{c_N}{s - \lambda_N} \tag{30}$$

which leads to the canonical diagonal form

$$\mathbf{A} = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_N \end{bmatrix} \qquad \mathbf{B} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \tag{31}$$

$$\mathbf{C} = \begin{bmatrix} c_1 & \ldots & c_N \end{bmatrix} \qquad \mathbf{D} = \begin{bmatrix} c_0 \end{bmatrix} \tag{32}$$

6.2. Exact exponential integration

The exact state trajectory is given by the integral

$$\tilde{\mathbf{x}}(t) = \tilde{\mathbf{x}}_h(t) + \tilde{\mathbf{x}}_e(t) = e^{\mathbf{A}t}\tilde{\mathbf{x}}_0 + \int_0^t e^{\mathbf{A}(t-\tau)}\mathbf{B}\tilde{u}(\tau)d\tau \tag{33}$$

as the sum of the homogeneous solution to the initial conditions $\tilde{\mathbf{x}}_h$ and the forced state-response with zero initial conditions $\tilde{\mathbf{x}}_e$ given by the convolution of the input with the kernel $e^{\mathbf{A}t}$.

Furthermore when $\mathbf{A}$ is diagonal we have

$$e^{\mathbf{A}t} = \begin{bmatrix} e^{\lambda_1 t} & & \\ & \ddots & \\ & & e^{\lambda_N t} \end{bmatrix} \tag{34}$$

which greatly simplifies the computation of the exponential map. In that case (33) can be evaluated component-wise as

$$\tilde{x}^i(t) = e^{\lambda_i t}\tilde{x}_0^i + \int_0^t e^{\lambda_i(t-\tau)}\tilde{u}(\tau)d\tau \quad i \in \{1 \ldots N\} \tag{35}$$

where we used the notation x^i to detonate the i-th coordinate of the vector $\mathbf{x}$

6.2.1. *Polynomial input*

With $\tilde{u}(t)$ being a polynomial of degree K in monomial[2] form and coefficients $\tilde{u}_k$

$$\tilde{u}(t) = \sum_{k=0}^{K} \tilde{u}_k \frac{t^k}{k!} \tag{36}$$

we can expand the forced response $\tilde{x}_e$ in (35) as a weighted sum

$$\int_0^t e^{\lambda_i(t-\tau)} \left(\sum_{k=0}^{K} \tilde{u}_k \frac{t^k(\tau)}{k!} \right) d\tau = \sum_{k=0}^{K} \tilde{u}_k \varphi_{k+1}(\lambda_i, t) \tag{37}$$

with the basis functions $\{\varphi_k\}$ being defined by the convolution

$$\varphi_k(\lambda, t) = \int_0^t e^{\lambda(t-\tau)} \frac{\tau^{k-1}}{(k-1)!} d\tau \qquad k \geq 1 \tag{38}$$

One of the main advantages of using a polynomial input (rather than a more general model) lies in the fact that these basis functions can be integrated exactly, avoiding the need of a quadrature

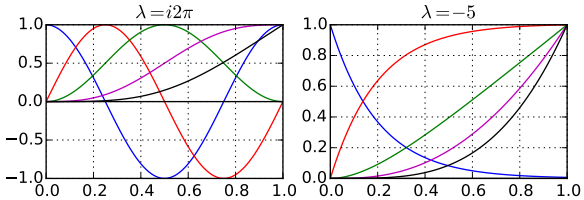


Figure 2: Normalized φ-functions for $k \in \{0 \dots 4\}$. The real parts of the impulse (blue), step (red), ramp (green), quadratic (magenta) and cubic (black) responses are shown for a complex pole $\lambda = i2\pi$ (left plot) and a real pole $\lambda = -5$ (right plot) over the unit interval $t \in [0, 1]$.

approximation formula. See Appendix 12 for a detailed derivation and a recursive formula, and Figure 2 for their temporal shapes.

Using those we can decompose the local state trajectories as.

$$\tilde{x}^i(t) = \tilde{x}^i_0 \varphi_0(\lambda_i, t) + \sum_{k=0}^{K} \tilde{u}_k \varphi_{k+1}(\lambda_i, t) \tag{39}$$

We note that the initial condition is equivalent to an impulsive input $\tilde{x}^i_0 \delta(t)$. This filtering scheme can thus be generalized to non polynomial impulsive inputs.

6.2.2. *Numerical update scheme*

Since we only wish to sample the trajectory on a fixed grid $t_n \in \mathbb{Z}$, we just need to evaluate the local state trajectory $\mathbf{x}(t)$ and the output $y(t)$ at $t = 1$ to finally get the following numerical scheme

$$\tilde{x}^i_{n+1} = \tilde{x}^i_n \varphi_0(\lambda_i) + \sum_{k=0}^{K} \tilde{u}_{k,n} \varphi_{k+1}(\lambda_i, 1) \tag{40}$$

$$\tilde{y}_{n+1} = \sum_{i=1}^{N} c_i \tilde{x}^i_{n+1} + c_0 \tilde{u}_n(1) \tag{41}$$

where the coefficients $\varphi_k(\lambda_i, 1)$ can be pre-computed and the components $\tilde{x}^i_{n+1}$ evaluated in parallel.

6.3. Filter examples

6.3.1. *Low-pass filter of order 1*

We consider a first order low-pass filter with transfer function $H(s) = \frac{a}{s+a}$. The temporal response to a piecewise polynomial input $\{t^2, 1-t, 0, 1\}$ is shown in Figure 3 for $a \in \{1, 3, 6, 10\}$.

6.3.2. *Butterworth Filter of order 3*

To further illustrate the non-band-limited representation capacity of piece-wise polynomials, and the effectiveness of the filtering scheme, we have shown in Figure 4 the response of a third-order Butterworth filter with cutoff $\omega_c = \pi$ to a triangular input signal. Its Laplace transfer function for a normalized pulsation $\omega_c = 1$ is given by $H(s) = \frac{1}{(s^2+s+1)(s+1)}$ with poles $\lambda_1 = \frac{-1-i\sqrt{3}}{2}$, $\lambda_2 = \frac{-1+i\sqrt{3}}{2}$, $\lambda_3 = -1$ and coefficients $c_0 = 0$, $c_1 = \frac{-3+i\sqrt{3}}{6}$, $c_2 = \frac{-3-i\sqrt{3}}{6}$, $c_3 = 1$.

[2] We use the monomial form here instead of Bernstein polynomials because this is the one that leads to the most straightforward and meaningful derivation.

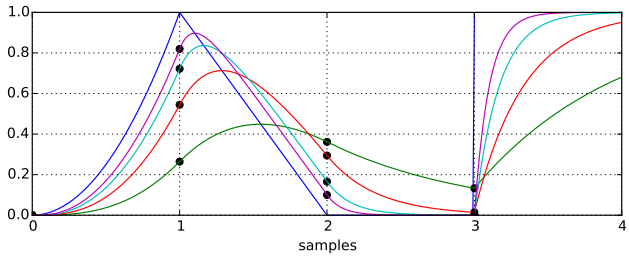


Figure 3: Exact continuous-time responses of a first order low-pass filter to a polynomial input (in blue).

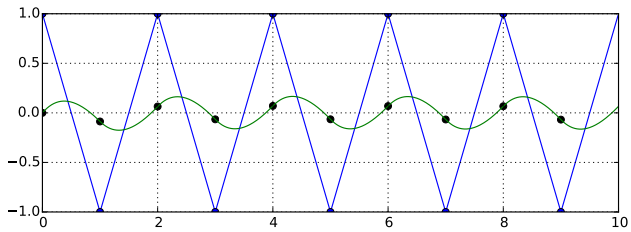


Figure 4: Exact continuous-time response of the order 3 Butterworth filter with cutoff pulsation $\omega_c = \pi$ to a triangle input at the Nyquist frequency.

7. APPLICATION: NONLINEAR LC OSCILLATOR

In order to illustrate the proposed method, we consider the simplest example having non linear dynamics. For that purpose, we use a parallel autonomous LC circuit with a linear inductor and a saturating capacitor with the Hamiltonian energy storage function given by

$$\mathcal{H}(q, \phi) = \frac{\ln(\cosh(q))}{C_0} + \frac{\phi}{2L} \tag{42}$$

where the state q is the charge of the capacitor and ϕ the flux in the inductor. Its circuit's schematic is shown in figure 5 and its energy storage law are displayed in 6

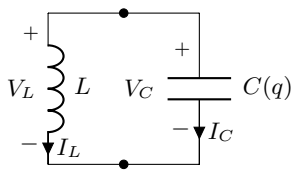


Figure 5: A nonlinear LC oscillator circuit

By partial differentiation of the Hamiltonian function $\mathcal{H}$ by respectively q and ϕ we get the capacitor's voltage and the inductor's current, while applying the temporal derivative on q, ϕ gives the capacitor's current and inductor's voltage.

$$V_C = \partial_q \mathcal{H} = \frac{\tanh(q)}{C_0} \qquad I_C = q' \tag{43}$$

$$I_L = \partial_\phi \mathcal{H} = \frac{\phi}{L} \qquad V_L = \phi' \tag{44}$$

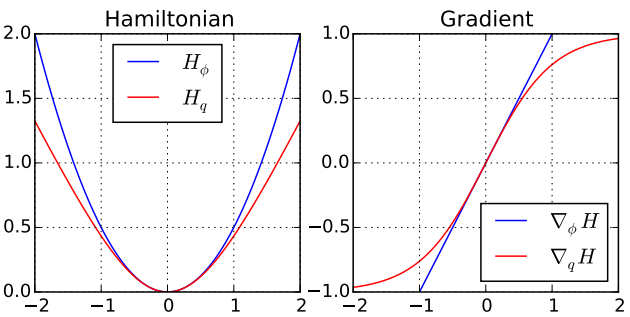


Figure 6: Respective energy storage functions (left plot) and their gradients (right plot), of the nonlinear capacitor (in red) and linear inductor (in blue), for $C = 1$, $L = 1$.

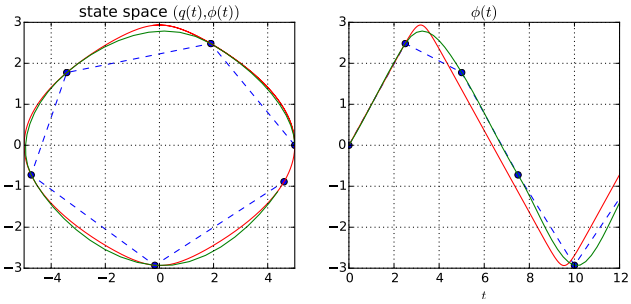


Figure 7: Comparison of simulated orbits with discrete points (in blue) computed using the AVF method, reconstructed cubic trajectory (in green) and reference trajectory computed at 10x sampling rate (in red).

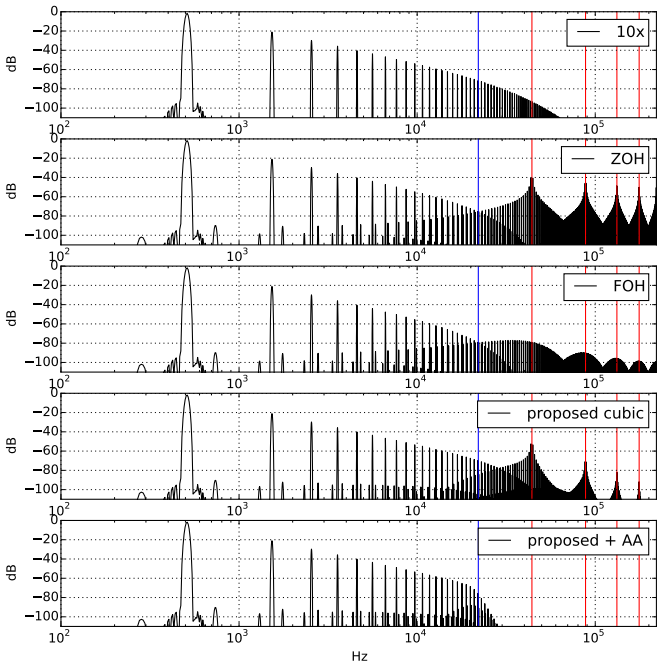


Figure 8: Continuous-time spectrum of the nonlinear LC circuit flux ϕ for a fundamental frequency of 500 Hz and a sampling frequency of 44.1 kHz. The 10x oversampled reference is compared to the AVF method's discrete output with zero-order hold (ZOH), first-order hold (FOH), the proposed method (proposed cubic) and its 12th order Butterworth filtered spectrum (proposed + AA). The Nyquist frequency is materialized in blue and the multiples of the sampling rate in red.

This gives the Branch Component Equations.

Applying Kirchhoff Current and Voltage Laws gives the constraints $I_C = -I_L$, $V_C = V_L$. We can summarize the previous equations with the conservative autonomous Hamiltonian system.

$$\mathbf{x}' = \mathbf{J}\nabla\mathcal{H}(\mathbf{x}) \tag{45}$$

with.

$$\mathbf{x} = \begin{bmatrix} q \\ \phi \end{bmatrix}, \qquad \mathbf{J} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}, \qquad \nabla\mathcal{H} = \begin{bmatrix} \partial_q \mathcal{H} \\ \partial_\phi \mathcal{H} \end{bmatrix} \tag{46}$$

Its state space and temporal trajectories are shown in Figure 7. We can see that the numerical scheme preserves the energy since the discrete points lie exactly on the orbit of the reference trajectory. The reconstructed state-space trajectory also shows a good match with the reference for most of the interpolated segments, except around transition regions at the bottom and top.

The spectrum of the flux ϕ is shown in Figure 8. One can see that the reference spectrum contains harmonics above twice the representable bandwidth where they pass below -90 dB.

The ZOH and FOH spectrums contains spectral images of the non bandlimited spectrum that decay respectively at -6dB/oct and -12dB/oct. Their aliased components in the audio bandwidth start around -80 dB at the Nyquist frequency and decay slowly toward approximately -100 dB at low frequencies.

Contrary, our method, informed by the dynamic, exhibits both reduced aliasing in the audio bandwidth and sharpened spectrum around the Nyquist frequency. It also has a higher spectral images decay rate thanks to its $\mathcal{C}^1$ regularity. Its aliased components start at -85 dB at the Nyquist frequency and decay much faster to reach -100 dB at about 14 kHz where they reach a kind of aliasing noise floor caused by higher harmonics fold-back.

Finally, as expected, the 12th-order Butterworth half-band lowpass filter removes components above the Nyquist frequency thanks to the piecewise continuous cubic input.

8. DISCUSSION

First, we highlight the fact that the vector field approximation in (17) acts as a first-order antialiasing filter: it is a projection of the vector field on a rectangular kernel. It prevents high-order spectral images from disturbing the low frequency dynamic during the numerical simulation and it is consistent with the underlying piecewise linear approximation model.

Second, the numerical scheme is energy-preserving. From a signal processing perspective, the lowpass filtering effect on the vector field is compensated by the finite difference approximation of the derivative. This is a direct generalization of the mid-point / bilinear methods to nonlinear differential equations.

Third, using the fact that the trajectory approximation has accuracy order $p = 2$ at the junctions, we can re-exploit the differential equation to reconstruct an informed $\mathcal{C}^1$-continuous cubic trajectory. It exhibits reduced aliasing in the passband and better

high-frequency resolution.

We observe that on the studied example, our method manages to reduce aliased components that are folded once into the audio band. However components caused by multiple folding of the spectrum cannot be removed anymore. This is related to the Papoulis generalized sampling expansion [24] who states that a band-limited function can be perfectly reconstructed from its values and derivatives sampled at half the Nyquist rate.

Some difficulties arise when trying to generalize the above ideas to higher order trajectories and filtering kernels. First, the line-integral (17) is no longer computable in closed form when the trajectory model is non-affine. Second, higher order kernels have longer temporal support which can lead to non-causal integrals.

9. CONCLUSION AND PERSPECTIVES

Our main contribution is an approach based on smooth piecewise defined trajectories coupled with a guaranteed-passive simulation. The method proceeds in three steps: 1) an energy-preserving passive numerical scheme is applied, 2) $\mathcal{C}^k$-continuous trajectories are reconstructed, 3) Exact continuous time lowpass filtering and sampling is performed. We have proposed a first instance of this method using the class of piecewise polynomials with regularity $k = 1$ and accuracy order $p = 2$ that exhibits reduced aliasing.

Further work will concern increasing the regularity k and accuracy order p, merging the numerical scheme and the interpolation steps by considering energy-preserving methods with a built-in regular continuous model and considering other classes of models such as rational and exponential functions.

In this regard, exponential integrators [25] that integrate the linear part of the dynamic exactly (as we have done in section 6) and rely on approximations for the nonlinear part are of great interest.

Finally we would like to further investigate the link between multi-stages / multi-derivatives general linear methods, their accuracy orders, numerical dispersion and internal bandwidth, and to analyze their behavior and representation capabilities within the framework of Reproducing Kernels Hilbert Spaces and generalized sampling theory [26] [27] [28].

10. ACKNOWLEDGMENTS

This work has been done at laboratory STMS, Paris, within the context of the French National Research Agency sponsored project INFIDHEM. Further information is available on the project web page.

11. REFERENCES

[1] E. Hairer, C. Lubich, and G. Wanner, *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations; 2nd ed.*, Springer, Dordrecht, 2006.

[2] A. van der Schaft and D. Jeltsema, "Port-hamiltonian systems theory: An introductory overview," *Foundations and Trends in Systems and Control*, vol. 1, no. 2-3, pp. 173–378, 2014.

[3] A. van der Schaft, "Port-hamiltonian systems: an introductory survey," in *Proceedings of the International Congress of Mathematicians Vol. III: Invited Lectures*, Madrid, Spain, 2006, pp. 1339–1365.

[4] A. Falaize and T. Hélie, "Simulation of an analog circuit of a wah pedal: a port-Hamiltonian approach," in *135th convention of the Audio Engineering Society*, New-York, United States, Oct. 2013, pp. –.

[5] A. Falaize and T. Hélie, "Passive simulation of the nonlinear port-Hamiltonian modeling of a Rhodes Piano," *Journal of Sound and Vibration*, vol. 390, pp. 289–309, Mar. 2017.

[6] N. Lopes and T. Hélie, "Energy Balanced Model of a Jet Interacting With a Brass Player's Lip," *Acta Acustica united with Acustica*, vol. 102, no. 1, pp. 141–154, 2016.

[7] A. Falaize and T. Hélie, "Passive simulation of electrodynamic loudspeakers for guitar amplifiers: a port- Hamiltonian approach," in *International Symposium on Musical Acoustics*, Le Mans, France, July 2014, pp. 1–5.

[8] A. Falaize and T. Hélie, "Passive guaranteed simulation of analog audio circuits: A port-hamiltonian approach," *Applied Sciences*, vol. 6, no. 10, 2016.

[9] J. P. Boyd, *Chebyshev and Fourier Spectral Methods*, Dover Books on Mathematics. Dover Publications, Mineola, NY, second edition, 2001.

[10] T. S. Stilson, *Efficiently-variable Non-oversampled Algorithms in Virtual-analog Music Synthesis: A Root-locus Perspective*, Ph.D. thesis, 2006.

[11] V. Zavalishin J. D. Parker and E. Le Bivic, "Reducing the aliasing of nonlinear waveshaping using continuous-time convolution," in *Proc. Digital Audio Effects (DAFx-16)*.

[12] S. Bilbao, F. Esqueda J. D. Parker, and V. Valimaki, "Antiderivative antialiasing for memoryless nonlinearities," in *IEEE Signal Processing Letters*, Nov. 2016.

[13] S. Sarkka and A. Huovilainen, "Accurate discretization of analog audio filters with application to parametric equalizer design," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2486–2493, Nov 2011.

[14] M. Unser, "Think analog, act digital," in *Seventh Biennial Conference, 2004 International Conference on Signal Processing and Communications (SPCOM'04)*, Bangalore, India, December 11-14, 2004.

[15] M. Unser, A. Aldroubi, and M. Eden, "B-Spline signal processing: Part I—Theory," *IEEE Transactions on Signal Processing*, vol. 41, no. 2, pp. 821–833, February 1993.

[16] M. Unser, A. Aldroubi, and M. Eden, "B-Spline signal processing: Part II—Efficient design and applications," *IEEE Transactions on Signal Processing*, vol. 41, no. 2, pp. 834–848, February 1993.

[17] M. Unser, "Cardinal exponential splines: Part II—Think analog, act digital," *IEEE Transactions on Signal Processing*, vol. 53, no. 4, pp. 1439–1449, April 2005.

[18] A. Falaize, N. Lopes, T. Hélie, D. Matignon, and B. Maschke, "Energy-balanced models for acoustic and audio systems: a port-Hamiltonian approach," in *Unfold Mechanics for Sounds and Music*, Paris, France, Sept. 2014.

[19] E. L. Mansfield and G. R. W. Quispel, "On the construction of discrete gradients," 2009.

[20] E. Celledoni, V. Grimm, R.I. McLachlan, D.I. McLaren, D. O'Neale, B. Owren, and G.R.W. Quispel, "Preserving

energy resp. dissipation in numerical PDEs using the 'average vector field' method," *Journal of Computational Physics*, vol. 231, no. 20, pp. 6770 – 6789, 2012.

[21] N. Lopes, T. Hélie, and A. Falaize, "Explicit second-order accurate method for the passive guaranteed simulation of port-Hamiltonian systems," in *5th IFAC Work- shop on Lagrangian and Hamiltonian Methods for Non Linear Control*, Lyon, France, July 2015, IFAC.

[22] E. Celledoni, R. I. McLachlan, D. I. McLaren, B. Owren, G. R. W. Quispel, and W. M. Wright, "Energy-preserving runge-kutta methods," *ESAIM: Mathematical Modelling and Numerical Analysis*.

[23] R. T. Farouki, "The bernstein polynomial basis: A centennial retrospective," *Comput. Aided Geom. Des.*, vol. 29, no. 6, pp. 379–419, Aug. 2012.

[24] A. Papoulis, "Generalized sampling expansion," *IEEE Transactions on Circuits and Systems*, vol. 24, no. 11, pp. 652–654, Nov 1977.

[25] M. Hochbruck and A. Ostermann, "Exponential integrators," *Acta Numerica*, vol. 19, pp. 209–286, 2010.

[26] D. Nehab and H. Hoppe, "A fresh look at generalized sampling," *Foundations and Trends in Computer Graphics and Vision*, vol. 8, no. 1, pp. 1–84, 2014.

[27] P.L. Dragotti, M. Vetterli, and T. Blu, "Sampling moments and reconstructing signals of finite rate of innovation: Shannon meets strang-fix," *IEEE Transactions on Signal Processing*, vol. 55, no. 5, pp. 1741–1757, May 2007.

[28] M. Unser, "Sampling-50 years after shannon," *Proceedings of the IEEE*, vol. 88, no. 4, pp. 569–587, April 2000.

12. APPENDIX: φ-FUNCTIONS

The φ-functions, that appear when doing exact integration of an LTI system with polynomial input given in monomial form, are defined by the convolution integral

$$\varphi_k(\lambda, t) = \int_0^t e^{\lambda(t-\tau)} \frac{\tau^{k-1}}{(k-1)!} d\tau \qquad k \geq 1 \tag{47}$$

and by definition

$$\varphi_0(\lambda, t) := e^{\lambda t} \tag{48}$$

For $\lambda = 0$ it is immediate that

$$\varphi_k(\lambda = 0, t) = \frac{t^k}{k!} \tag{49}$$

12.1. Recurrence relation

We first prove that they satisfy the recurrence formula

$$\varphi_{k+1}(\lambda, t) = \frac{\varphi_k(\lambda, t) - \varphi_k(0, t)}{\lambda} \quad \lambda \neq 0 \tag{50}$$

Proof. Using integration by parts

$$\int_a^b u(\tau)v'(\tau)d\tau = [uv]_a^b - \int_a^b u'(\tau)v(\tau)d\tau$$

with $[a, b] = [0, t]$, $u(\tau) = e^{\lambda(t-\tau)}$, $v'(\tau) = \frac{\tau^{k-1}}{(k-1)!}$ and its primitive $v(\tau) = \frac{\tau^k}{k!}$ gives

$$\begin{aligned}\varphi_k(\lambda, t) &= \left[e^{\lambda(t-\tau)} \frac{\tau^k}{k!}\right]_0^t + \lambda \int_0^t e^{\lambda(t-\tau)} \frac{\tau^k}{k!} d\tau \\ &= \frac{t^k}{k!} + \lambda \varphi_{k+1}(\lambda, t)\end{aligned}$$

which after using (49) and identification gives

$$\varphi_{k+1}(\lambda, t) = \frac{\varphi_k(\lambda, t) - \varphi_k(0, t)}{\lambda}$$

□

12.2. Explicit form

Using (50) recursively for $\lambda \neq 0$, the first basis functions are given by

$$\varphi_0(\lambda, t) = e^{\lambda t} \tag{51}$$

$$\varphi_1(\lambda, t) = \frac{e^{\lambda t} - 1}{\lambda} \tag{52}$$

$$\varphi_2(\lambda, t) = \frac{e^{\lambda t} - (1 + \lambda t)}{\lambda^2} \tag{53}$$

$$\varphi_3(\lambda, t) = \frac{e^{\lambda t} - (1 + \lambda t + \frac{(\lambda t)^2}{2!})}{\lambda^3} \tag{54}$$

$$\varphi_4(\lambda, t) = \frac{e^{\lambda t} - (1 + \lambda t + \frac{(\lambda t)^2}{2!} + \frac{(\lambda t)^3}{3!})}{\lambda^4} \tag{55}$$

this suggests the following explicit form

$$\varphi_k(\lambda, t) = \frac{1}{\lambda^k} \left(e^{\lambda t} - \sum_{n=0}^{k-1} \frac{(\lambda t)^n}{n!} \right), \quad \lambda \neq 0 \tag{56}$$

Proof. It is immediate to verify that (56) is satisfied for $k = 0$. Then assuming that (56) is true for some $k \in \mathbb{N}$ and using the recurrence (50) we prove

$$\begin{aligned}\varphi_{k+1}(\lambda, t) &= \frac{\varphi_k(\lambda, t) - \varphi_k(0, t)}{\lambda} \\ &= \frac{1}{\lambda^{k+1}} \left(e^{\lambda t} - \sum_{n=0}^{k-1} \frac{(\lambda t)^n}{n!} \right) - \frac{1}{\lambda} \frac{t^k}{k!} \\ &= \frac{1}{\lambda^{k+1}} \left(e^{\lambda t} - \sum_{n=0}^{k} \frac{(\lambda t)^n}{n!} \right)\end{aligned}$$

that (56) is also true for $k + 1$. By induction (56) is thus satisfied for all $k \in \mathbb{N}$. □

The φ-functions represent thus the tail of the truncated taylor series expansion of $e^{\lambda t}$ up to a scaling factor. This is clear when rewriting (56) as

$$e^{\lambda t} = \sum_{n=0}^{k-1} \frac{(\lambda t)^n}{n!} + \lambda^k \varphi_k(\lambda, t) \tag{57}$$

POWER-BALANCED MODELLING OF CIRCUITS AS SKEW GRADIENT SYSTEMS

Rémy Müller
IRCAM-STMS (UMR 9912)
Sorbonne University
Paris, France
remy.muller@ircam.fr

Thomas Hélie *
IRCAM-STMS (UMR 9912)
Sorbonne University
Paris, France
thomas.helie@ircam.fr

ABSTRACT

This article is concerned with the power-balanced simulation of analog audio circuits, governed by nonlinear differential algebraic equations (DAE). The proposed approach is to combine principles from the port-Hamiltonian and Brayton-Moser formalisms to yield a skew-symmetric gradient system. The practical interest is to provide a solver, using an average discrete gradient, that handles differential and algebraic relations in a unified way, and avoids having to pre-solve the algebraic part. This leads to a structure-preserving method that conserves the power balance and total energy. The proposed formulation is then applied on typical nonlinear audio circuits to study the effectiveness of the method.

1. INTRODUCTION

The need for stable, accurate and power-balanced simulation of nonlinear multi-physical systems is ubiquitous in the modelling of electronic circuits or mechanical systems and the natural setting for electronic circuits leads to Differential-Algebraic Equations.

Standard methods of solving electronic circuits are the State-variable [1], Modified Nodal Analysis [2], Sparse Tableau Analysis [3] and Wave Digital Filters (WDF) [4] according to the choice of variables the system is solved for. More recently, in the audio signal processing field, it has led to the Nodal DK method [5], nonlinear state-space [6] and extension of WDF to handle multi-port nonlinearities [7].

However, the underlying geometric structure and power-balance are often lost in the process. Furthermore, most numerical schemes either introduce or dissipate energy artificially, yielding unexpected, unstable or over-damped results.

To get rid of such artefacts, a very active research is focused on geometric numerical integration methods [8] that provide a theoretical framework for structure-preserving or invariant-preserving integration of dynamical systems. Among those methods, the Port-Hamiltonian (PHS) [9] [10] and Brayton-Moser (BM) [11] [12] formalisms are dual representations [13] [14] generalizing the Hamiltonian and Lagrangian formalisms to open dynamical systems with algebraic constraints (including dissipation).

PHS have been applied successfully to the modelling of the wah-wah pedal [15], Fender Rhodes [16], brass instruments [17] and loudspeaker nonlinearities [18]. Furthermore, automated generation of the PHS equations from the graph incidence matrix of a circuit's netlist has been investigated in [19] and leads to a skew-symmetric DAE form.

This paper considers this formulation as a starting point and proposes to combine the Brayton-Moser and Port-Hamiltonian viewpoints to represent all the constitutive laws as deriving from a single potential.

The presentation is organized as follows: first, in section 2, results about power balance, passivity, and duality of flow and effort spaces are recalled and it is shown how the power-balance can be represented by Dirac structures. Section 3 shows how, for both dynamic and algebraic components, the flow and effort variables can be derived from a single power potential involving the Hamiltonian and the algebraic content and co-content potentials [20] [21]. Section 4, then shows how to perform a power-balanced structure-preserving discretization of the system using a discrete gradient [22] [23]. Section 5 shows how to solve the resulting algebraic system using Newton iteration. Finally the method is applied to some example circuits in section 6 to show the effectiveness of the approach.

2. POWER BALANCE AND DIRAC STRUCTURES

For an electronic circuit, the Tellegen theorem [24] states that the sum of powers absorbed by all circuit elements is balanced.

$$P(\mathbf{e},\mathbf{f}) := \mathbf{e}^{\mathsf{T}}\mathbf{f} = \sum_n e_n f_n = 0 \tag{1}$$

where $\mathbf{e},\mathbf{f}$ are respectively the effort and flow variables of the circuit's branch components. This is an instance of the conservation of energy principle made famous by Lavoisier with the statement *nothing is lost, nothing is created, everything is transformed*.

This principle can be formalized mathematically by Dirac structures[1] that encodes the conservative power exchange in the circuit.

2.1. Power space

For an n-port element, let $\mathcal{F}$ be an n-dimensional real vector space and denote its dual $\mathcal{E} := \mathcal{F}^*$ (the space of linear functions on $\mathcal{F}$). We call $\mathcal{F}$ the space of flows $\mathbf{f}$ and $\mathcal{E}$ the space of efforts $\mathbf{e}$. On the product space $\mathcal{P} := \mathcal{F}\times\mathcal{E}$, power is defined by the non-degenerate bilinear form

$$P(\mathbf{e},\mathbf{f}) = \langle \mathbf{e} \,|\, \mathbf{f}\rangle, \quad \forall(\mathbf{f},\mathbf{e}) \in \mathcal{P} = \mathcal{F}\times\mathcal{E} \tag{2}$$

where $\langle \mathbf{e} \,|\, \mathbf{f}\rangle$ denotes the duality product, that is the linear function $\mathbf{e} \in \mathcal{E} = \mathcal{F}^*$ acting on $\mathbf{f} \in \mathcal{F}$. If $\mathcal{F}$ is equipped with an inner product $\langle\cdot,\cdot\rangle_{\mathcal{F}}$, then $\mathcal{E} = \mathcal{F}^*$ can be identified with $\mathcal{F}$ such that $\langle \mathbf{e} \,|\, \mathbf{f}\rangle = \langle \mathbf{e},\mathbf{f}\rangle_{\mathcal{F}}$, for all $\mathbf{f} \in \mathcal{F}$, $\mathbf{e} \in \mathcal{E} \sim \mathcal{F}$. If for example, $\mathcal{F}$ is the space of currents and $\mathcal{E}$ the space of voltages, then $\langle \mathbf{e} \,|\, \mathbf{f}\rangle = \langle \mathbf{e},\mathbf{f}\rangle_{\mathcal{F}} = \mathbf{e}^{\mathsf{T}}\mathbf{f}$ denote the electrical power.

* The author acknowledges the support of the ANR-DFG (French-German) project INFIDHEM ANR-16-CE92-0028.

[1] The Kirchoff Current and Voltage laws are special cases of Dirac structures when all the components share either the same current (series connection) or the same voltage (parallel connection).

2.2. Passivity and Dirac structures

In the $2n$-dimensional space $\mathcal{P}$, a passive linear n-port can be represented as an n-dimensional subspace $\mathcal{S} \subset \mathcal{P}$ defined by n linear constraints which admits the kernel representation

$$\mathcal{S} = \{(\mathbf{f}, \mathbf{e}) \in \mathcal{P} \mid \mathbf{F}\mathbf{f} + \mathbf{E}\mathbf{e} = 0\} \tag{3}$$

with $\mathrm{rank}([\mathbf{F}\ \mathbf{E}]) = n$. Furthermore, a linear subspace $\mathcal{D} \subset \mathcal{P}$ is said to be power-conserving if

$$\langle \mathbf{e} \,|\, \mathbf{f} \rangle = 0, \quad \forall (\mathbf{f}, \mathbf{e}) \in \mathcal{D} \tag{4}$$

It becomes a (constant) Dirac structure [25] [26] if and only if it is a maximal subspace of $\mathcal{P}$ with that property i.e. $\dim(\mathcal{D}) = \dim(\mathcal{F}) = \dim(\mathcal{E})$ and it admits the following matrix representations.

Definition 2.1 (Kernel representation)**.** *The kernel form of a Dirac structure is given by the subspace*

$$\mathcal{D} = \{(\mathbf{f}, \mathbf{e}) \in \mathcal{P} \mid \mathbf{F}\mathbf{f} + \mathbf{E}\mathbf{e} = 0,\ \mathbf{E}^\mathsf{T}\mathbf{F} + \mathbf{F}\mathbf{E}^\mathsf{T} = 0\} \tag{5}$$

where $\mathbf{F}, \mathbf{E} \in \mathbb{R}^{n \times n}$ *satisfy* $\mathrm{rank}([\mathbf{F}\ \mathbf{E}]) = n$.

Definition 2.2 (Hybrid skew-symmetric representation)**.** *Let $\mathcal{D}$ be given as in* (5)*, suppose there exists a permutation of the flow and efforts variables* $\pi : (\mathbf{F}, \mathbf{E}, \mathbf{f}, \mathbf{e}) \to (\tilde{\mathbf{F}}, \tilde{\mathbf{E}}, \tilde{\mathbf{f}}, \tilde{\mathbf{e}})$ *such that* $\tilde{\mathbf{F}}$ *is invertible then*

$$\mathcal{D} = \{(\tilde{\mathbf{f}}, \tilde{\mathbf{e}}) \in \mathcal{P} \quad | \quad \tilde{\mathbf{f}} = \mathbf{J}\tilde{\mathbf{e}}, \quad \mathbf{J} = -\tilde{\mathbf{F}}^{-1}\tilde{\mathbf{E}}\} \tag{6}$$

where $\mathbf{J} = -\mathbf{J}^\mathsf{T}$ *is skew-symmetric.*

Conversely, for any skew-symmetric matrix $\mathbf{J}$, the subspace $\mathcal{D}$ is a Dirac structure and one can verify that the power balance (1) is encoded by the skew-symmetry of $\mathbf{J}$:

$$P(\tilde{\mathbf{e}}, \tilde{\mathbf{f}}) = \tilde{\mathbf{e}}^\mathsf{T}\tilde{\mathbf{f}} = \tilde{\mathbf{e}}^\mathsf{T}\mathbf{J}\tilde{\mathbf{e}} = 0. \tag{7}$$

The skew-symetric form (6) will be used in the rest of the article.

3. GRADIENT DESCRIPTION OF COMPONENTS

Circuits are then categorized into dynamical, and algebraic components where algebraic components are further separated into dissipative and external sources because the later have degenerated constitutive laws. We show how the mixed effort $\tilde{\mathbf{e}}$ can be uniformly represented as the gradient of the scalar power potential (1).

3.1. Dynamic components: Hamiltonian potential

For dynamic components with state variable $\mathbf{x}$, flow variables are defined as the time-derivative of the state ($\mathbf{f} := \dot{\mathbf{x}}$) and the effort by a constitutive law $\mathbf{e} := \hat{e}(\mathbf{x})$. It is assumed that the constitutive law derives from the gradient of an energy storage function $H(\mathbf{x}(t))$ such that by definition $\hat{e}(\mathbf{x}) := \nabla H(\mathbf{x})$ and the power is

$$P(\mathbf{e}, \mathbf{f}) = \mathbf{e}^\mathsf{T}\mathbf{f} = \nabla H(\mathbf{x}) \cdot \dot{\mathbf{x}} = \frac{\mathrm{d}}{\mathrm{d}t} H(\mathbf{x}(t)). \tag{8}$$

The Hamiltonian function can then be found using the line integral.

$$H(\mathbf{x}) = \int \underbrace{\nabla H(\mathbf{x})}_{\mathbf{e}} \cdot \underbrace{\dot{\mathbf{x}}}_{\mathbf{f}} \,\mathrm{d}t = \int \nabla H(\mathbf{x}) \cdot \mathrm{d}\mathbf{x} \tag{9}$$

This idea is illustrated with the important cases of the linear capacitor and inductor. We then show how to handle a nonlinear component with an integrable constitutive law.

3.1.1. Capacitor

For a capacitor, the state variable is given by the charge $x_C = q$, with the flow $f = i_C = \dot{q}$, and effort $e = v_C = \frac{q}{C}$. This gives the Hamiltonian

$$H(q) = \int \frac{q}{C} \cdot \dot{q} \,\mathrm{d}t = \frac{1}{C} \int q \,\mathrm{d}q = \frac{q^2}{2C} \tag{10}$$

3.1.2. Inductor

Similarily for an inductor, the state variable is given by the flux-linkage $x_L = \phi$, the flow[2] by its time-derivative $f = \dot{\phi} = v_L$ and the dual effort by $e = i_L = \frac{\phi}{L}$ with an Hamiltonian function

$$H(\phi) = \int \frac{\phi}{L} \cdot \dot{\phi} \,\mathrm{d}t = \frac{1}{L} \int \phi \cdot \mathrm{d}\phi = \frac{\phi^2}{2L} \tag{11}$$

3.1.3. Nonlinear dynamic component

For a nonlinear dynamic component with state variable x, flow $f = \dot{x}$ and a constitutive law $e = \hat{e}(x) = \tanh(x)$, its Hamiltonian storage function is given by

$$H(x) = \int_0^t \hat{e}(x) \cdot \dot{x} \,\mathrm{d}t = \int_0^x \hat{e}(\bar{x}) \cdot \mathrm{d}\bar{x} = \ln(\cosh(x)) \tag{12}$$

3.2. Algebraic components: current and voltage potentials

If we consider the power differential $\mathrm{d}P$, using the product rule,

$$\mathrm{d}P(\mathbf{e}, \mathbf{f}) = \mathrm{d}(\mathbf{e} \cdot \mathbf{f}) = \mathbf{e} \cdot \mathrm{d}\mathbf{f} + \mathbf{f} \cdot \mathrm{d}\mathbf{e}. \tag{13}$$

Integration over a path Γ gives the integration by parts formula

$$\mathbf{e} \cdot \mathbf{f}\Big|_{\partial\Gamma} = \int_\Gamma \mathbf{e} \cdot \mathrm{d}\mathbf{f} + \int_\Gamma \mathbf{f} \cdot \mathrm{d}\mathbf{e}. \tag{14}$$

So, for components defined by algebraic constitutive laws $\Gamma = \{(\mathbf{e}, \mathbf{f}) \in \mathcal{P} \mid \mathbf{f} = \hat{f}(\mathbf{e})\}$, (respectively $\mathbf{e} = \hat{e}(\mathbf{f})$), the flow and effort potentials[3] are defined by the line integrals

$$D(\mathbf{f}) := \int_0^{\mathbf{f}} \hat{e}(\bar{\mathbf{f}}) \cdot \mathrm{d}\bar{\mathbf{f}}, \qquad D^*(\mathbf{e}) := \int_0^{\mathbf{e}} \hat{f}(\bar{\mathbf{e}}) \cdot \mathrm{d}\bar{\mathbf{e}}. \tag{15}$$

And according to (14), the instantaneous power is given, for $(\mathbf{e}, \mathbf{f}) \in \Gamma$, by (see figure 1 for a geometric interpretation and proof)

$$P(\mathbf{e}, \mathbf{f}) = \mathbf{e} \cdot \mathbf{f} = D(\mathbf{f}) + D^*(\mathbf{e}). \tag{16}$$

The flow and efforts can then be respectively obtained by partial derivatives of the power potential as

$$\mathbf{e} = \frac{\partial P}{\partial \mathbf{f}} = \nabla D(\mathbf{f}), \qquad \text{or} \qquad \mathbf{f} = \frac{\partial P}{\partial \mathbf{e}} = \nabla D^*(\mathbf{e}). \tag{17}$$

So in the case of a flow (resp. effort) controlled component the power can be expressed as a function of a single variable using either

$$P(\mathbf{e}) = \mathbf{e} \cdot \nabla D^*(\mathbf{e}) \qquad \text{or} \qquad P(\mathbf{f}) = \nabla D(\mathbf{f}) \cdot \mathbf{f}. \tag{18}$$

[2] Note that according to the energy domain (electric, magnetic, ...), the roles of flow and efforts need not necessarily be associated to the current and voltage. The convention adopted here, is that the flow of dynamic components is given by the time-derivative of the energy variable, while the effort is given by the gradient of the energy potential.

[3] These potentials are also called the content and co-content [20] [21].

3.2.1. *Linear resistor*

For a current-controlled (resp. voltage-controlled) resistor, the constitutive law is $v = \hat{e}(i) = Ri$ (resp. $i = \hat{f}(v) = v/R$). By consequence its current and voltage potentials are given by

$$D(i) = \int_0^i \hat{e}(f)\,\mathrm{d}f = \int_0^i Rf\,\mathrm{d}f = \frac{Ri^2}{2} \tag{19}$$

$$D^*(v) = \int_0^v \hat{f}(e)\,\mathrm{d}e = \int_0^v \frac{e}{R}\,\mathrm{d}e = \frac{v^2}{2R}. \tag{20}$$

Introduce function P as $P(v,i) = D(i) + D^*(v)$, then, for all (v,i) belonging on the characteristic curve, the power can be given by $v \cdot i$ (product-type), $P(v,i)$ (sum-type), $P(v,\hat{f}(v))$ (voltage-controlled) and $P(\hat{e}(i),i)$ (current-controlled), that is

$$P(v,i) = v{\cdot}i = D(i){+}D^*(v) = \frac{1}{2}\left(Ri^2 + \frac{v^2}{R}\right) = \frac{v^2}{R} = Ri^2. \tag{21}$$

In this particular case, we have $D(i) = D^*(v) = Ri^2$ because of linearity (for $v = Ri$) but this result should not be extrapolated as the next example will show.

3.2.2. *P-N Diode*

For a voltage controlled P-N diode, the constitutive law is given by

$$i = \hat{f}(v) = I_S\left(\exp\left(\frac{v}{nV_T}\right) - 1\right) \tag{22}$$

where I_S is the saturation current, n the ideality factor and V_T the thermal voltage. Its voltage potential is given by

$$D^*(v) = \int_0^v \hat{f}(e)\,\mathrm{d}e = nV_T I_S\left(\exp\left(\frac{v}{nV_T}\right) - \frac{v}{nV_T} - 1\right). \tag{23}$$

Direct integration for the current potential does not lead to an easily integrable primitive, however because of bijectivity, we can evaluate it indirectly by using the inverse map

$$v = \hat{e}(i) = \hat{f}^{-1}(i) = nV_T \ln\left(1 + \frac{i}{I_S}\right), \quad i > -I_S \tag{24}$$

and the Legendre transform $D(i) = \left[vi - D^*(v)\right]_{v=\hat{f}^{-1}(i)}$:

$$D(i) = nV_T I_S\left(\left(1 + \frac{i}{I_S}\right)\ln\left(1 + \frac{i}{I_S}\right) - \frac{i}{I_S}\right) \tag{25}$$

Using the above definitions, the current and voltage potentials being known, the component can be used as being either flow or effort-driven according to the constraints imposed by the circuit interconnections.

3.3. External sources

For external voltage (resp. current) sources, the constitutive laws $v = \hat{e}(i) = V$, (resp. $i = \hat{f}(v) = I$) are independent of the current (resp. voltage) variables and not bijective, with V (resp. I) being the source parameter. This gives the powers

$$P_V(v,i) = Vi = D(i), \qquad P_I(v,i) = vI = D^*(v). \tag{26}$$

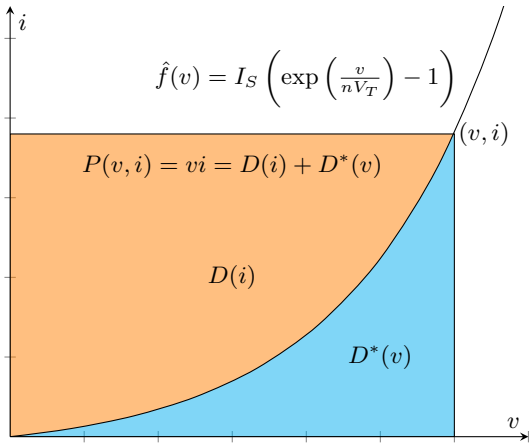


Figure 1: The areas occupied by the diode power $P(v,i)$ and the current and voltage potentials $D(i)$ and $D^*(v)$ are shown in the (v,i) plane for $I_S = 1, nV_T = 1$. It is geometrically clear that the current and voltage potentials are complimentary and their sum equals the power vi. It is also clear that in the nonlinear case $D(i) \neq D^*(v)$.

By consequence, for voltage (resp. current) sources, the voltage potential $D^*(v)$ (resp. current potential $D(i)$) is degenerate and null.

3.4. Summary

Using an appropriate permutation π (cf definition 2.2), the mixed flow $\tilde{\mathbf{f}}$ and its dual $\tilde{\mathbf{e}}$ can be parametrized by a state variable $\mathbf{x} \in \mathbb{R}^n$, a dissipative variable $\mathbf{w} \in \mathbb{R}^p$ and an output $\mathbf{y} \in \mathbb{R}^m$, where the potential $Z(\mathbf{w})$ (resp. $S(\mathbf{y})$) is an appropriate choice among the dissipative (resp. external) current and voltage potentials imposed by the permutation π. (Please refer to [19] for more details.)

$$\tilde{\mathbf{f}} := [\dot{\mathbf{x}}, \mathbf{w}, \mathbf{y}]^\mathsf{T} \tag{27}$$

$$\tilde{\mathbf{e}} := [\nabla H(\mathbf{x}), \nabla Z(\mathbf{w}), \nabla S(\mathbf{y})]^\mathsf{T} \tag{28}$$

The power potential[4] (1) can then be expressed as

$$P(\tilde{\mathbf{e}}, \tilde{\mathbf{f}}) = \tilde{\mathbf{e}}^\mathsf{T}\tilde{\mathbf{f}} = \underbrace{\nabla H(\mathbf{x})^\mathsf{T}\dot{\mathbf{x}}}_{P_c} + \underbrace{\nabla Z(\mathbf{w})^\mathsf{T}\mathbf{w}}_{P_d} + \underbrace{\nabla S(\mathbf{y})^\mathsf{T}\mathbf{y}}_{P_e}. \tag{29}$$

Combining the definitions (27) and (28), with the Dirac structure (6), leads to the skew-symmetric gradient form of Differential-Algebraic Port-Hamiltonian equations as

$$\underbrace{\begin{bmatrix}\dot{\mathbf{x}}\\ \mathbf{w}\\ \mathbf{y}\end{bmatrix}}_{\tilde{\mathbf{f}}} = \mathbf{J}\underbrace{\begin{bmatrix}\nabla H(\mathbf{x})\\ \nabla Z(\mathbf{w})\\ \nabla S(\mathbf{y})\end{bmatrix}}_{\tilde{\mathbf{e}}} \iff \frac{\partial P}{\partial \tilde{\mathbf{e}}} = \mathbf{J}\frac{\partial P}{\partial \tilde{\mathbf{f}}} \tag{30}$$

[4] Note that because of the uniform usage of the *receiver convention* for each component (including sources), the power potentials represent the *absorbed* power by each component. This means that dissipative components will absorb *positive power*, while sources will, on average, absorb *negative power* to compensate for losses (but can temporarily receive power).

Integrating (29) over a time interval $[t_0, t_1]$ combined with the power balance (7), leads to the conservation of the total energy

$$\Delta E = H(\mathbf{x})\Big|_{t_0}^{t_1} + \int_{t_0}^{t_1} P_d(t)\,\mathrm{d}t + \int_{t_0}^{t_1} P_e(t)\,\mathrm{d}t = 0. \quad (31)$$

4. STRUCTURE-PRESERVING INTEGRATION SCHEME

The main objective of the numerical scheme is first and foremost, to provide a structure-preserving method that conserves the invariant (31) in discrete-time over each time-step. This offers the strong guarantee that no artificial energy is either consumed or created by the numerical scheme.To achieve this goal, thanks to the unified representation of DAE circuits as gradient systems introduced in section 3, it is now possible to generalize the usage of discrete gradient methods [22] [23] for *both* dynamic and algebraic components.

4.1. Discrete Gradients

Given a scalar potential $H : \mathbb{R}^n \mapsto \mathbb{R}$, a point $\mathbf{x} \in \mathbb{R}^n$ and a variation $\delta\mathbf{x} \in \mathbb{R}^n$, a necessary and sufficient condition for a function $\overline{\nabla} H(\mathbf{x}, \delta\mathbf{x}) : \mathbb{R}^n \times \mathbb{R}^n \mapsto \mathbb{R}^n$ to be a discrete gradient is given by

$$\overline{\nabla} H(\mathbf{x}, \delta\mathbf{x}) \cdot \delta\mathbf{x} = H(\mathbf{x} + \delta\mathbf{x}) - H(\mathbf{x}) \quad (32)$$

$$\overline{\nabla} H(\mathbf{x}, 0) = \nabla H(\mathbf{x}) \quad (33)$$

Definition 4.1 (Average Discrete Gradient)**.** *Let $\mathbf{x}, \delta\mathbf{x} \in \mathbb{R}^n$, and $H : \mathbb{R}^n \mapsto \mathbb{R}$ be a scalar potential. The average discrete gradient is defined for an affine trajectory model $\hat{\mathbf{x}}(\tau) = \mathbf{x} + \tau\delta\mathbf{x}$ by*

$$\overline{\nabla} H(\mathbf{x}, \delta\mathbf{x}) := \int_0^1 \nabla H(\mathbf{x} + \tau\delta\mathbf{x})\,\mathrm{d}\tau \quad (34)$$

Furthermore, using the gradient theorem, for separable potentials of the form

$$H(\mathbf{x}) = \sum_{i=1}^{N} H_i(x_i), \quad (35)$$

the discrete gradient can be computed *exactly* by finite differences on each scalar potential. It is given component-wise by

$$[\overline{\nabla} H(\mathbf{x}, \delta\mathbf{x})]_i := \begin{cases} \dfrac{H_i(x_i + \delta x_i) - H_i(x_i)}{\delta x_i} & \delta x_i \neq 0 \\ \dfrac{\partial H_i}{\partial x_i}(x_i) & \delta x_i = 0 \end{cases} \quad (36)$$

Finally, and *only in the case of quadratic potentials* of the form $H(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{W}\mathbf{x}$ with $\mathbf{W} = \mathbf{W}^T \succeq 0$, does the discrete gradient correspond to evaluation of the gradient at the mid-point.

$$\overline{\nabla} H(\mathbf{x}, \delta\mathbf{x}) = \nabla H\left(\mathbf{x} + \frac{1}{2}\delta\mathbf{x}\right) = \mathbf{W}\left(\mathbf{x} + \frac{1}{2}\delta\mathbf{x}\right) \quad (37)$$

The following result will also be exploited in the next section.

Property 4.1. *Given a separable potential $H : \mathbb{R}^n \mapsto \mathbb{R}$, as in* (35) *of class $\mathcal{C}^2$, a point $\mathbf{x} \in \mathbb{R}^n$, a variation $\boldsymbol{\nu} \in \mathbb{R}^n$ and its discrete gradient $\overline{\nabla} H(\mathbf{x}, \boldsymbol{\nu})$ defined as* (36)*, the derivative of the discrete gradient with respect to the variation $\boldsymbol{\nu}$ is the diagonal matrix $\partial_{\boldsymbol{\nu}}\overline{\nabla} H : (\mathbf{x}, \boldsymbol{\nu}) \in \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^{n\times n}$ with entries*

$$\left[\partial_{\boldsymbol{\nu}}\overline{\nabla} H\right]_{i,i} = \begin{cases} \dfrac{\nabla H_i(x_i + \nu_i) - \overline{\nabla} H_i(x_i, \nu_i)}{\nu_i} & \nu_i \neq 0 \\ \dfrac{1}{2}\dfrac{\partial^2 H_i}{\partial x_i^2}(x_i) & \nu_i = 0 \end{cases} \quad (38)$$

Proof. see Appendix A. □

4.2. Averaged System

Assuming over each time step $\Omega_n = [t_n, t_n + h]$, an affine trajectory model

$$\mathbf{z}(t_n + h\tau) = \mathbf{z}_n + \tau\delta\mathbf{z}_n \quad (39)$$

where $\mathbf{z} = [\mathbf{x}, \mathbf{w}, \mathbf{y}]^T$, and integrating (30) over Ω_n, we obtain the discrete structure-preserving system

$$\begin{bmatrix} \delta\mathbf{x}_n/h \\ \bar{\mathbf{w}}_n \\ \bar{\mathbf{y}}_n \end{bmatrix} = \mathbf{J} \begin{bmatrix} \overline{\nabla} H(\mathbf{x}_n, \delta\mathbf{x}_n) \\ \overline{\nabla} Z(\mathbf{w}_n, \delta\mathbf{w}_n) \\ \overline{\nabla} S(\mathbf{y}_n, \delta\mathbf{y}_n) \end{bmatrix} \quad (40)$$

where $\bar{\mathbf{w}}_n = \mathbf{w}_n + \delta\mathbf{w}_n/2$, $\bar{\mathbf{y}}_n = \mathbf{y}_n + \delta\mathbf{y}_n/2$. The DAE system (30) has been converted to an algebraic system that needs to be to solved for the average variation $\delta\mathbf{z}_n = [\delta\mathbf{x}_n, \delta\mathbf{w}_n, \delta\mathbf{y}_n]^\mathsf{T}$.

5. NEWTON ITERATION

Denote the variation $\boldsymbol{\nu} = \delta\mathbf{z}_n$, solving the discrete algebraic system (40) can be rewritten as the root-finding problem

$$F(\boldsymbol{\nu}^*) = 0 \quad (41)$$

where $\boldsymbol{\nu}^*$ is the looked for solution and F is defined by

$$F(\boldsymbol{\nu}) := \mathbf{D}_0\mathbf{z}_n + \mathbf{D}_1\boldsymbol{\nu} - \mathbf{J}\overline{\nabla}_{\tilde{\mathbf{f}}} P(\mathbf{z}_n, \boldsymbol{\nu}), \quad (42)$$

with $\mathbf{D}_0 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \mathbf{I}_p & 0 \\ 0 & 0 & \mathbf{I}_m \end{bmatrix}$, $\mathbf{D}_1 = \begin{bmatrix} \mathbf{I}_n/h & 0 & 0 \\ 0 & \mathbf{I}_p/2 & 0 \\ 0 & 0 & \mathbf{I}_m/2 \end{bmatrix}$, where $\mathbf{I}_n$ denote the $n \times n$ identity matrix and $\overline{\nabla}_{\tilde{\mathbf{f}}} P = [\overline{\nabla} H, \overline{\nabla} Z, \overline{\nabla} S]^\mathsf{T}$.

5.1. Newton update

For an estimate $\boldsymbol{\nu}_k$ and a perturbation $\Delta\boldsymbol{\nu}_k$, the true solution $\boldsymbol{\nu}^*$ of (41) can be written as $\boldsymbol{\nu}^* = \boldsymbol{\nu}_k + \Delta\boldsymbol{\nu}_k$. Taylor series expansion of F around $\boldsymbol{\nu}_k$, with $\|\Delta\boldsymbol{\nu}_k\|$ sufficiently small yields

$$0 = F(\boldsymbol{\nu}_k + \Delta\boldsymbol{\nu}_k) = F(\boldsymbol{\nu}_k) + [F'(\boldsymbol{\nu}_k)](\Delta\boldsymbol{\nu}_k) + \mathcal{O}(\|\Delta\boldsymbol{\nu}_k\|^2). \quad (43)$$

If the Jacobian F' is invertible, neglecting high-order terms and solving for $\Delta\boldsymbol{\nu}$ leads to the Newton update

$$\Delta\boldsymbol{\nu}_k := -F'(\boldsymbol{\nu}_k)^{-1}F(\boldsymbol{\nu}_k), \quad \boldsymbol{\nu}_{k+1} := \boldsymbol{\nu}_k + \Delta\boldsymbol{\nu}_k, \quad (44)$$

where the Jacobian of F is given by

$$F'(\boldsymbol{\nu}) = \mathbf{D}_1 - \mathbf{J}\left(\partial_{\boldsymbol{\nu}}\overline{\nabla}_{\tilde{\mathbf{f}}} P(\mathbf{z}_n, \boldsymbol{\nu})\right). \quad (45)$$

For a separable potential P, using property (4.1), $\partial_{\boldsymbol{\nu}}\overline{\nabla}_{\tilde{\mathbf{f}}} P$ is a diagonal matrix that can be computed from the knowledge of the gradient, Hessian and discrete gradient of the potential.

5.2. Convergence and stiffness

If the eigenvalues of the matrix $\mathbf{A} = \mathbf{D}_1^{-1}\mathbf{J}\left(\partial_{\boldsymbol{\nu}}\overline{\nabla}_{\tilde{\mathbf{f}}}P(\mathbf{z}_n,\boldsymbol{\nu})\right)$ are such that $\|\mathbf{A}\|_2 = \max(|\lambda_i|) < 1$, the fixed-point induced by (40) is contracting. The Banach fixed-point theorem guarantees existence and unicity of the solution. It is then possible to approximate the inverse of the Jacobian with the Neumann series identity

$$(\mathbf{I}-\mathbf{A})^{-1} = \sum_{k=0}^{\infty}\mathbf{A}^k \approx \mathbf{I}+\mathbf{A}+\mathbf{A}^2+\dots \tag{46}$$

to get the first (or any higher) order approximation

$$F'(\boldsymbol{\nu})^{-1} \approx \left(\mathbf{I}+\mathbf{D}_1^{-1}\mathbf{J}\left(\partial_{\boldsymbol{\nu}}\overline{\nabla}_{\tilde{\mathbf{f}}}P(\mathbf{z}_n,\boldsymbol{\nu})\right)\right)\mathbf{D}_1^{-1} \tag{47}$$

If $\max|\lambda_i| \geq 1$, the system is said to be stiff, the series (46) is divergent, and the approximation (47) is no longer valid. Solving the system then requires a matrix inversion for each iteration. Using the Newton-Kantorovich theorem, for a starting point $\boldsymbol{\nu}_0$, if there exists positive constants β_0, γ, h_0, such that $\|F'(\boldsymbol{\nu}_0)^{-1}\| \leq \beta_0$, $F'(\boldsymbol{\nu})$ is locally γ-Lipschitz and $h_0 := \|\Delta\boldsymbol{\nu}_0\|\,\beta_0\gamma < 1/2$, then the sequence $\{\boldsymbol{\nu}_k\}$ converges quadratically to some unique $\boldsymbol{\nu}^*$ such that $F(\boldsymbol{\nu}^*) = 0$. Please refer to [27] for more details.

6. CIRCUIT EXAMPLES

6.1. Envelope Follower

We consider the envelope follower circuit shown in figure 3 with parameters $C = 100$ pF, $I_S = 2.52$ nA, $V_T = 23$ mV and $n = 1.96$. Kirchoff laws leads to the following Dirac structure:

$$\underbrace{\begin{bmatrix} i_C \\ v_D \\ i_S \end{bmatrix}}_{\tilde{\mathbf{f}}} = \underbrace{\begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 1 \\ 0 & -1 & 0 \end{bmatrix}}_{\mathbf{J}} \underbrace{\begin{bmatrix} v_C \\ i_D \\ v_S \end{bmatrix}}_{\tilde{\mathbf{e}}}. \tag{48}$$

For this circuit we have $\mathbf{x} = [q]$, $\mathbf{w} = [v_D]$, $\mathbf{y} = [i_S]$, $\tilde{\mathbf{f}} = [\dot{q}, v_D, i_S]^\mathsf{T}$ and the following potentials

$$H(q) = \frac{q^2}{2C}, \tag{49}$$

$$Z(v_D) = nV_T I_S\left(\exp\left(\frac{v_D}{nV_T}\right)-1\right) - v_D I_S, \tag{50}$$

$$S(i_S) = V i_S. \tag{51}$$

Taking their gradients gives the right-hand side vector

$$\tilde{\mathbf{e}} = \begin{bmatrix} v_C \\ i_D \\ v_S \end{bmatrix} = \begin{bmatrix} \nabla H(q) \\ \nabla Z(v_D) \\ \nabla S(i_S) \end{bmatrix} = \begin{bmatrix} q/C \\ I_S\left(\exp\left(\frac{v_D}{nV_T}\right)-1\right) \\ V \end{bmatrix} \tag{52}$$

and the product $\tilde{\mathbf{e}}^\mathsf{T}\tilde{\mathbf{f}}$ gives the power balance potential

$$P(\tilde{\mathbf{e}},\tilde{\mathbf{f}}) = \underbrace{\nabla H(q)\dot{q}}_{P_C(q)} + \underbrace{\nabla Z(v_D)v_D}_{P_D(v_D)} + \underbrace{\nabla S(i_S)i_S}_{P_S(i_S)}. \tag{53}$$

For the capacitor and voltage source, we obtain the discrete gradients

$$\overline{\nabla}H(q,\delta q) = \frac{1}{C}\left(q+\frac{\delta q}{2}\right), \qquad \overline{\nabla}S(i,\delta i) = V, \tag{54}$$

and after some algebraic manipulations (see appendix B), the discrete gradient of the diode potential can be expressed as

$$\overline{\nabla}Z(v,\delta v) = I_S\left(\exp\left(\frac{v+\delta v/2}{nV_T}\right)\operatorname{sinhc}\left(\frac{\delta v}{2nV_T}\right)-1\right). \tag{55}$$

where the sinhc term ($\operatorname{sinhc} := \sinh(x)/x$) acts as a correction compared to evaluation of the gradient at the mid-point.

6.2. Diode Clipper

We consider the diode clipper circuit shown in figure 5 with parameters $R = 1$ kΩ, $C = 100$ nF, $I_S = 2.52$ fA, $V_T = 23$ mV and $n = 1$. For the two diodes, with $v_D := v_{D_1}$ and the diodes current $i_D := i_{D_1} - i_{D_2}$, the constitutive law is

$$i_D = \hat{f}(v_D) = 2I_S\sinh\left(\frac{v_D}{nV_T}\right). \tag{56}$$

Its integration gives the voltage potential

$$D_D^*(v_D) = \int_0^{v_D}\hat{f}(v)dv = 2nV_T I_S\left(\cosh\left(\frac{v_D}{nV_T}\right)-1\right). \tag{57}$$

Application of Kirchoff laws leads to the following Dirac structure:

$$\underbrace{\begin{bmatrix} i_C \\ v_R \\ v_D \\ i_S \end{bmatrix}}_{\tilde{\mathbf{f}}} = \underbrace{\begin{bmatrix} 0 & 1 & -1 & 0 \\ -1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{bmatrix}}_{\mathbf{J}} \underbrace{\begin{bmatrix} v_C \\ i_R \\ i_D \\ v_S \end{bmatrix}}_{\tilde{\mathbf{e}}}. \tag{58}$$

For this circuit, $\mathbf{x} = [q]$, $\mathbf{w} = [v_R, v_D]^\mathsf{T}$, $\mathbf{y} = [i_S]$, $\tilde{\mathbf{f}} = [\dot{q}, v_R, v_D, i_S]^\mathsf{T}$ and the potentials are

$$H(q) = \frac{q^2}{2C}, \quad Z(v_R,v_D) = \frac{v_R^2}{2R} + D_D^*(v_D), \quad S(i_S) = V i_S. \tag{59}$$

Their gradients regenerates the mixed effort

$$\tilde{\mathbf{e}} = \begin{bmatrix} v_C \\ i_R \\ i_D \\ v_S \end{bmatrix} = \begin{bmatrix} \nabla H \\ \nabla Z_R \\ \nabla Z_D \\ \nabla S \end{bmatrix} = \begin{bmatrix} q/C \\ v_R/R \\ 2I_S\sinh\left(\frac{v_D}{nV_T}\right) \\ V \end{bmatrix} \tag{60}$$

and the product $\tilde{\mathbf{e}}^\mathsf{T}\tilde{\mathbf{f}}$ gives the power balance potential

$$P(\tilde{\mathbf{e}},\tilde{\mathbf{f}}) = \underbrace{\nabla H(q)\dot{q}}_{P_C(q)} + \underbrace{\nabla Z_R(v_R)v_R}_{P_R(v_R)} + \underbrace{\nabla Z_D(v_D)v_D}_{P_D(v_D)} + \underbrace{\nabla S(i_S)i_S}_{P_S(i_S)}. \tag{61}$$

Similarily as in the envelope follower case, we have the discrete gradients (54) for the capacitor and voltage source, with

$$\overline{\nabla}Z_R(v,\delta v) = \frac{1}{R}\left(v+\frac{\delta v}{2}\right) \tag{62}$$

for the resistor, and after some algebraic manipulations, the discrete gradient of the diodes potential can be expressed as

$$\overline{\nabla}Z_D(v,\delta v) = 2I_S\sinh\left(\frac{v+\delta v/2}{nV_T}\right)\operatorname{sinhc}\left(\frac{\delta v}{2nV_T}\right). \tag{63}$$

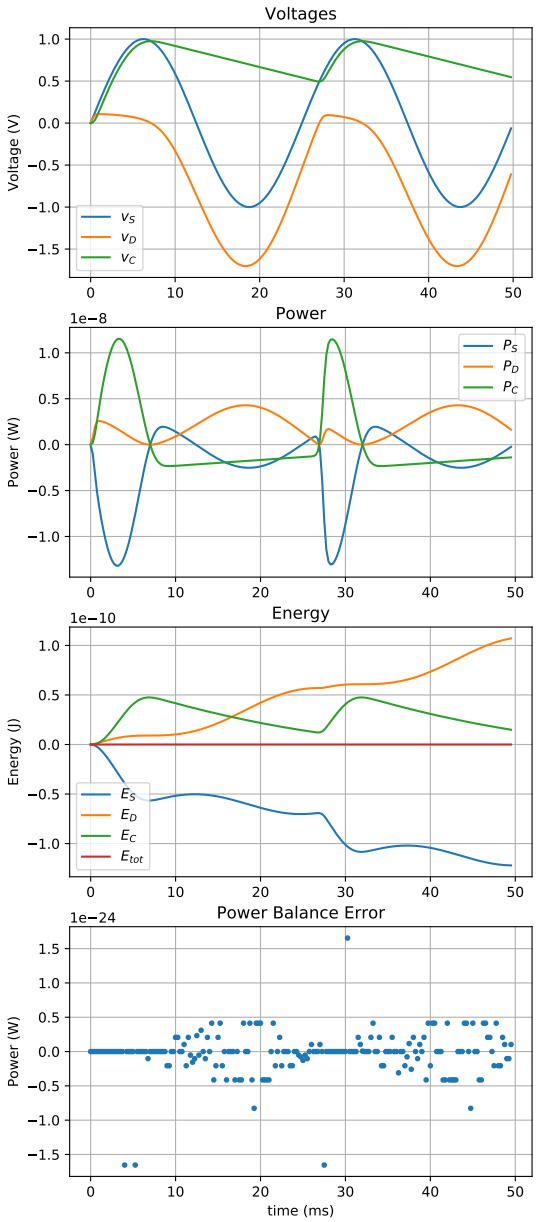


Figure 2: Envelope follower circuit driven by a 1V sinusoidal input with fundamental frequency $f = 40$ Hz, $f_s = 4$ kHz.

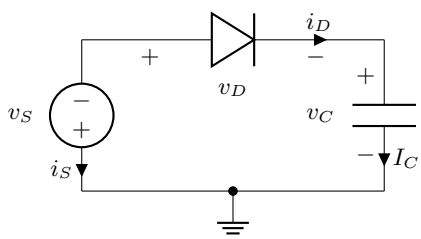


Figure 3: Envelope Follower circuit

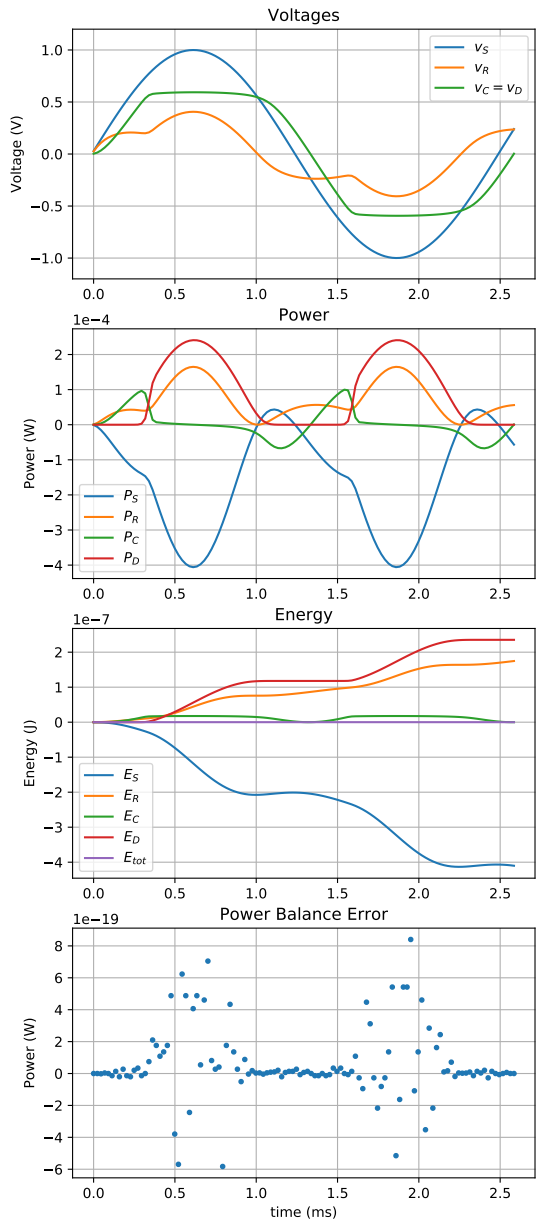


Figure 4: Diode clipper circuit driven by a 1V sinusoidal input with fundamental frequency $f = 400$ Hz, $f_s = 44.1$ kHz.

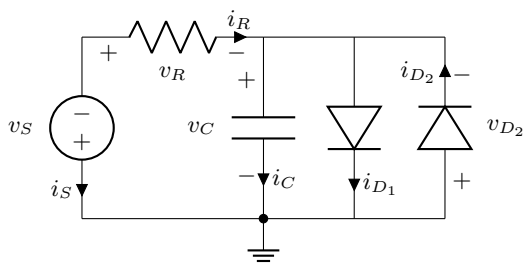


Figure 5: Diode Clipper circuit

6.3. Analysis

Simulation results for both circuits are shown in figure 2 and figure 4 with respective sampling frequencies 4 kHz and 44.1 kHz. We remark that in both cases, the power balance is satisfied with high precision. The relative error is of the order of the machine epsilon ($\epsilon = 2^{-53} \approx 1.11 \cdot 10^{-16}$). This results in a vanishing total energy variation.

For dissipative components, the absorbed power is always positive; the dissipated energy is thus monotonously increasing. For dynamic components and sources, the power is alternatively absorbed and released, the difference being that sources have a decreasing average energy trend to compensate for losses in the dissipative components.

Existence and uniqueness of the fixed points are guaranteed if $h < C/\gamma_D$ for the envelope follower and if $h < C/\max(\gamma_D, \gamma_R)$ for the diode clipper (proof is ommited) where γ_K stands for the local Lipschitz constants $\gamma_K = \max_\nu |\partial_\nu \overline{\nabla} Z_K(v_{K_0}, \nu)|$ of the diode and resistor components in a neighborhood around $\boldsymbol{\nu}_0$.

For the diode clipper circuit, the fixed-point does not converge, but the Newton iteration does. We can remark that each time the diodes are saturating, the precision of the power balance is slightly deteriorated. This can be explained by two facts: the dissipated power is also increasing during saturation and the system becomes stiff, thus the numerical conditioning of the Jacobian in the Newton iteration gets worse.

7. CONCLUSION

The main contribution of this paper consists in a) using the power-balance as the core object from which all quantities in the system are derived, b) generalizing the usage of potentials and their gradients to represent the flow and effort variables for both dynamic an algebraic components, c) keeping the sparse skew-symmetric structure matrix $\mathbf{J}$ until numerical simulation, d) integration of the system using the average discrete gradient. This leads to a consistent structure-preserving approximation that conserves the form of the original system in discrete-time.

It is also shown that the Jacobian of the Newton iteration has a special structure that only involves diagonal and skew-symmetric matrices. It can be computed only from the knowledge of the potentials associated with each component and stiffness can be inferred by inspection of the derivatives of the discrete gradient. Furthermore the structure-preserving approach offers a valuable tool to monitor the quality of our approximations with respect to the power balance.

The main drawback of the approach is a direct consequence from its strength. Indeed, the preservation of the power balance, prevents the use of L-stable integrators (which limit the stiffness by introducing artificial numerical dissipation) such as the Backward Difference Formulas or Radau IIa methods [28] [29]. This imposes some restrictions on the step size or the use of adaptive strategies. However, since the average integration of the system can be interpreted as a lowpass projector and first-order anti-aliasing filter [30], parasitic oscillations at the Nyquist frequency which are typical of stiff systems are attenuated during the simulation.

Further perspectives include the use of higher-order trajectory models, exponential integrators [31] which have shown to be effective in the simulation of stiff systems and more generally Lie-group integrators [32] [33] whose trajectories belong, by construction, to the system manifold.

8. ACKNOWLEDGMENTS

The second author acknowledges the support of the ANR-DFG (French-German) project INFIDHEM ANR-16-CE92-0028.

9. REFERENCES

[1] E. S. Kuh and R. A. Rohrer, "The state-variable approach to network analysis," *Proceedings of the IEEE*, vol. 53, no. 7, pp. 672–686, 1965.

[2] C.-W. Ho, A. Ruehli, and P. Brennan, "The modified nodal approach to network analysis," *IEEE Transactions on circuits and systems*, vol. 22, no. 6, pp. 504–509, 1975.

[3] G. Hachtel, R. Brayton, and F. Gustavson, "The sparse tableau approach to network analysis and design," *IEEE Transactions on circuit theory*, vol. 18, no. 1, pp. 101–113, 1971.

[4] K. Meerkotter and R. Scholz, "Digital simulation of nonlinear circuits by wave digital filter principles," in *Circuits and Systems*. IEEE, 1989, pp. 720–723.

[5] D. T. Yeh, J. S. Abel, and J. O. Smith, "Automated physical modeling of nonlinear audio circuits for real-time audio effects–part i: Theoretical development," *IEEE transactions on audio, speech, and language processing*, vol. 18, no. 4, pp. 728–737, 2010.

[6] M. Holters and U. Zölzer, "A generalized method for the derivation of non-linear state-space models from circuit schematics," in *Signal Processing Conference (EUSIPCO), 2015 23rd European*. IEEE, 2015, pp. 1073–1077.

[7] K. J. Werner, V. Nangia, J. O. Smith III, and J. S. Abel, "Resolving wave digital filters with multiple/multiport nonlinearities," in *Proc. 18th Conf. Digital Audio Effects*, 2015, pp. 387–394.

[8] E. Hairer, C. Lubich, and G. Wanner, *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations; 2nd ed.* Dordrecht: Springer, 2006.

[9] A. van der Schaft and D. Jeltsema, "Port-hamiltonian systems theory: An introductory overview," *Foundations and Trends in Systems and Control*, vol. 1, no. 2-3, pp. 173–378, 2014.

[10] A. van der Schaft, "Port-hamiltonian systems: an introductory survey," in *Proceedings of the International Congress of Mathematicians Vol. III: Invited Lectures*, Madrid, Spain, 2006, pp. 1339–1365.

[11] R. Brayton and J. Moser, "A theory of nonlinear networks. i," *Quarterly of Applied Mathematics*, vol. 22, no. 1, pp. 1–33, 1964.

[12] ——, "A theory of nonlinear networks. ii," *Quarterly of applied mathematics*, vol. 22, no. 2, pp. 81–104, 1964.

[13] A. J. van der Schaft, "On the relation between port-hamiltonian and gradient systems," *IFAC Proceedings Volumes*, vol. 44, no. 1, pp. 3321–3326, 2011.

[14] D. Jeltsema and J. M. Scherpen, "A dual relation between port-hamiltonian systems and the brayton–moser equations for nonlinear switched rlc circuits," *Automatica*, vol. 39, no. 6, pp. 969–979, 2003.

[15] A. Falaize and T. Hélie, "Simulation of an analog circuit of a wah pedal: a port-Hamiltonian approach," in *135th convention of the Audio Engineering Society*, New-York, United States, Oct. 2013, pp. –.

[16] ——, "Passive simulation of the nonlinear port-Hamiltonian modeling of a Rhodes Piano," *Journal of Sound and Vibration*, vol. 390, pp. 289–309, Mar. 2017.

[17] N. Lopes and T. Hélie, "Energy Balanced Model of a Jet Interacting With a Brass Player's Lip," *Acta Acustica united with Acustica*, vol. 102, no. 1, pp. 141–154, 2016.

[18] A. Falaize and T. Hélie, "Passive simulation of electrodynamic loudspeakers for guitar amplifiers: a port- Hamiltonian approach," in *International Symposium on Musical Acoustics*, Le Mans, France, Jul. 2014, pp. 1–5.

[19] A. Falaize and T. Hélie, "Passive guaranteed simulation of analog audio circuits: A port-hamiltonian approach," *Applied Sciences*, vol. 6, no. 10, 2016.

[20] W. Millar, "Some general theorems for non-linear systems possessing resistance," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 42, no. 333, pp. 1150–1160, 1951.

[21] C. Cherry, "Some general theorems for non-linear systems possessing reactance," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 42, no. 333, pp. 1161–1177, 1951.

[22] R. I. McLachlan, G. Quispel, and N. Robidoux, "Geometric integration using discrete gradients," *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 357, no. 1754, pp. 1021–1045, 1999.

[23] E. Celledoni, V. Grimm, R. McLachlan, D. McLaren, D. O'Neale, B. Owren, and G. Quispel, "Preserving energy resp. dissipation in numerical PDEs using the 'average vector field' method," *Journal of Computational Physics*, vol. 231, no. 20, pp. 6770 – 6789, 2012.

[24] B. D. Tellegen, "A general network theorem, with applications," *Philips Res Rep*, vol. 7, pp. 256–269, 1952.

[25] I. Y. Dorfman, "Dirac structures of integrable evolution equations," *Physics Letters A*, vol. 125, no. 5, pp. 240–246, 1987.

[26] T. Courant and A. Weinstein, "Beyond poisson structures," *Action hamiltoniennes de groupes. Troisieme théoreme de Lie (Lyon, 1986)*, vol. 27, pp. 39–49, 1988.

[27] P. Deuflhard, *Newton methods for nonlinear problems: affine invariance and adaptive algorithms*. Springer, 2011, vol. 35.

[28] G. Wanner and E. Hairer, *Solving ordinary differential equations II: Stiff and Differential-Algebraic Problems*. Springer, 1991, vol. 14.

[29] J. C. Butcher, *Numerical methods for ordinary differential equations*. John Wiley & Sons, 2016.

[30] R. Müller and T. Hélie, "Trajectory anti-aliasing on guaranteed-passive simulation of nonlinear physical systems," in *Proc. 20th Conf. Digital Audio Effects*, 2017.

[31] M. Hochbruck and A. Ostermann, "Exponential integrators," *Acta Numerica*, vol. 19, pp. 209–286, 2010.

[32] E. Celledoni, H. Marthinsen, and B. Owren, "An introduction to lie group integrators–basics, new developments and applications," *Journal of Computational Physics*, vol. 257, pp. 1040–1061, 2014.

[33] A. Iserles, H. Z. Munthe-Kaas, S. P. Nørsett, and A. Zanna, "Lie-group methods," *Acta numerica*, vol. 9, pp. 215–365, 2000.

A. DISCRETE GRADIENT DERIVATIVE

Proof. To prove property 4.1 for $H(x)$ a scalar potential, when the variation $\nu \neq 0$, using a) the quotient rule, b) the chain rule and c) identification with the discrete gradient definition (36), we obtain

$$\begin{aligned}\frac{\partial\overline{\nabla}H}{\partial\nu} &\overset{a}{=} \frac{[\frac{\partial}{\partial\nu}(H(x+\nu)-H(x))]\nu - [H(x+\nu)-H(x)]\frac{\partial\nu}{\partial\nu}}{\nu^2}\\ &\overset{b}{=} \frac{1}{\nu}\left(\frac{\partial H}{\partial x}(x+\nu)\frac{\partial(x+\nu)}{\partial v} - \frac{H(x+\nu)-H(x)}{\nu}\right)\\ &\overset{c}{=} \frac{\nabla H(x+\nu)-\overline{\nabla}H(x,\nu)}{\nu}.\end{aligned}$$

When $\nu \to 0$, using a) the definition of the discrete gradient (36) with b) Taylor series expansion about x and neglecting high order terms when passing to the limit leads to

$$\begin{aligned}\frac{\partial\overline{\nabla}H}{\partial\nu}(x,0) &:= \lim_{\nu\to 0}\frac{\nabla H(x+\nu)-\overline{\nabla}H(x,\nu)}{\nu}\\ &\overset{a}{=} \lim_{\nu\to 0}\frac{\nabla H(x+\nu)}{\nu} - \frac{H(x+\nu)-H(x)}{\nu^2}\\ &\overset{b}{=} \lim_{\nu\to 0}\frac{H'(x)+H''\nu}{\nu} - \frac{H'(x)\nu+H''(x)\nu^2/2!}{\nu^2}\\ &= \frac{1}{2}\frac{\partial^2 H}{\partial x^2}(x)\end{aligned}$$

□

B. DISCRETE GRADIENT OF THE DIODE POTENTIAL

Proof. Using a) the definition of the discrete gradient (36), b) the definition of the diode potential (23) followed by c) factorization of the mid-point exponential term, then d) identification of the $\sinh$ and e) sinhc functions, the discrete gradient of the diode voltage potential can be expressed as

$$\begin{aligned}\overline{\nabla}D^*(v,\delta v) &:\overset{a}{=} \frac{D_D^*(v+\delta v)-D_D^*(v)}{\delta v}\\ &\overset{b}{=} \frac{nV_T I_S}{\delta v}\left(\exp\left(\frac{v+\delta v}{nV_T}\right) - \exp\left(\frac{v}{nV_T}\right) - \frac{\delta v}{nV_T}\right)\\ &\overset{c}{=} I_S\left(\frac{nV_T}{\delta v}\exp\left(\frac{v+\delta v/2}{nV_T}\right)\left(e^{\frac{\delta v}{2nV_T}} - e^{-\frac{\delta v}{2nV_T}}\right) - 1\right)\\ &\overset{d}{=} I_S\left(\frac{2nV_T}{\delta v}\exp\left(\frac{v+\delta v/2}{nV_T}\right)\sinh\left(\frac{\delta v}{2nV_T}\right) - 1\right)\\ &\overset{e}{=} I_S\left(\exp\left(\frac{v+\delta v/2}{nV_T}\right)\mathrm{sinhc}\left(\frac{\delta v}{2nV_T}\right) - 1\right)\end{aligned}$$

and since $\mathrm{sinhc}(0) = 1$, $\overline{\nabla}D^*(v,0) = \nabla D^*(v)$ satisfies eq (33). □

A MINIMAL PASSIVE MODEL OF THE OPERATIONAL AMPLIFIER : APPLICATION TO SALLEN-KEY ANALOG FILTERS

Rémy Müller

IRCAM-STMS (UMR 9912)
Sorbonne University
Paris, France
remy.muller@ircam.fr

Thomas Hélie *

IRCAM-STMS (UMR 9912)
Sorbonne University
Paris, France
thomas.helie@ircam.fr

ABSTRACT

This papers stems from the fact that, whereas there are passive models of transistors and tubes, a minimal passive model of the operational amplifier does not seem to exist. A new behavioural model is presented that is memoryless, fully described by its interaction ports, with a minimal number of equations, for which a passive power balance can be defined. The proposed model handles saturation, asymmetric power supply, and can be used with non-ideal voltage references. To illustrate the model in audio applications, the non-inverting voltage amplifier and a saturating Sallen-Key lowpass filter are considered.

1. INTRODUCTION

Operational Amplifier (OPA) models can be roughly categorized into a) Controlled Source (CS) models, b) white box macro models and c) Nullor models .

In CS models (see [1]), the power supplies are lumped within the OPA and controlled sources can provide an infinite amount of power. It has the advantage of being simple and hides most of the internal complexity. This is the method of choice used by students to study the functional behaviour of OPA circuits. The main drawback comes from the absence of external supply ports. This results in non passive models, and forbids simulations with non-ideal voltage sources (e.g. in low-budget guitar stomboxes).

White box macro models (see references [2] [3] [4]) use dozens of transistors to accurately reproduce the inner structure and non-ideal characteristics of particular devices. While this is appropriate for offline simulation and circuit design, the main drawback of this approach comes from the high number of (implicit) nonlinear equations which makes it often unsuitable for real-time simulation.

Nullors (see references [5] [6] [7] [8]), are singular two-port elements where the input flow and effort variables are both zero: $e_1 = f_1 = 0$, while the output flow and effort variables e_2, f_2 are unconstrained. One drawback is the lack of flow / effort duality. In addition, similar to CS, Nullors have no explicit power supply ports and thus are not passive devices, inheriting the same drawbacks mentioned above.

For audio applications, dedicated Wave Digital Filters (WDF) models of the OPA for specific circuit topologies have been proposed in [9], more recently, using Modified Nodal Analysis to WDF adaptors, both Nullor and CS general purpose models of the OPA and OTA have been proposed in [10] [11] and Sallen-key filters have been modelled with WDF in [12].

We propose a passive, quasi-ideal, black-box, behavioural model of the OPA, simple enough for realtime simulation, with explicit power supply and modelling nonlinear saturation. In particular, a by-product of this research is to have a model compatible with the port-Hamiltonian formalism [13].

The paper is structured as follows. First a general purpose passive model of the OPA is proposed in section 2, then it is illustrated by treating the non-inverting voltage amplifier circuit in section 3, finally a detailed study and simulation of a saturating Sallen-Key lowpass filter is presented in section 4.

2. OPERATIONAL AMPLIFIER MODEL

The objective of this paper is to find the simplest class of Operational Amplifier models satisfying the following properties:

a) Memoryless: infinite bandwidth, infinite slew rate,

b) Passivity: the power dissipated by the OPA is non-negative (i.e. hidden sources of energy are forbidden),

c) Quasi-ideal behaviour: infinite input impedance, zero output impedance, infinite common-mode rejection ratio,

d) Finite output voltage range and saturation: explicit non-constant power-supply ports,

e) Minimal: behavioural model with a minimum number of equations (i.e. not a white box model containing dozen of transistors).

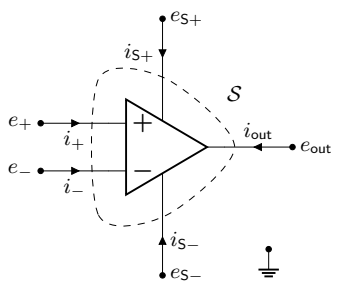


Figure 1: Circuit diagram of an Operational Amplifier (OPA) with currents drawn in receiver convention. The gaussian surface $\mathcal{S}$ enclosing the component is shown in dashed line.

* The author acknowledges the support of the ANR-DFG (French-German) project INFIDHEM ANR-16-CE92-0028.

Copyright: © 2019 Rémy Müller et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 Unported License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

2.1. Notations

The OPA shown on figure 1 is modelled as a 5-port device with node voltages being measured relatively to the ground, node currents directed toward the element using the receiver convention and pins labelled $\mathcal{P} = \{+, -, \mathsf{S}+, \mathsf{S}-, \mathsf{out}\}$. In this paper, we assume that the ports of the OPA can be partitioned into a voltage-driven set $\mathcal{T}$, and a current-controlled co-set $\mathcal{T}^*$

$$\mathcal{T} := \{+, -, \mathsf{S}+, \mathsf{S}-\}, \quad \mathcal{T}^* := \{\mathsf{out}\}, \quad \mathcal{T} \cup \mathcal{T}^* = \mathcal{P}. \quad (1)$$

The respective inputs and outputs are collected into the vectors

$$\mathbf{u} := [\mathbf{e}_{\mathcal{T}}, \mathbf{i}_{\mathcal{T}^*}]^\mathsf{T} = [e_+, e_-, e_{\mathsf{S}+}, e_{\mathsf{S}-}, i_{\mathsf{out}}]^\mathsf{T}, \quad (2)$$

$$\mathbf{y} := [\mathbf{i}_{\mathcal{T}}, \mathbf{e}_{\mathcal{T}^*}]^\mathsf{T} = [i_+, i_-, i_{\mathsf{S}+}, i_{\mathsf{S}-}, e_{\mathsf{out}}]^\mathsf{T}, \quad (3)$$

Finally, the common supply, the differential supply and the differential input voltages are respectively defined by

$$V_{\mathsf{cm}} = \frac{e_{\mathsf{S}+} + e_{\mathsf{S}-}}{2}, \quad V_{\mathsf{dm}} = \frac{e_{\mathsf{S}+} - e_{\mathsf{S}-}}{2}, \quad \epsilon = e_+ - e_-. \quad (4)$$

2.2. Constitutive equations

Since there are 5 ports with dual flow and efforts variables, 5 independent equations are required to specify the device:

1-2) **Non-energetic input ports**: the current entering the pins $\{+, -\}$ is zero (infinite input impedance)

$$i_+ = i_- = 0, \quad (5)$$

3) **Conservation of charge**: Kirchoff Current Law applied over the gaussian surface[1] $\mathcal{S}$ enclosing the AOP implies that the sum of all currents is zero

$$\sum_{\ell \in \mathcal{P}} i_\ell = 0, \quad (6)$$

4) **Passivity**: the power absorbed by the OPA is greater or equal to zero

$$P_{\mathsf{diss}} = \mathbf{y}^\mathsf{T}\mathbf{u} = \sum_{\ell \in \mathcal{P}} e_\ell \cdot i_\ell \geq 0, \quad (7)$$

5) **Differential gain and saturation**: the tensions are tied by a continuous relation $e_{\mathsf{out}} = f(e_+, e_-, e_{\mathsf{S}+}, e_{\mathsf{S}-})$ such that

$$\begin{cases} \dfrac{\partial f}{\partial \epsilon} \geq 0, & \text{monotonicity} \\ \max\left(\dfrac{\partial f}{\partial \epsilon}\right) = K, & \text{differential gain} \\ \max(f) = e_{\mathsf{S}+},\ \epsilon \to +\infty & \text{positive saturation} \\ \min(f) = e_{\mathsf{S}-},\ \epsilon \to -\infty & \text{negative saturation} \end{cases} \quad (8)$$

This gives 4 equalities and 1 inequality

$$i_+ = 0 \quad (9)$$

$$i_- = 0 \quad (10)$$

$$i_{\mathsf{S}+} + i_{\mathsf{S}-} + i_{\mathsf{out}} = 0 \quad (11)$$

$$P_{\mathsf{diss}} = i_{\mathsf{S}+} \cdot e_{\mathsf{S}+} + i_{\mathsf{S}-} \cdot e_{\mathsf{S}-} + i_{\mathsf{out}} \cdot e_{\mathsf{out}} \geq 0 \quad (12)$$

$$f(e_{\mathsf{S}+}, e_{\mathsf{S}-}, e_+, e_-) - e_{\mathsf{out}} = 0 \quad (13)$$

Since there is an inequality and the relation f is not specified yet, there is an infinite class of models satisfying these equations. A particular instance is chosen as follows.

[1] The Gaussian surface $\mathcal{S}$ is shown on figure 1. For more details see [1].

2.3. Toward a unique model

Substituting (4) into the passivity equation (12), using the conservation of charge (11) and simplifying by i_{out} gives the constraint[2]

$$V_{\mathsf{cm}} + V_{\mathsf{dm}} \left(\frac{i_{\mathsf{S}+} - i_{\mathsf{S}-}}{i_{\mathsf{S}+} + i_{\mathsf{S}-}} \right) = e_{\mathsf{out}} - \frac{P_{\mathsf{diss}}}{i_{\mathsf{out}}}, \quad (i_{\mathsf{out}} \neq 0) \quad (14)$$

which imposes a lot of structure on the form of the output function. In order to specify a unique model, the following choices are made.

2.3.1. *Differential input transistor pair*

First, motivated by the typical structure of an OPA, composed of a differential pair of transistors, gain stages and a push-pull output (see [14] p.707), the adimensioned modulation factor [3]

$$\rho(\epsilon) := -\frac{i_{\mathsf{S}+}}{i_{\mathsf{out}}} = \frac{\exp(x)}{\exp(x) + \exp(-x)}, \quad x = \frac{K\epsilon}{V_{\mathsf{dm}}}, \quad (15)$$

is introduced and shown on figure 2. According to the conservation of charge (11), this leads to the symmetric current splitting

$$i_{\mathsf{S}+} = -\rho(\epsilon) i_{\mathsf{out}}, \qquad i_{\mathsf{S}-} = -\rho(-\epsilon) i_{\mathsf{out}}. \quad (16)$$

2.3.2. *The conservative OPA choice*

Second, among all passive OPA models, the conservative ones are chosen, neglecting internal dissipation:

$$P_{\mathsf{diss}} = 0. \quad (17)$$

The power supply ports provide the amount of power necessary to balance the power consumed at the output port. This is an instance of a nonlinear nonenergic n-port [15].

2.3.3. *Final model*

Substituting (16) and (17) into (14) uniquely defines the output function (a similar result was also derived in [16])

$$e_{\mathsf{out}} = V_{\mathsf{cm}} + V_{\mathsf{dm}} \tanh\left(\frac{K\epsilon}{V_{\mathsf{dm}}}\right). \quad (18)$$

Expressed as a function of $e_{\mathsf{S}+}, e_{\mathsf{S}-}$ this gives

$$e_{\mathsf{out}} = \rho(+\epsilon) e_{\mathsf{S}+} + \rho(-\epsilon) e_{\mathsf{S}-}. \quad (19)$$

Finally gathering equations (5) (16) (19) in matrix form reveals the modulated hybrid Dirac structure[4] of the conservative OPA model given by the skew-symmetric matrix $\mathbf{J}(\mathbf{u})$:

$$\underbrace{\begin{bmatrix} i_+ \\ i_- \\ i_{\mathsf{S}+} \\ i_{\mathsf{S}-} \\ e_{\mathsf{out}} \end{bmatrix}}_{\mathbf{y}} = \underbrace{\left[\begin{array}{cc|ccc} \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \hline \cdot & \cdot & \cdot & \cdot & -\rho(+\epsilon) \\ \cdot & \cdot & \cdot & \cdot & -\rho(-\epsilon) \\ \cdot & \cdot & \rho(\epsilon) & \rho(-\epsilon) & \cdot \end{array}\right]}_{\mathbf{J}(\mathbf{u})} \underbrace{\begin{bmatrix} e_+ \\ e_- \\ e_{\mathsf{S}+} \\ e_{\mathsf{S}-} \\ i_{\mathsf{out}} \end{bmatrix}}_{\mathbf{u}}. \quad (20)$$

The singularity of the structure matrix $\mathbf{J}$ encodes the conservation of the so-called Casimir invariants $i_+ = i_- = 0$, in addition to the conservative power-balance

$$P_{\mathsf{diss}} = \mathbf{u}^\mathsf{T}\mathbf{y} = \mathbf{u}^\mathsf{T}\mathbf{J}(\mathbf{u})\mathbf{u} = 0, \quad (\mathbf{J} = -\mathbf{J}^\mathsf{T}). \quad (21)$$

[2] see appendix A for a detailed proof.

[3] Different choices can be made here to adapt to other transistors types.

[4] Please refer to the references [17] [18] [13] for more details on Dirac structures and to [1] for hybrid parameters.

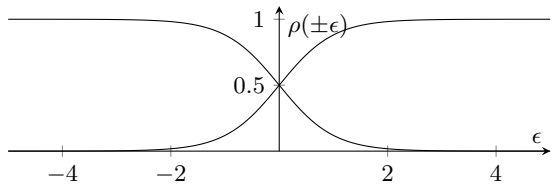
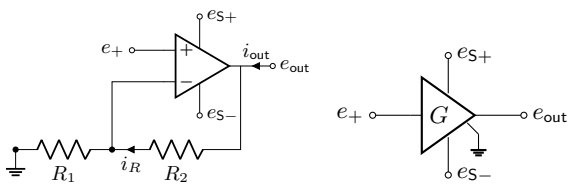


Figure 2: The modulation factor $\rho(\pm\epsilon)$, for $K = 1$, $V_{\mathsf{dm}} = 1$.

3. CASE STUDY

To study the behaviour of the proposed model in practical applications, the case of the voltage amplifier is examined in section 3.1. Then as a pedagogical example, the voltage amplifier is driven by a sinusoidal voltage source and asymmetrically powered by a single capacitor to simulate a discharging battery in section 3.2. The voltage amplifier will be used as a building block of the Sallen-Key lowpass filter shown in section 4.

3.1. The non-inverting voltage amplifier



(a) (b)

Figure 3: a) a non-inverting voltage amplifier circuit with explicit alimentation ports and b) its symbol.

A non-inverting voltage amplifier (figure 3) is achieved by feeding back the output e_{out} to the negative input e_- through a voltage divider

$$\epsilon = e_+ - \frac{e_{\mathsf{out}}}{G}, \qquad G = \frac{R_1 + R_2}{R_1} = 1 + \frac{R_2}{R_1}. \tag{22}$$

The instantaneous feedback makes the circuit act as a proportional corrector with high proportional gain K in order to satisfy the constraint $e_{\mathsf{out}} \approx Ge_+$ within the range $e_{\mathsf{out}} \in [e_{\mathsf{S}+}, e_{\mathsf{S}-}]$.

The voltage divider induces an internal current $i_R = e_{\mathsf{out}}/R$, where $R = R_1 + R_2$, and the current splitting (16) becomes

$$i_{\mathsf{S}+} = -\rho(\epsilon)(i_{\mathsf{out}} - i_R), \quad i_{\mathsf{S}-} = -\rho(-\epsilon)(i_{\mathsf{out}} - i_R). \tag{23}$$

This results in the following law for the voltage amplifier

$$\begin{bmatrix} i_+ \\ i_{\mathsf{S}+} \\ i_{\mathsf{S}-} \\ e_{\mathsf{out}} \end{bmatrix} = \left[\begin{array}{c|cc|c} \cdot & \cdot & \cdot & \cdot \\ \hline \cdot & g_+(\epsilon) & g_\pm(\epsilon) & -\rho(\epsilon) \\ \cdot & g_\pm(\epsilon) & g_-(\epsilon) & -\rho(-\epsilon) \\ \hline \cdot & \rho(\epsilon) & \rho(-\epsilon) & \cdot \end{array}\right] \begin{bmatrix} e_+ \\ e_{\mathsf{S}+} \\ e_{\mathsf{S}-} \\ i_{\mathsf{out}} \end{bmatrix}. \tag{24}$$

with conductances

$$g_+(\epsilon) = \frac{\rho(\epsilon)^2}{R}, \; g_-(\epsilon) = \frac{\rho(-\epsilon)^2}{R}, \; g_\pm(\epsilon) = \frac{\rho(\epsilon)\rho(-\epsilon)}{R}. \tag{25}$$

In the following, it is assumed that $R \to \infty$ such that internal losses are negligible. In particular, this is the case of the classical voltage follower circuit for which $R_2 = 0$, and $R_1 = \infty$.

3.1.1. Implicit constraint

The relation (24) is still implicitly defined since ϵ depends on both input and output variables e_+ and e_{out}. To avoid apparent difficulties with discontinuous functions, consider the curve

$$\mathcal{F} = \left\{(u, y) \in \mathbb{R}^2 \mid F(u, y) = 0\right\}, \tag{26}$$

specified by the function

$$F(u, y) = V_{\mathsf{cm}} + V_{\mathsf{dm}} \tanh\left(\frac{K}{V_{\mathsf{dm}}}\left(u - \frac{y}{G}\right)\right) - y, \tag{27}$$

and given e_+, look for e_{out} such that $(e_+, e_{\mathsf{out}}) \in \mathcal{F}$.

Since the output function is monotonous with respect to ϵ and bounded in $[e_{\mathsf{S}-}, e_{\mathsf{S}+}]$, a unique solution exists within that range. A global method such as the bisection method is guaranteed to find it, whereas, since K is typically about 10^6, it is very difficult to use either fixed-point or derivative-based methods because of bad numerical conditioning. Numerical simulations are shown on figure 4.

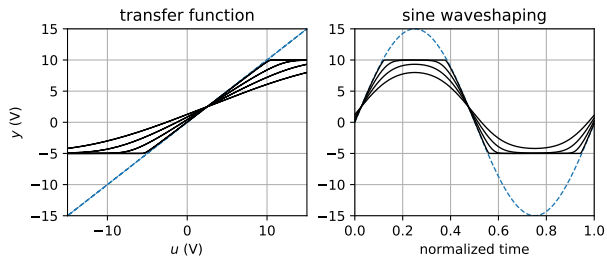


Figure 4: Transfer function of the voltage amplifier for $G = 1$, $K \in \{1, 2, 5, 50\}$, $e_{\mathsf{S}+} = 10$V, $e_{\mathsf{S}-} = -5$V. Smaller values than the typical OPA gain $K \approx 10^6$ are used for visualisation purposes.

3.1.2. Explicit representation

Taking the limit when $K \to \infty$ gives an explicit representation of $\mathcal{F}$ as the piecewise continuous curve

$$\mathcal{F}_\infty = \lim_{K\to\infty} \mathcal{F} : \begin{cases} y = e_{\mathsf{S}+}, & Gu > y \\ y = e_{\mathsf{S}-}, & Gu < y \\ y \in [e_{\mathsf{S}-}, e_{\mathsf{S}+}], & y = Gu \end{cases}. \tag{28}$$

One can see on figure 4 that convergence to $\mathcal{F}_\infty$ is very fast even for moderate values of K. This justifies the use of this limit process in following developments.

For $(e_+, e_{\mathsf{out}}) \in \mathcal{F}_\infty$ this gives the explicit form

$$e_{\mathsf{out}} = V_{\mathsf{cm}} + V_{\mathsf{dm}} \operatorname{sat}\left(\frac{Ge_+ - V_{\mathsf{cm}}}{V_{\mathsf{dm}}}\right), \tag{29}$$

where

$$\operatorname{sat}(x) = \min(\max(x, -1), 1). \tag{30}$$

Alternatively one can represent this function as

$$e_{\mathsf{out}} = \mu_+(e_+, V_{\mathsf{cm}}, V_{\mathsf{dm}})e_{\mathsf{S}+} + \mu_-(e_+, V_{\mathsf{cm}}, V_{\mathsf{dm}})e_{\mathsf{S}-} \tag{31}$$

where the implicit modulation factor $\rho(\pm\epsilon)$ in (24) has been replaced by the explicit one

$$\mu_\pm(e_+, V_{\mathsf{cm}}, V_{\mathsf{dm}}) = \frac{1 \pm \operatorname{sat}(x)}{2}, \quad x = \frac{Ge_+ - V_{\mathsf{cm}}}{V_{\mathsf{dm}}}. \tag{32}$$

3.2. A single-rail voltage follower powered by a capacitor

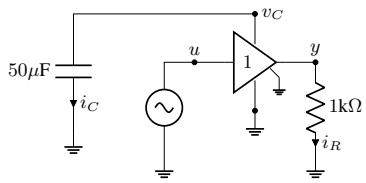


Figure 5: A single-rail voltage amplifier powered by a capacitor.

To illustrate one of the practical interest of having explicit power supply ports, the voltage amplifier is used with the negative supply port grounded, and the positive supply port powered by a capacitor to simulate a discharging battery (figure 5).

Using (20) with $V_{\mathsf{cm}} = V_{\mathsf{dm}} = q/(2C)$, and $i_{\mathsf{out}} = -y/R$, yields the algebro-differential equations

$$\begin{cases} \dot{q} = -\eta(u,q)\dfrac{y}{R}, \\ y = \eta(u,q)\dfrac{q}{C} \end{cases}, \quad \eta(u,q) = \mu_+\left(u, \frac{q}{2C}, \frac{q}{2C}\right). \tag{33}$$

The energy stored in the capacitor is $H(q) = q^2/2C$. Then its differential equation is governed by the monotonous discharge

$$\frac{\mathrm{d}}{\mathrm{d}t}H(q) = \frac{\partial H}{\partial q}\frac{\mathrm{d}q}{\mathrm{d}t} = -\frac{q}{C}\eta(q,u)\frac{y}{R} = -\frac{y^2}{R}. \tag{34}$$

The circuit acts as a half-wave rectifier with a positive clipping threshold governed by the discharge of the capacitor as shown on figure 6.

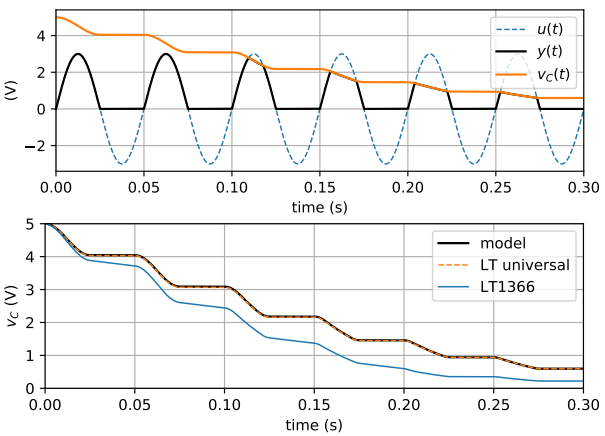


Figure 6: Time domain simulation of the capacitor-powered single rail voltage amplifier with $v_C(0) = 5$V and $|u| = 3V$. Top plot: proposed model. Bottom plot: comparison of discharge rate with LTspice's Universal OPA level.2 and the LT1366 [19].

Remark (Comparison between models)

As expected, with the proposed model, the capacitor does not discharge during negative saturation (energy-preservation), and has a monotonous discharge otherwise. Comparison with LTspice's universal model shows that the two simulations are very close. Finally with the LT1366, the discharge is monotonous and qualitatively similar, but decays faster due to internal dissipation.

4. SALLEN-KEY ANALOG LOWPASS FILTER

The class of Sallen-Key Filters (SKF), introduced in [20], is perhaps one of the most common analog filter design topology. It is used for the realization of analog biquadratic filters, for example in parametric equalisers. It is also the basis of the multimode Steiner filter [21], the Korg MS-20 [22] and the Buchla Lowpass-Gate [23].

A Sallen-Key lowpass filter schematic is shown on figure 8a. The linear regime and its control parameters are studied in 4.1, the circuit is then converted into equations in 4.2. Discretization is performed using the Average Vector Field method in 4.3, finally simulation results are shown in 4.4.

4.1. Linear behaviour and control parameters

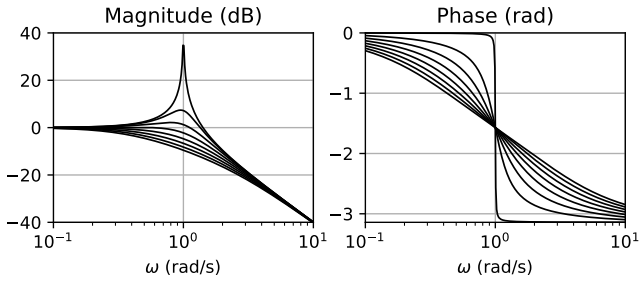


Figure 7: Bode plot of the Sallen-Key filter for $\omega = 1$, $G \in [0, 3]$

It is recalled that the Laplace transfer function (shown on figure 7) of a second order resonant lowpass filters with pulsation ω and quality factor Q is

$$H_{\mathsf{LP}}(s) = \frac{1}{1 + \frac{1}{Q}\left(\frac{s}{\omega}\right) + \left(\frac{s}{\omega}\right)^2}, \tag{35}$$

In the linear regime, the Laplace transfer function of the lowpass Sallen-Key filter is

$$H_{\mathsf{SK}}(s) = \mathcal{L}\left\{\frac{y_{\mathsf{SK}}}{v_{\mathsf{IN}}}\right\} = \frac{1}{1 + a_1 s + a_2 s^2}, \tag{36}$$

where

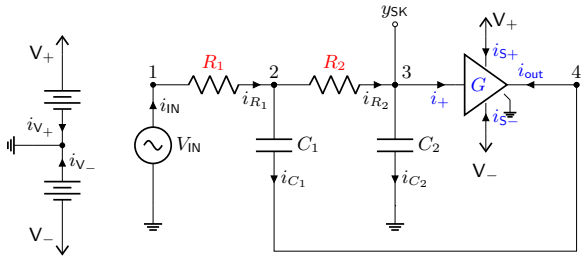
$$a_1 = \left((1-G)R_1C_1 + (R_1+R_2)C_2\right), \tag{37}$$
$$a_2 = C_1C_2R_1R_2. \tag{38}$$

Since there are only two target controls (ω, Q), for 5 design parameters (R_1, R_2, C_1, C_2, G), there are many possible design decisions that are often decided according to electronic constraints.

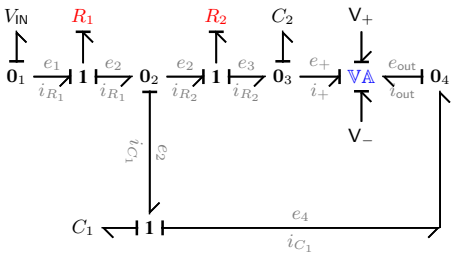
In this paper, the Steiner filter parametrization is used with $R_1 = R_2 = R$, and $C_1 = C_2 = C$ because of its simplicity. The transfer function (36) simplifies to

$$H_{\mathsf{SK}}(s) = \frac{1}{1 + (3-G)\left(\frac{s}{\omega}\right) + \left(\frac{s}{\omega}\right)^2}, \tag{39}$$

with $\omega = 1/(RC)$, and $Q = 1/(3-G)$. In simulations, capacitances are both set to $C = 4.7$nF and the resistors are adjusted to achieve the target cutoff frequencies.



(a) Circuit



(b) Bondgraph

$$
\begin{array}{c}
 \\ i_{\mathsf{IN}} \\ i_{C_1} \\ i_{C_2} \\ i_{\mathsf{V}_+} \\ i_{\mathsf{V}_-} \\ v_{R_1} \\ v_{R_2} \\ e_+ \\ e_{\mathsf{S}+} \\ e_{\mathsf{S}-} \\ i_{\mathsf{out}}
\end{array}
\begin{array}{c}
\begin{array}{ccccccccccc}
V_{\mathsf{IN}} & v_{C_1} & v_{C_2} & \mathsf{V}_+ & \mathsf{V}_- & i_{R_1} & i_{R_2} & i_+ & i_{\mathsf{S}+} & i_{\mathsf{S}-} & e_{\mathsf{out}}
\end{array} \\
\begin{pmatrix}
 & & & & & -1 & . & . & . & . & \\
 & & & & & 1 & -1 & . & . & . & \\
 & & & & & . & 1 & -1 & . & . & \\
 & & & & & . & . & . & -1 & . & \\
 & & & & & . & . & . & . & -1 & \\
1 & -1 & . & . & . & & & & & & -1 \\
. & 1 & -1 & . & . & & & & & & 1 \\
. & . & 1 & . & . & & & & & & . \\
. & . & . & 1 & . & & & & & & . \\
. & . & . & . & 1 & & & & & & . \\
 & & & & & 1 & -1 & . & . & . &
\end{pmatrix}
\end{array}
$$

(c) Skew-symmetric Dirac structure (KCL+KVL)

$$
\begin{pmatrix} i_{C_1} \\ i_{C_2} \\ \hline i_{\mathsf{IN}} \\ i_{\mathsf{out}} \end{pmatrix}
= \underbrace{\frac{1}{R}\left(\begin{array}{cc|cc} -2 & 1 & 1 & -2 \\ 1 & -1 & 0 & 1 \\ \hline 1 & 0 & -1 & 1 \\ -2 & 1 & 1 & -2 \end{array}\right)}_{Y}
\begin{pmatrix} v_{C_1} \\ v_{C_2} \\ \hline v_{\mathsf{IN}} \\ e_{\mathsf{out}} \end{pmatrix}
$$

(d) Reduced model (admittance form)

Figure 8: a) The original Sallen-Key lowpass filter circuit, b) its corresponding bondgraph (see references [24] [25] [26]) with computational causality assignment. c) the skew-symmetric Dirac structure representing Kirchoff conservation laws. d) the reduced dynamical model.

4.2. Modelling

To model the Sallen-Key filter, the following systematic approach is used:

- **Bondgraph**: The circuit 8a is first converted to an equivalent bondgraph 8b using the rules in [25]. A bond between two ports $\mathbf{A} \longrightarrow \mathbf{B}$ stands for a pair of dual port-variables (e, f). The half-arrow indicates the power sign convention $P = ef \geq 0$. **0** denotes a parallel junction where all bonds share the same voltage, and **1** denotes a serial junction where all bonds share the same current.

- **Causality assignment**: to convert an acausal bidirectional bondgraph to a causal, computable, block-diagram, one needs to partition the flows and efforts into inputs and outputs. The convention uses a vertical stroke $\mathbf{A} \longmapsto \mathbf{B}$ next to ports that are effort-controlled. *Computational causalities* can be assigned graphically by propagating the following rules: voltage sources and capacitors have an effort-out causality, **0** junctions can only have one input effort, while the dual **1** junctions can only have one output effort.

- **Dirac Structure**: given the causality assignment, shown on 8b, into inputs and outputs, it is now straightforward to fill the Dirac Structure matrix 8c by inspecting circuit 8a and expressing Kirchoff's current and voltage laws.

- **Reduced model**: one can reduce the model by solving trivial equalities like $e_+ = v_{C_2}$, $e_{\mathsf{S}+} = V_+$, $e_{\mathsf{S}-} = V_-$, treating $V_\pm$ as constants and replacing the linear resistive currents (i_{R_1}, i_{R_2}) by their constitutive laws. This results in the reduced admittance model shown on figure 8d.

4.2.1. *Nonlinear feedback*

To separate the linear and nonlinear feedback, one can write

$$\hat{e}_{\mathsf{out}}(v) = Gv - \nabla N(v) \tag{40}$$

where the nonlinear law is

$$\begin{aligned} \nabla N(v) &:= Gv - \hat{e}_{\mathsf{out}}(v) \\ &= \min(0, Gv - e_{\mathsf{S}-}) + \max(0, Gv - e_{\mathsf{S}+}). \end{aligned} \tag{41}$$

and its algebraic potential (figure 9) is given by the line integral

$$\begin{aligned} N(v) &:= \int_0^v \nabla N(s) \cdot \mathrm{d}s \\ &= \frac{\min(0, Gv - e_{\mathsf{S}-})^2}{2G} + \frac{\max(0, Gv - e_{\mathsf{S}+})^2}{2G}. \end{aligned} \tag{42}$$

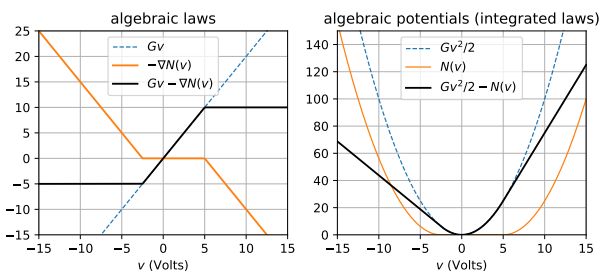


Figure 9: Algebraic feedback laws and their potentials shown for $G = 2$, $e_{\mathsf{S}+} = 10\text{V}$, $e_{\mathsf{S}-} = -5\text{V}$.

4.2.2. State-space model

Finally replacing the flow and effort variables by their constitutive laws, and only considering the input-state-output, one gets

$$\begin{cases} \dot{\mathbf{x}} = \omega \left[\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} - \mathbf{F}\nabla N(\mathbf{C}\mathbf{x}) \right] \\ \mathbf{y} = \mathbf{C}\mathbf{x} \end{cases}, \tag{43}$$

where $\mathbf{u} = v_{\mathsf{IN}}$, $\mathbf{y} = y_{\mathsf{SK}}$, $\mathbf{x} = [v_{C_1}, v_{C_2}]^{\mathsf{T}}$, $\omega = 1/(RC)$ and

$$\mathbf{A} = \begin{bmatrix} -2 & 1-2G \\ 1 & -1+G \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \tag{44}$$

$$\mathbf{C} = \begin{bmatrix} 0 & 1 \end{bmatrix}, \qquad \mathbf{F} = \begin{bmatrix} -2 \\ 1 \end{bmatrix}. \tag{45}$$

Using the co-energy variables v_{C_1}, v_{C_2} instead of the energy variables q_{C_1}, q_{C_2} is justified here by the fact that the capacitors are linear and time-invariant, i.e. the co-energy $H^*(v) = Cv^2/2$ equals the energy $H(q) = q^2/(2C)$ for the linear law $v = q/C$.

4.3. Discretization using the AVF method

The Average Vector Field (AVF) method is used to discretize (43) because of its structure-preserving properties: it preserves the energy (resp. dissipativity) of conservative (resp. dissipative) systems (see [27]). One can also refer to [28] where it has been shown that the bilinear transform doesn't always guarantee the dissipativity of nonlinear filters (whether time-varying or not).

As an important side-effect, the AVF method can also be interpreted as a first-order instance of anti-derivative antialiasing [29].

4.3.1. The Average Vector Field method

Let $\Omega = [t_0, t_0 + h]$ be a time-step, $\mathbf{x} : \Omega \to \mathbb{R}^n$ a locally affine trajectory parametrized by the normalized variable $\tau \in [0, 1]$

$$\mathbf{x}(t_0 + h\tau) = \mathbf{x}_0 + \tau(\mathbf{x}_1 - \mathbf{x}_0). \tag{46}$$

Introduce the averaging operator $\mathcal{A}$, defined for all functions $f : \mathbb{R}^n \to \mathbb{R}^n$ or operators $f : \mathcal{H} \to \mathcal{H}$, where $\mathcal{H}$ is a functional space from $\Omega \to \mathbb{R}^n$, by

$$(\mathcal{A}\, f)(\mathbf{x}) := \int_0^1 f(\mathbf{x}(t_0 + h\tau))\, \mathrm{d}\tau. \tag{47}$$

For the time derivative and identity operators, one gets

$$\overline{\dot{\mathbf{x}}} := \left(\mathcal{A}\tfrac{\mathrm{d}}{\mathrm{d}t}\right)\mathbf{x} = \frac{\mathbf{x}_1 - \mathbf{x}_0}{h}, \quad \bar{\mathbf{x}} := (\mathcal{A}\mathcal{I})\mathbf{x} = \frac{\mathbf{x}_0 + \mathbf{x}_1}{2}. \tag{48}$$

Using the gradient theorem, this gives the average discrete gradient

$$\begin{aligned} \overline{\nabla} N(v_0, v_1) &:= (\mathcal{A}\nabla N)(v_0 + \tau(v_1 - v_0)) \\ &= \begin{cases} \dfrac{N(v_1) - N(v_0)}{v_1 - v_0} & v_0 \neq v_1 \\ \nabla N(v_0) & v_0 = v_1 \end{cases}. \end{aligned} \tag{49}$$

Computing its derivative with respect to v_1 leads to

$$\frac{\partial \overline{\nabla} N}{\partial v_1}(v_0, v_1) = \begin{cases} \dfrac{\nabla N(v_1) - \overline{\nabla} N(v_0, v_1)}{v_1 - v_0} & v_0 \neq v_1 \\ \dfrac{1}{2}\nabla^2 N(v_0) & v_0 = v_1 \end{cases}. \tag{50}$$

One can refer to [30], where the discrete gradient's derivative is also used for numerical simulation.

4.3.2. Averaged system

Applying the averaging operator $\mathcal{A}$ to (43), leads to the structure-preserving discrete algebraic system

$$\begin{cases} \overline{\dot{\mathbf{x}}} = \omega \left[\mathbf{A}\bar{\mathbf{x}} + \mathbf{B}\bar{\mathbf{u}} - \mathbf{F}\,\overline{\nabla} N(\mathbf{C}\mathbf{x}_0, \mathbf{C}\mathbf{x}_1) \right] \\ \bar{\mathbf{y}} = \mathbf{C}\bar{\mathbf{x}} \end{cases}. \tag{51}$$

Solving the linear part for $\mathbf{x}_1$ gives the discrete state-space update

$$\mathbf{x}_1 = \mathbf{A}_d\mathbf{x}_0 + \mathbf{B}_d\bar{\mathbf{u}} - \mathbf{F}_d\,\overline{\nabla} N(\mathbf{C}\mathbf{x}_0, \mathbf{C}\mathbf{x}_1), \tag{52}$$

with the normalised pulsation $\omega_d = h\omega$ and

$$\begin{aligned} \mathbf{A}_d &= \mathbf{D}^{-1}\left(\mathbf{I} + \frac{\omega_d}{2}\mathbf{A}\right), & \mathbf{B}_d &= \mathbf{D}^{-1}(\omega_d\mathbf{B}), \\ \mathbf{D} &= \left(\mathbf{I} - \frac{\omega_d}{2}\mathbf{A}\right), & \mathbf{F}_d &= \mathbf{D}^{-1}(\omega_d\mathbf{F}). \end{aligned} \tag{53}$$

4.4. Simulation

Simulation results[5] are shown on figures 10 and 11 and exhibit a very close match with offline simulations performed in LTspice. To solve (52), one can either use the simple fixed-point iteration, or Newton's method.

4.4.1. Fixed-point iteration

A simple numerical scheme is to look for the fixed-point $\mathbf{x}_1 = \phi(\mathbf{x}_1)$ of the pre-conditioned fixed-point function

$$\phi(\mathbf{x}_1) := \mathbf{A}_d\mathbf{x}_0 + \mathbf{B}_d\bar{\mathbf{u}} - \mathbf{F}_d\,\overline{\nabla} N(\mathbf{C}\mathbf{x}_0, \mathbf{C}\mathbf{x}_1), \tag{54}$$

with the fixed-point iteration

$$\mathbf{x}_1^{k+1} = \phi\left(\mathbf{x}_1^k\right), \quad \mathbf{x}_1^0 = \mathbf{x}_0. \tag{55}$$

A sufficient convergence condition is detailed in appendix B.

In practice, thanks to the non linear feedback splitting in (40), when the OPA is in the linear regime, $\nabla N = 0$. Then the iteration reduces to an explicit one-step trapezoidal integrator and converges in only one iteration.

4.4.2. Newton iteration

To accelerate convergence, one can use Newton's method [31] as follows: define the auxiliary function

$$\varphi(\mathbf{x}_1) = \mathbf{x}_1 - \phi(\mathbf{x}_1), \tag{56}$$

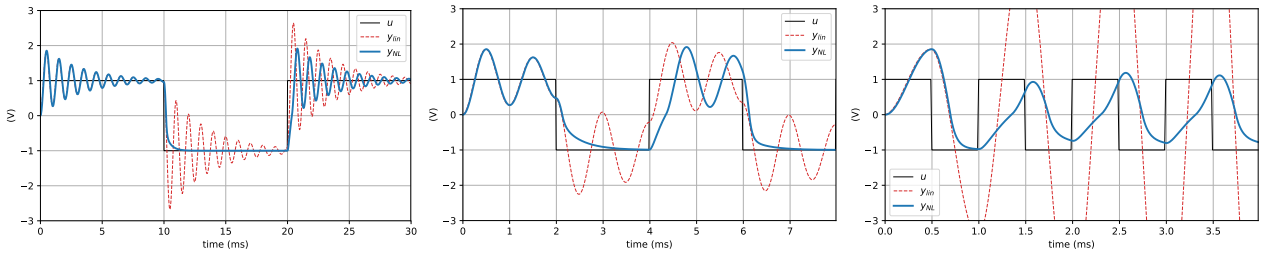
and look for the root $\mathbf{x}_1^*$ such that $\varphi(\mathbf{x}_1^*) = 0$ with the Newton iteration

$$\mathbf{x}_1^{k+1} = \mathbf{x}_1^k - \left(\varphi'(\mathbf{x}_1^k)\right)^{-1}\varphi(\mathbf{x}_1^k), \quad \mathbf{x}_1^0 = \mathbf{x}_0. \tag{57}$$

where the Jacobian of φ is given by

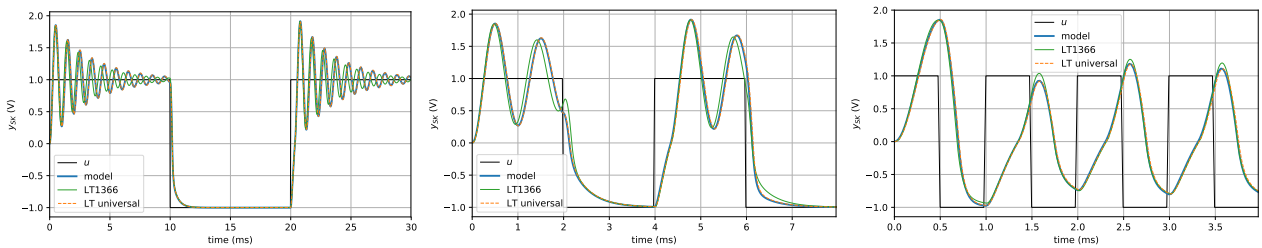
$$\varphi'(\mathbf{x}_1) = \mathbf{I} + \mathbf{F}_d\mathbf{C}\,\frac{\partial \overline{\nabla} N}{\partial v_1}(\mathbf{C}\mathbf{x}_0, \mathbf{C}\mathbf{x}_1). \tag{58}$$

[5] Sound examples and LTspice files are available at the accompanying website: `https://github.com/remymuller/dafx19-opa`.



(a) $f_0 = 50$Hz (b) $f_0 = 250$Hz (c) $f_0 = 1$kHz

Figure 10: SKF filter response to a square wave input with sampling frequency $f_s = 44.1$kHz, $C = 4.7$nF, cutoff $f_c = 1$kHz ($R = 33.8k\Omega$), $Q = 10$, asymmetric saturation $V_+ = 15$V, $V_- = 0$V and different fundamental frequencies. The non linear SKF response is shown in solid blue, with the linear SKF response in dashed red for reference.



(a) $f_0 = 50$Hz (b) $f_0 = 250$Hz (c) $f_0 = 1$kHz

Figure 11: Comparison between the proposed model, LTspice's universal OPA level.2 and the LT1366 opamp. The proposed model output is almost indistinguishable from LTspice's universal model, whereas the tuning of the LT1366 is slightly different because of dissipation.

5. CONCLUSIONS AND PERSPECTIVES

In this paper, a static, passive, black-box model of the operational amplifier with explicit power supply has been examined. It is suitable for the modelling of audio circuits and simple enough for real-time simulation. Furthermore the explicit modelling of external power supply ports allows the use of non-ideal voltage sources.

The choice has been made to ignore internal dissipation to keep the model minimal. However, non-ideal characteristics such as input and output impedance or power supply voltage drop can be achieved by modular composition of the model with other circuit elements. This will be the topic of further research.

The non inverting amplifier is also derived as a dedicated building block. Numerical simulations justify the use of an infinite OPA gain to get an explicit formulation. Having a pre-solved amplifier model also greatly simplifies its use in electronic circuits, avoiding numerical stiffness and high index DAE.

Finally, the amplifier is used for audio simulations to model a saturating Sallen-Key lowpass filter of second order. A reduced state-space model is derived from the circuit schematic, and a structure-preserving discretization is performed using the average vector field method. A comparison with LTspice shows that our results are very close to those of more complex macro models.

The perspectives of this study are a) modelling other non-ideal OPA characteristics such as finite slew-rate and bandwidth, current and voltage offsets, non-zero common-mode input gain... b) studying the behaviour of the model in other typical circuits (oscillator, rectifier, comparator) and c) experimental comparison with specific devices such as the common μA741, or TL072 audio OPAs.

6. REFERENCES

[1] L. O. Chua, C. A. Desoer, and E. S. Kuh, *Linear and nonlinear circuits*. McGraw-Hill College, 1987.

[2] G. R. Boyle, D. Pederson, B. Cohn, and J. E. Solomon, "Macromodeling of integrated circuit operational amplifiers," *IEEE Journal of Solid-State Circuits*, vol. 9, no. 6, pp. 353–364, 1974.

[3] B. Carter and T. R. Brown, *Handbook of operational amplifier applications*. Texas Instruments Dallas, Tex, USA, 2001.

[4] M. Alexander and D. F. Bowers, "Spice-compatible op amp macro-models," *Analog Devices, Application Note, AN-138*, 1990.

[5] H. Carlin, "Singular network elements," *IEEE Transactions on circuit theory*, vol. 11, no. 1, pp. 67–72, 1964.

[6] B. Tellegen, "On nullators and norators," *IEEE Transactions on circuit theory*, vol. 13, no. 4, pp. 466–469, 1966.

[7] L. Odess and H. Ur, "Nullor equivalent networks of non-ideal operational amplifiers and voltage-controlled sources," *IEEE Transactions on Circuits and Systems*, vol. 27, no. 3, pp. 231–235, 1980.

[8] G. Martinelli, "On the nullor," *Proceedings of the IEEE*, vol. 53, no. 3, pp. 332–332, 1965.

[9] R. C. Paiva, S. D'Angelo, J. Pakarinen, and V. Valimaki, "Emulation of operational amplifiers and diodes in audio distortion circuits," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 59, no. 10, pp. 688–692, 2012.

[10] K. J. Werner, W. R. Dunkel, M. Rest, M. J. Olsen, and J. O. Smith, "Wave digital filter modeling of circuits with operational amplifiers," in *2016 24th European Signal Processing Conference (EUSIPCO)*, pp. 1033–1037, IEEE, 2016.

[11] Ó. Bogason and K. J. Werner, "Modeling circuits with operational transconductance amplifiers using wave digital filters," in *Proc. 20th Int. Conf. Digital Audio Effects, Edinburgh, UK*, pp. 130–137, 2017.

[12] M. Verasani, A. Bernardini, and A. Sarti, "Modeling sallen-key audio filters in the wave digital domain," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 431–435, IEEE, 2017.

[13] A. Schaft, "Port-hamiltonian systems: an introductory survey," 2006.

[14] A. S. Sedra and K. C. Smith, *Microelectronic circuits*. New York: Oxford University Press, 1998.

[15] J. L. Wyatt and L. Chua, "A theory of nonenergic n-ports," *International Journal of Circuit Theory and Applications*, vol. 5, no. 2, pp. 181–208, 1977.

[16] J. Mačák, *Real-time digital simulation of guitar amplifiers as audio effects*. PhD thesis, Brno University of Technology, Brno, 2012.

[17] T. J. Courant, "Dirac manifolds," *Transactions of the American Mathematical Society*, vol. 319, no. 2, pp. 631–661, 1990.

[18] A. J. Van Der Schaft, *L2-gain and passivity techniques in nonlinear control*, vol. 3. Springer, 2000.

[19] A. Devices, "Lt1366 datasheet." `https://www.analog.com/en/products/lt1366.html`, 2019. Online; accessed: 2019-03-22.

[20] R. P. Sallen and E. L. Key, "A practical method of designing rc active filters," *IRE Transactions on Circuit Theory*, vol. 2, no. 1, pp. 74–85, 1955.

[21] N.Steiner, "Voltage-tunable active filter features, low, high and bandpass modes," in *Electronic design 25, December 6*, 1974.

[22] T. Stinchcombe, "A study of the korg ms10 & ms20 filters." `http://www.timstinchcombe.co.uk/synth/MS20_study.pdf`, 2006. Online; accessed: 2019-03-22.

[23] J. Parker and S. D'Angelo, "A digital model of the buchla lowpass-gate," in *Proc. Int. Conf. Digital Audio Effects (DAFx-13), Maynooth, Ireland*, pp. 278–285, 2013.

[24] H. M. Paynter, *Analysis and design of engineering systems*. MIT press, 1961.

[25] P. C. Breedveld, "A systematic method to derive bond graph models," in *Second European Simulation Congress, Antwerp, Belgium*, 1986.

[26] J. F. Broenink, "Introduction to physical systems modelling with bond graphs," *SiE whitebook on simulation methodologies*, vol. 31, 1999.

[27] E. Celledoni, V. Grimm, R. I. McLachlan, D. McLaren, D. O'Neale, B. Owren, and G. Quispel, "Preserving energy resp. dissipation in numerical pdes using the average vector field method," *Journal of Computational Physics*, vol. 231, no. 20, pp. 6770–6789, 2012.

[28] T. Hélie, "Lyapunov stability analysis of the moog ladder filter and dissipativity aspects in numerical solutions," in *Proceedings of the 14th International Conference on Digital Audio Effects DAFx-11, Paris, France*, pp. 19–23, 2011.

[29] S. Bilbao, F. Esqueda, J. D. Parker, and V. Välimäki, "Antiderivative antialiasing for memoryless nonlinearities," *IEEE Signal Processing Letters*, vol. 24, no. 7, pp. 1049–1053, 2017.

[30] R. Muller and T. Hélie, "Power-balanced modelling of circuits as skew gradient systems," in *21 st International Conference on Digital Audio Effects (DAFx-18)*, 2018.

[31] P. Deuflhard, *Newton methods for nonlinear problems: affine invariance and adaptive algorithms*, vol. 35. Springer Science & Business Media, 2011.

A. STRUCTURE OF THE OUTPUT EQUATION

Using the passivity equation (12), then introducing $V_{\mathsf{cm}}, V_{\mathsf{dm}}$ using (4), factoring $V_{\mathsf{cm}}, V_{\mathsf{dm}}$, finally, for $i_{\mathsf{out}} \neq 0$, dividing by i_{out} and using (11) one gets the general form for the output equation (14).

Proof.

$$\begin{aligned} & i_{\mathsf{S}+} \cdot e_{\mathsf{S}+} + i_{\mathsf{S}-} \cdot e_{\mathsf{S}-} = -i_{\mathsf{out}} \cdot e_{\mathsf{out}} - P_{\mathsf{diss}} \\ \iff\ & i_{\mathsf{S}+}(V_{\mathsf{cm}} + V_{\mathsf{dm}}) + i_{\mathsf{S}-}(V_{\mathsf{cm}} - V_{\mathsf{dm}}) = -i_{\mathsf{out}} \cdot e_{\mathsf{out}} - P_{\mathsf{diss}} \\ \iff\ & V_{\mathsf{cm}}(i_{\mathsf{S}+} + i_{\mathsf{S}-}) + V_{\mathsf{dm}}(i_{\mathsf{S}+} - i_{\mathsf{S}-}) = -i_{\mathsf{out}} \cdot e_{\mathsf{out}} - P_{\mathsf{diss}} \\ \overset{i_{\mathsf{out}} \neq 0}{\iff}\ & V_{\mathsf{cm}} + V_{\mathsf{dm}} \left(\frac{i_{\mathsf{S}+} - i_{\mathsf{S}-}}{i_{\mathsf{S}+} + i_{\mathsf{S}-}} \right) = e_{\mathsf{out}} - \frac{P_{\mathsf{diss}}}{i_{\mathsf{out}}}. \end{aligned}$$

□

B. FIXED-POINT CONVERGENCE

According to the Banach fixed-point theorem, existence and uniqueness of the solution are guaranteed if the fixed point (55) is contracting, i.e. there exists a Lipschitz constant $\alpha \in [0, 1)$ such that

$$\left\| \phi(\mathbf{x}_1) - \phi(\mathbf{x}_0) \right\| \leq \alpha \|\mathbf{x}_1 - \mathbf{x}_0\| \,. \tag{59}$$

A sufficient (but conservative) condition is given by

$$\alpha = 1.162\, G\omega_d < 1. \tag{60}$$

Proof. Using (54), then the derivative of the discrete gradient (50), (bounded by $G/2$), and using the matrix norm of $\mathbf{F}_d\mathbf{C}$, one gets

$$\begin{aligned} \left\| \phi(\mathbf{x}_1) - \phi(\mathbf{x}_0) \right\|_2 &= \left\| \mathbf{F}_d \left(\overline{\nabla} N(\mathbf{C}\mathbf{x}_0, \mathbf{C}\mathbf{x}_1) - \nabla N(\mathbf{C}\mathbf{x}_0) \right) \right\|_2 \\ &\leq \left\| \mathbf{F}_d \frac{\partial \overline{\nabla} N}{\partial v_1} \mathbf{C} \right\|_2 \|\mathbf{x}_1 - \mathbf{x}_0\|_2 \\ &\leq \|\mathbf{F}_d\mathbf{C}\|_2 \sup_{v_1} \left| \frac{\partial \overline{\nabla} N}{\partial v_1} (v_0, v_1) \right| \|\mathbf{x}_1 - \mathbf{x}_0\|_2 \\ &\leq \frac{2\omega_d \sqrt{\omega_d^2 + 8\omega_d + 20}}{\left|\omega_d^2 + 2(3 - G)\omega_d + 4\right|} \frac{G}{2} \|\mathbf{x}_1 - \mathbf{x}_0\|_2 \\ &\leq 1.162\, G\omega_d \|\mathbf{x}_1 - \mathbf{x}_0\|_2 \end{aligned}$$

where the bound 1.162 is obtained numerically by majorizing over $G \in [0, 3]$ and $\omega_d \geq 0$. □

FULLY-IMPLICIT ALGEBRO-DIFFERENTIAL PARAMETRIZATION OF CIRCUITS

Rémy Müller and Thomas Hélie

IRCAM-STMS (UMR 9912)
Sorbonne University
Paris, France
remy.muller@ircam.fr

ABSTRACT

This paper is concerned with the conception of methods tailored for the numerical simulation of power-balanced systems that are well-posed but implicitly described. The motivation is threefold: some electronic components (such as the ideal diode) can only be implicitly described, arbitrary connection of components can lead to implicit topological constraints, finally stable discretization schemes also lead to implicit algebraic equations.

In this paper we start from the representation of circuits using a power-balanced Kirchhoff-Dirac structure, electronic components are described by a local state that is observed through a pair of power-conjugated algebro-differential operators (V, I) to yield the branch voltages and currents, the arc length is used to parametrize switching and non-Lipschitz components, and a power balanced functional time-discretization is proposed. Finally, the method is illustrated on two simple but non-trivial examples.

1. INTRODUCTION

Network analysis of circuits and expression of Kirchhoff laws, naturally leads to implicit differential algebraic equations (DAE). Indeed in the most general form, the branch equations are not described by functions but by *relations* (in the voltage-current plane for algebraic components, voltage-charge for capacitor, current-flux for inductors ...). One of the most general approach is the Sparse Tableau analysis [1] which involves both the nodes and branch variables.

In the study of power-balanced systems, and more generally in the field of geometrical numerical integration, one is not only concerned with the quantitative accuracy of numerical simulations, but also with the qualitative preservation of structural invariants during discretization [2]. It has been shown that the symplectic structure of Hamiltonian systems, responsible for energy preservation, can be generalized to open systems with algebraic constraints by the notion of a Dirac structure [3] [4]. It can even be extended to infinite-dimensional systems such as partial differential equations using a Stokes-Dirac [5] structure. It has been shown in [6] (see also [7] [8]) that Kirchhoff laws generates a Kirchhoff-Dirac structure. Recent work [9] also study the properties and numerical discretization of Port-Hamiltonian DAE systems in descriptor form.

Usually, when possible, DAE are reduced to ordinary differential equations (ODE) or semi-explicit index-1 DAE [10] [8] for which a rich literature of results from system theory and numerical analysis is available to study stability, conservation laws, attraction points, existence and uniqueness of solutions ...

In these reduction processes, a choice has to be made regarding the variables the system is solved for. Choosing the node voltages leads to the Nodal Analysis (NA) method. But it is not sufficient to represent all systems, adding some branch currents leads to the popular Modified Nodal Analysis (MNA) [11]. The importance of state variable choices for computable numerical simulations can be found in [12]. Similar issues are addressed for wave digital filters in [13]. A procedure to guide these choices is the Sequential Causality Assignment Procedure (SCAP) in the bond-graph literature [14]. In the case of switching-circuits, such as those containing ideal diodes or discontinuous laws (see [15]) an approach is to solve for different variables according to the switching state of the system, but the number of such states becomes exponential in the number of switching components.

Since after time discretization, one is left with an algebraic system of (nonlinear) equations which has to be solved by an iterative scheme anyway, the goal of this article, is to propose a structure-preserving power-balanced numerical method capable of dealing with the implicit nature of the network equations.

Section 2 recalls how any electronic circuit can be represented by a Kirchhoff-Dirac structure, uniquely determined by the circuit's incidence matrix. Section 3 describes how to parametrize the (possibly implicit) relation imposed by any circuit component. Power-conjugated voltages and currents (v, i) are obtained by the application of a pair of nonlinear algebro-differential operators (V, I) to a parameter $\boldsymbol{x}$ which stands for the component's local state. In Section 4 arc-length and pseudo arc-length parameterizations[1] are proposed to overcome computational causality problems that arise in switching components and reduce numerical stiffness caused by high Lipschitz constants. In Section 5 a power-balanced and structure preserving time-discretization is presented using a functional framework. This leads to a nonlinear system of algebraic equations which is solved using Newton iteration. Finally two tests circuits are studied in Section 6, a stiff switching diode clipper and a conservative (nonlinear) LCLC circuit with an implicit topological constraint.

2. KIRCHHOFF-DIRAC STRUCTURES FOR CIRCUIT GRAPHS

From a circuit theory perspective, a *Dirac structure* is simply a multi-port that doesn't generate or dissipate power i.e.

$$P = \langle \mathbf{i} \,|\, \mathbf{v} \rangle = 0.$$

Considering components and their interconnections separately, because of Kirchhoff laws, the multi-port connecting all components

[1] Curvilinear coordinates for multi-ports are possible but not addressed.

Copyright: © 2020 Rémy Müller et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 Unported License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

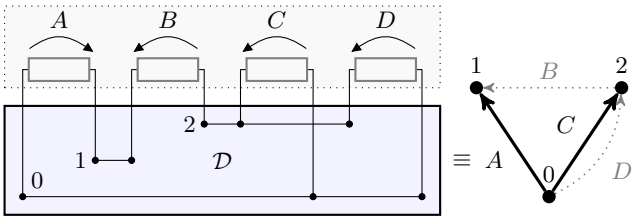


Figure 1: *Dirac structure example with edges* $\mathcal{E} = \{A, B, C, D\}$, *nodes* $\mathcal{N} = \{0, 1, 2\}$ *and chosen spanning tree* $T = \{A, C\}$.

(the PCB) is necessarily a Dirac structure. To formalize it for circuits, we borrow and slightly adapt the notations from [6] [5] [9].

2.1. Circuit Graphs

A directed circuit graph $\mathcal{G}(\mathcal{N}, \mathcal{E})$ is defined by a set of n nodes $\mathcal{N} = \{\eta_1, \dots, \eta_n\}$ and a set $\mathcal{E} = \{\epsilon_1, \dots, \epsilon_m\}$ of m directed edges (links, branches) with no self-loops. Edges are ordered pairs of nodes $\epsilon_i = (\eta_{i,0}, \eta_{i,1})$. Over each node ($k = 0$) and edge ($k = 1$)[2], *using the receiver convention for both*, we define conjugated current and voltages

$$(\mathbf{i}_k, \mathbf{v}_k) \in \mathcal{I}_k \times \mathcal{V}_k =: \mathcal{B}_k, \quad k \in \{0, 1\} \tag{1}$$

where $\mathcal{V}_0 \sim \mathbb{R}^n$, $\mathcal{V}_1 \sim \mathbb{R}^m$ are the spaces of voltages over the nodes $\mathcal{N}$ (resp. the edges $\mathcal{E}$) and $\mathcal{I}_0 = \mathcal{V}_0^*$, $\mathcal{I}_1 = \mathcal{V}_1^*$ the dual spaces of functionals $\mathcal{V}_0 \to \mathbb{R}$, $\mathcal{V}_1 \to \mathbb{R}$. The spaces $\mathcal{B}_0$ and $\mathcal{B}_1$ are respectively the spaces of *bonds* corresponding to the nodes and edges such that power is given by the duality pairings

$$\langle \mathbf{i}_k \,|\, \mathbf{v}_k \rangle_{\mathcal{B}_k} := \mathbf{i}_k^\mathsf{T} \mathbf{v}_k, \qquad k \in \{0, 1\}. \tag{2}$$

Note that since the spaces are finite-dimensional, one can identify each space with its dual $\mathcal{V}_0 \sim \mathcal{I}_0 = \mathbb{R}^n$, $\mathcal{V}_1 \sim \mathcal{I}_0 = \mathbb{R}^m$.

Furthermore, the directed graph is uniquely specified by its (reduced) co-incidence matrix $\mathbf{D}$ given by

$$\mathbf{D} = [d_{ij}]_{m \times n}, \quad d_{i,j} = \begin{cases} 1 & \epsilon_{i,1} = \eta_j \\ -1 & \epsilon_{i,0} = \eta_j \\ 0 & \text{otherwise} \end{cases}. \tag{3}$$

Kirchhoff Current (KCL) and Voltage laws (KVL) [3] can be expressed with an elegant duality (see [16] p.710) using the incidence and coincidence matrices by

$$\mathbf{v}_1 = \mathbf{D}\mathbf{v}_0, \qquad \mathbf{i}_0 = -\mathbf{D}^\mathsf{T}\mathbf{i}_1 = 0. \tag{4}$$

i.e. we have the following diagram.

$$\begin{array}{ccc} \mathbf{v}_0 \in \mathcal{V}_0 & \xrightarrow{\mathbf{D}} & \mathbf{v}_1 \in \mathcal{V}_1 \\ \updownarrow \langle \cdot \,|\, \cdot \rangle_{\mathcal{B}_0} & & \updownarrow \langle \cdot \,|\, \cdot \rangle_{\mathcal{B}_1} \\ \mathbf{i}_0 \in \mathcal{I}_0 & \xleftarrow[-\mathbf{D}^\mathsf{T}]{} & \mathbf{i}_1 \in \mathcal{I}_1 \end{array} \tag{5}$$

[2]This notation is convenient to make the link with automated circuit to Bond-graph algorithms [14]: **0**-junctions (shared voltage, parallel connection) for nodes and **1**-junctions for branches (shared current, serial connection) see Figures 5 and 6 for examples. It is also a mnemonic to remember that lumped circuit equations arise from the spatial discretization of electro-magnetic 1-forms for branches and 0-forms for nodes.

[3]The minus sign in front of $\mathbf{i}_0$ comes from the consistent use of the receiver convention for both nodes and branches: the sum of edge currents $\mathbf{i}_0$ *entering* each node has to be zero.

2.2. Kirchhoff-Dirac structure

Written in matrix form, one obtains the canonical Kirchhoff-Dirac structure $\mathcal{D}$ (with a structure very similar to the ones obtained for partial differential equations (PDE) [17] [5])

$$\mathcal{D} : \begin{bmatrix} \mathbf{i}_0 \\ \mathbf{v}_1 \end{bmatrix} = \begin{bmatrix} \mathbf{0} & -\mathbf{D}^\mathsf{T} \\ \mathbf{D} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{v}_0 \\ \mathbf{i}_1 \end{bmatrix}, \quad \mathbf{i}_0 = 0. \tag{6}$$

i.e. Kirchhoff Current Laws can be interpreted as *zero boundary conditions* on the node currents, and the co-incidence matrix $\mathbf{D}$ as a (lumped) differential operator. Left multiplying by $[\mathbf{v}_0 \;\; \mathbf{i}_1]$, the duality products and skew-symmetry leads to the power balance

$$P = \langle \mathbf{i}_0 \,|\, \mathbf{v}_0 \rangle + \langle \mathbf{i}_1 \,|\, \mathbf{v}_1 \rangle = \begin{bmatrix} \mathbf{v}_0 & \mathbf{i}_1 \end{bmatrix} \begin{bmatrix} \mathbf{0} & -\mathbf{D}^\mathsf{T} \\ \mathbf{D} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{v}_0 \\ \mathbf{i}_1 \end{bmatrix} = 0.$$

Furthermore since we have conservation of charge $\mathbf{i}_0 = 0$ on the nodes $\mathcal{N}$, this yields the Tellegen theorem over the edges[4] $\mathcal{E}$

$$\langle \mathbf{i}_1 \,|\, \mathbf{v}_1 \rangle = \sum_{\epsilon \in \mathcal{E}} \langle i_\epsilon \,|\, v_\epsilon \rangle = 0.$$

We also remark that the node voltages $\mathbf{v}_0$ can be interpreted as Lagrange multipliers parametrizing the sub-manifold defined by the linear constraints $\mathbf{i}_0 = 0$.

2.3. (Reduced) Hybrid Dirac structure

Whereas MNA solves the system for node voltages and branch currents, in Hybrid Analysis [16] and skew-gradient DAE [7] [8], the node voltages are eliminated. First a spanning tree T is chosen, this yields a partition of the branch currents and voltages into tree $(\mathbf{v}_T, \mathbf{i}_T)$ and link variables $(\mathbf{v}_L, \mathbf{i}_L)$. Partitioning equations according to the spanning tree, Kirchhoff laws (4) are rewritten as

$$\begin{bmatrix} \mathbf{v}_T \\ \mathbf{v}_L \end{bmatrix} = \begin{bmatrix} \mathbf{D}_T \\ \mathbf{D}_L \end{bmatrix} \mathbf{v}_0, \qquad \begin{bmatrix} \mathbf{D}_T^\mathsf{T} & \mathbf{D}_L^\mathsf{T} \end{bmatrix} \begin{bmatrix} \mathbf{i}_T \\ \mathbf{i}_L \end{bmatrix} = 0. \tag{7}$$

From graph theory, having a spanning tree ensures that the matrix $\mathbf{D}_T \in \mathbb{R}^{n \times n}$ is invertible. So we can eliminate the node voltages $\mathbf{v}_0$ using $\mathbf{v}_0 = \mathbf{D}_T^{-1}\mathbf{v}_T$. This yields a reduced Hybrid Dirac structure specified by its link-cutset matrix $\mathbf{C} = (\mathbf{D}_L\mathbf{D}_T^{-1})^\mathsf{T}$

$$\mathcal{D} : \begin{bmatrix} \mathbf{i}_T \\ \mathbf{v}_L \end{bmatrix} = \begin{bmatrix} \mathbf{0} & -\mathbf{C} \\ \mathbf{C}^\mathsf{T} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{v}_T \\ \mathbf{i}_L \end{bmatrix}. \tag{8}$$

Traditionally, the spanning tree is chosen to be a proper tree (i.e. containing all current-driven branches: Voltages Sources, Capacitors, ...) such that $\mathbf{v}_T$ is current-driven by $\mathbf{i}_T$ (i.e. computable from $\mathbf{i}_T$). However topological constraints such as in example 6.2 may prevent a proper tree to be found. Since the proposed method is fully-implicit by nature, it does not have such a requirement. Either the Kirchhoff-Dirac structure or *any* reduced Hybrid Dirac structure can be used for simulation.

For a formal definition of Dirac structures in the broader context of multi-physical networks, pleaser refer to [6] and references therein. A generic example of a Dirac structure and its graph, emphasizing the node-edge incidence structure, is shown on Figure 1. Detailed case-study are shown on Figures 5 and 6 and studied in Section 6.

[4]Indeed (see [16] p. 30) any two of KCL, KVL and Tellegen theorem implies the third one.

3. ALGEBRO-DIFFERENTIAL PARAMETRIZATION OF COMPONENT LAWS

From now on, for functional discretization purpose, we adopt a Hilbert space viewpoint, and lift Dirac structures over time steps. Consider a time interval $\Omega \subset \mathbb{R}$, the branch voltage and current spaces are lifted to the dual Hilbert spaces $\mathcal{I}_1 \sim \mathcal{V}_1 \subseteq L^2(\Omega)^m$ (L^2 being a pivot space) equipped with the inner (duality) product

$$\langle \mathbf{u} \,|\, \mathbf{v} \rangle := \frac{1}{|\Omega|} \int_\Omega \mathbf{u}(t)^\mathsf{T} \mathbf{v}(t) \,\mathrm{d}t. \tag{9}$$

We assume that branch equations can be parametrized locally by a state $\boldsymbol{x} \in \mathcal{X}_1 \subseteq L^2(\Omega)^m$, *nonlinear differential-algebraic operators* $\boldsymbol{I}_1 : \mathcal{X}_1 \to \mathcal{I}_1, \boldsymbol{V}_1 : \mathcal{X}_1 \to \mathcal{V}_1$ and a law

$$\begin{array}{rccl} F & : & \mathcal{X}_1 & \longrightarrow & \mathcal{B}_1 := \mathcal{I}_1 \times \mathcal{V}_1 \\ & & \boldsymbol{x} & \longmapsto & (\boldsymbol{I}_1(\boldsymbol{x}), \boldsymbol{V}_1(\boldsymbol{x})) \end{array}. \tag{10}$$

Likewise the KCL node boundary conditions (4) can be parametrized by the vector of node voltages $\boldsymbol{\lambda} \in \mathcal{X}_0 \subseteq L^2(\Omega)^n$ and the linear constraint

$$\begin{array}{rccl} B & : & \mathcal{X}_0 & \longrightarrow & \mathcal{B}_0 := \mathcal{I}_0 \times \mathcal{V}_0 \\ & & \boldsymbol{\lambda} & \longmapsto & (\boldsymbol{I}_0, \boldsymbol{V}_0)(\boldsymbol{\lambda}) = (0, \boldsymbol{\lambda}) \end{array}. \tag{11}$$

Composing (6) with (10) (11) we obtain the fully implicit algebro-differential formulation of a Port-Hamiltonian system (PHS)

$$\Sigma = \left\{ \begin{array}{l} (\boldsymbol{I}_0, \boldsymbol{V}_0, \boldsymbol{I}_1, \boldsymbol{V}_1)(\boldsymbol{x}, \boldsymbol{\lambda}) \in \mathcal{B}_1 \times \mathcal{B}_0; \\ N(\boldsymbol{x}) = 0, \quad \forall (\boldsymbol{x}, \boldsymbol{\lambda}) \in \mathcal{X}_1 \times \mathcal{X}_0 \end{array} \right\} \tag{12}$$

defined by the operator $N : \mathcal{X}_0 \times \mathcal{X}_1 \to L^2(\Omega)^{m+n}$

$$N(\boldsymbol{x}, \boldsymbol{\lambda}) = \begin{bmatrix} \mathbf{0} \\ \boldsymbol{V}_1(\boldsymbol{x}) \end{bmatrix} - \begin{bmatrix} \mathbf{0} & -\mathbf{D}^\mathsf{T} \\ \mathbf{D} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\lambda} \\ \boldsymbol{I}_1(\boldsymbol{x}) \end{bmatrix}. \tag{13}$$

For the reduced Hybrid Dirac structure one gets

$$\Sigma = \left\{ (\boldsymbol{I}_T, \boldsymbol{V}_T, \boldsymbol{I}_C, \boldsymbol{V}_C)(\boldsymbol{x}) \in \mathcal{B}_1 \,|\, N(\boldsymbol{x}) = 0, \, \forall \boldsymbol{x} \in \mathcal{X}_1 \right\} \tag{14}$$

with the algebro-differential operator $N : \mathcal{X}_1 \to L^2(\Omega)^m$

$$N(\boldsymbol{x}) = \begin{bmatrix} \boldsymbol{I}_T(\boldsymbol{x}) \\ \boldsymbol{V}_C(\boldsymbol{x}) \end{bmatrix} - \begin{bmatrix} \mathbf{0} & -\mathbf{C} \\ \mathbf{C}^\mathsf{T} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{V}_T(\boldsymbol{x}) \\ \boldsymbol{I}_C(\boldsymbol{x}) \end{bmatrix}. \tag{15}$$

We note that for differential components, the state space is given by the Sobolev space $\mathcal{X} \subseteq H^1(\Omega) \subset L^2(\Omega)$ defined by

$$\mathcal{X} = \left\{ x \in L^2(\Omega) \,\middle|\, \dot{x} \in L^2(\Omega); \; x(t) = x_0 + \int_0^t \dot{x}(s) \,\mathrm{d}s \right\}, \tag{16}$$

whereas for algebraic components, no additional smoothness is implied so $\mathcal{X} \sim L^2(\Omega)$.

The differential-algebraic operators corresponding to common electronic components are summarized in Table 1 and the case of implicitly parametrized algebraic components is now further detailed in Section 4.

4. (PSEUDO) ARC-LENGTH PARAMETRIZATION

We study here implicit arc-length and pseudo arc-length parameterizations of algebraic components whose laws cannot be represented as functions of either current or voltage (or such that unbounded Lipschitz constants may cause numerical problems during simulations). As an example we consider the cases of the ideal diode, a nonlinear resistor and the Shockley diode.

4.1. The ideal diode

An ideal diode law is determined by the set (see [15])

$$\mathcal{R}_D = \left\{ (v, i) \in \mathbb{R}^2 \,\middle|\, \begin{cases} v = 0 & i \in \mathbb{R}^+, \\ i = 0 & v \in \mathbb{R}^-. \end{cases} \right\} \tag{17}$$

It has the numerical disadvantage of being alternatively voltage and current controlled. In the hybrid formulation, computational causality assignment [14] would imply that a different Dirac structure such as (8) should be used according to the current state of the circuit. Furthermore, when the number of switching components grows, the number of switch configurations of the circuit grows exponentially. A solution around this problem is to consider the parametrization $R_D : \lambda \mapsto (V_D(\lambda), I_D(\lambda))$ with arc-length

$$\lambda(v, i) = \begin{cases} i/I_0 & v = 0, \; i \in \mathbb{R}^+ \\ v/V_0 & i = 0, \; v \in \mathbb{R}^- \end{cases}, \tag{18}$$

for arbitrarily chosen positive reference current and voltages I_0, V_0. Inverting the relation, one obtains the algebraic operators

$$V_D(\lambda) = V_0 \min(\lambda, 0), \qquad I_D(\lambda) = I_0 \max(\lambda, 0). \tag{19}$$

with $V_D'(\lambda) = V_0 \cdot \mathbf{1}_{\mathbb{R}^-}(\lambda)$, and $I_D'(\lambda) = I_0 \cdot \mathbf{1}_{\mathbb{R}^+}(\lambda)$, where $\mathbf{1}_A(\lambda)$ denotes the indicator function of a set A.

Differential	x	$V(x)$	$I(x)$	$H(x)$
Capacitor	q	q/C	$\dot{q}$	$q^2/2\mathsf{C}$
Inductor	ϕ	$\dot{\phi}$	ϕ/L	$\phi^2/2\mathsf{L}$
Nonlinear Capacitor	q	$\nabla H(q)$	$\dot{q}$	$H(q)$
Nonlinear Inductor	ϕ	$\dot{\phi}$	$\nabla H(\phi)$	$H(\phi)$

Algebraic	x	$V(x)$	$I(x)$	$P(x)$
Resistor	i	$\mathsf{R}i$	i	$\mathsf{R}i^2$
Conductor	v	v	$\mathsf{G}v$	$\mathsf{G}v^2$
Nonlinear Resistor	i	$z(i)$	i	$i \cdot z(i)$
Nonlinear Conductor	v	v	$z(v)$	$v \cdot z(v)$
Voltage source	i	V	i	$\mathsf{V} \cdot i$
Current source	v	v	I	$\mathsf{I} \cdot v$

Table 1: *Differential and Algebraic components. H (energy), P (power), q (charge), ϕ (flux), z (non linear function).*

4.2. A Hard Clipping resistor

We now consider the case of a hard clipping resistor (it will be used in example 6.1) whose (v, i) graph is described by the set

$$\mathcal{R}_D = \left\{ (v,i) \in \mathbb{R}^2;\ \begin{cases} i \in \mathbb{R}^- \setminus \{0\} & v \in \{-1\} \\ i \in \{0\} & v \in (-1,1) \\ i \in \mathbb{R}^+ \setminus \{0\} & v \in \{1\} \end{cases} \right\}. \quad (20)$$

We parametrize it continuously using (see Figure 4 page 8)

$$\mathcal{R}_D = \left\{ (v,i) \in \mathbb{R}^2; \quad (v,i) = (V(\lambda), I(\lambda)), \quad \forall \lambda \in \mathbb{R} \right\} \quad (21)$$

with the voltage and current operators

$$V(\lambda) = V_0 \operatorname{clip}_{[-1,1]}(\lambda), \quad (22)$$

$$I(\lambda) = I_0 \left(\min(0, \lambda + 1) + \max(0, \lambda - 1) \right). \quad (23)$$

For arbitarily chosen positive reference voltage and currents V_0, I_0.

4.3. The Shockley diode

We finally consider the Shockley diode model [5].

$$I(v) = I_S \left(\exp\left(\frac{v}{V_T} \right) - 1 \right), \quad (24)$$

where I_S is the saturation current, $V_T = k_b T / q_e$ the thermal voltage, with temperature T, Boltzmann constant k_b and electron charge q_e. It is $\mathcal{C}^\infty$-continuous, but not globally Lipschitz.

For a chosen reference resistance R_0, the true arc-length of the graph $(v, R_0 I(v))$ is determined by $\mathrm{d}\lambda^2 = (1 + (R_0 I'(v))^2)\, \mathrm{d}v^2$ but it is not practical to manipulate. Instead, introducing the diode cutoff point (V_0, I_0) as the point of unit slope

$$R_0 I'(V_0) = 1, \qquad I_0 = I(V_0), \quad (25)$$

where $V_0 = V_T \ln\left(\frac{V_T}{R_0 I_S}\right)$, $I_0 = V_T / R_0 - I_S$. Remarking that for $v \ll V_0$, $\mathrm{d}\lambda \approx \mathrm{d}v$ and for $v \gg V_0$, $\mathrm{d}\lambda \approx R_0 I'(v)\, \mathrm{d}v$, one can introduce the $\mathcal{C}^0$ pseudo arc-length differential

$$\mathrm{d}\tilde{\lambda}(v) = \begin{cases} \mathrm{d}v & v < V_0 \\ R_0 I'(v)\, \mathrm{d}v & v \geq V_0 \end{cases}. \quad (26)$$

Integrating $\tilde{\lambda}(v) := \int_0^v \mathrm{d}\tilde{\lambda}$ one obtains the $\mathcal{C}^1$ pseudo-arclength

$$\tilde{\lambda}(v) = \begin{cases} v & v < V_0 \\ V_0 + R_0 (I(v) - I_0), & v \geq V_0. \end{cases} \quad (27)$$

Inverting the relation leads to the algebraic operators

$$V_D(\lambda) = \begin{cases} \lambda & \lambda < V_0, \\ V_T \ln\left(1 + \dfrac{I_0 + (\lambda - V_0)/R_0}{I_S}\right) & \lambda \geq V_0, \end{cases} \quad (28)$$

$$I_D(\lambda) = \begin{cases} I(\lambda) & \lambda < V_0, \\ I_0 + \dfrac{\lambda - V_0}{R_0} & \lambda \geq V_0 \end{cases}. \quad (29)$$

[5] Anti-parallel Shockley diodes will be simulated in example 6.1

such that by construction, Lipschitz constants are unitary (this property is key to deal with convergence and numerical stiffness)

$$L_V = \sup_\lambda \left| V_D' \right| = 1, \qquad L_I = \sup_\lambda \left| R_0 I_D' \right| = 1. \quad (30)$$

5. FUNCTIONAL DISCRETIZATION AND NUMERICAL SOLVER

We now use the functional framework presented in Section 3 to discretize the system with a finite number of parameters per time step, (see the reference [18] for the representation of non band-limited signals having a *finite rate of innovation*).

Our time discretisation scheme can be interpreted as an extension of (spectral) time-finite elements methods [19] to DAE. It is based on the following theorem which proves that a weak PHS is preserved over the chosen approximation subspace.

Theorem 5.1 (Weak PHS)**.** *Let Ω be a time step, $\boldsymbol{x} \in \mathcal{X} \subseteq L^2(\Omega)^m$ a functional state, two operators $\boldsymbol{b} : \mathcal{X} \to L^2(\Omega)^m$, $\boldsymbol{a} : \mathcal{X} \to L^2(\Omega)^m$ and a skew-symmetric matrix $\mathbf{J}$ defining the PHS operator*

$$N(\boldsymbol{x}) = \boldsymbol{b}(\boldsymbol{x}) - \mathbf{J}\boldsymbol{a}(\boldsymbol{x}) = 0, \quad \mathbf{J} = -\mathbf{J}^*. \quad (31)$$

Let $\boldsymbol{P} : L^2(\Omega) \to R(\boldsymbol{P}) \subseteq L^2(\Omega)^m$ be a projector ($\boldsymbol{P}^2 = \boldsymbol{P}$) satisfying the skew-adjoint commutation $\boldsymbol{P}\mathbf{J} = \mathbf{J}\boldsymbol{P}^*$, for the L^2 inner product (9)*, then the projected operator*

$$\boldsymbol{P} \circ N(\boldsymbol{x}) = 0 \quad (32)$$

defines a weak PHS which preserves the power balance.

$$\left\langle \boldsymbol{a}(\boldsymbol{x}) \,\middle|\, \boldsymbol{P} \,\middle|\, \boldsymbol{b}(\boldsymbol{x}) \right\rangle = 0. \quad (33)$$

Proof. Using (32), taking the inner product with $\boldsymbol{a}(\boldsymbol{x})$, and using the fact that 1) $\boldsymbol{P}^2 = \boldsymbol{P}$ (idempotence), 2) we have the commutation $\boldsymbol{P}\mathbf{J} = \mathbf{J}\boldsymbol{P}^*$ and 3) $\boldsymbol{P}\mathbf{J}\boldsymbol{P}^*$ is skew-adjoint, we obtain

$$\begin{aligned} & \left\langle \boldsymbol{a}(\boldsymbol{x}) \,\middle|\, \boldsymbol{P} \,\middle|\, N(\boldsymbol{x}) \right\rangle = 0 \\ \iff & \left\langle \boldsymbol{a}(\boldsymbol{x}) \,\middle|\, \boldsymbol{P} \,\middle|\, \boldsymbol{b}(\boldsymbol{x}) \right\rangle = \left\langle \boldsymbol{a}(\boldsymbol{x}) \,\middle|\, \boldsymbol{P}\mathbf{J} \,\middle|\, \boldsymbol{a}(\boldsymbol{x}) \right\rangle \\ & \overset{1}{=} \left\langle \boldsymbol{a}(\boldsymbol{x}) \,\middle|\, \boldsymbol{P}^2\mathbf{J} \,\middle|\, \boldsymbol{a}(\boldsymbol{x}) \right\rangle \\ & \overset{2}{=} \left\langle \boldsymbol{a}(\boldsymbol{x}) \,\middle|\, \boldsymbol{P}\mathbf{J}\boldsymbol{P}^* \,\middle|\, \boldsymbol{a}(\boldsymbol{x}) \right\rangle \overset{3}{=} 0. \end{aligned}$$

□

Remark (Energy conservation)**.** As an immediate consequence, for a conservative Hamiltonian system given by the operator

$$N(\boldsymbol{x}) = \frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} - \mathbf{J}\nabla H(\boldsymbol{x}) = 0, \quad \mathbf{J} = -\mathbf{J}^T. \quad (34)$$

discretized such that $\dot{\boldsymbol{x}} = \boldsymbol{P}\mathbf{J}\nabla H(\boldsymbol{x})$, then the Hamiltonian energy H is preserved over a time-step $\Omega = (t_0, t_1)$,

$$H(\boldsymbol{x}(t_1)) - H(\boldsymbol{x}(t_0)) = 0. \quad (35)$$

Indeed, let $\boldsymbol{b} = \frac{\mathrm{d}}{\mathrm{d}t}$ and $\boldsymbol{a} = \nabla H$, from the gradient theorem and using the same arguments as the previous proof, it follows that

$$\begin{aligned} H(\boldsymbol{x}(t_1)) - H(\boldsymbol{x}(t_0)) &= \left\langle \nabla H(\boldsymbol{x}) \,\middle|\, \dot{\boldsymbol{x}} \right\rangle \\ &= \left\langle \nabla H(\boldsymbol{x}) \,\middle|\, \boldsymbol{P}\mathbf{J}\boldsymbol{P}^* \,\middle|\, \nabla H(\boldsymbol{x}) \right\rangle = 0. \end{aligned}$$

5.1. Piecewise constant and affine polynomial spaces

In this article we will restrict ourselves to constant and affine polynomial spaces $\mathbb{P}^0, \mathbb{P}^1$ for which we have *exact closed-form* expression of the projected operators. (Higher-order polynomial spaces require the use of approximate quadratures rules [2] [9]). Results are exposed without proof except when the proof is not available elsewhere (see [8]).

Consider a unit time step $\Omega = (0, 1)$, for the normalized time variable $\tau \in (0, 1)$ and two orthogonal polynomials

$$\ell_0(\tau) = 1, \qquad \ell_1(\tau) = \tau - \frac{1}{2}.$$

The operator $P_K : L^2(\Omega) \to \mathbb{P}^K(\Omega) \subset L^2(\Omega)$, $K \in \{0, 1\}$ defined by

$$(P_K\, u)(\tau) = \sum_{i=0}^{K} \ell_i(\tau) \frac{\langle \ell_i \mid u \rangle}{\langle \ell_i \mid \ell_i \rangle} \tag{36}$$

is an orthogonal projector. i.e. P_K is self-adjoint ($P_K = P_K^*$) and idempotent ($P_K^2 = P_K$). For notational simplicity, we define the following notation. Let $A : L^2(\Omega) \to L^2(\Omega)$ be an operator, the *projected operator* $\bar{A}_K : L^2(\Omega) \to \mathbb{P}^K(\Omega)$ is defined by

$$\bar{A}_K := P_K \circ A, \quad \bar{A} := \bar{A}_0. \tag{37}$$

By extension, for a vectorized projector $\boldsymbol{P} := P_K \otimes \mathbf{I}_n$, it yields the projected PHS operator

$$\overline{N}(\boldsymbol{x}) := \boldsymbol{P} \circ N(\boldsymbol{x}) \tag{38}$$

Because of the tensor product construction, we also have the commutation $\boldsymbol{P}\mathbf{J} = \mathbf{J}\boldsymbol{P} = \mathbf{J}\boldsymbol{P}^*$ such that $\boldsymbol{P}$ satisfies Theorem 5.1.

For numerical computations, it is necessary to compute the polynomial coefficients of the image of a trajectory through a nonlinear function. This is possible thanks to the following property

Property 5.1 (Projected function)**.** Let $f : \mathbb{R} \to \mathbb{R}$ be a semi-continuous function with known antiderivative F and a function

$$x(\tau) = \ell_0(\tau)\bar{x} + \ell_1(\tau)\delta x \;\in\; \mathbb{P}^1(\Omega), \tag{39}$$

parametrized by its mean and variation $\Theta = (\bar{x}, \delta x) \in \mathbb{R}^2$

Then the projected function $P_1 \circ f \circ x$ has the projection coefficients $\bar{\mathbf{f}} : \mathbb{R}^2 \to \mathbb{R}^2$ defined by

$$\bar{\mathbf{f}}_i := \langle \ell_i \mid f \circ x \rangle / \|\ell_i\|^2 . \tag{40}$$

They are given in closed form by

$$\bar{\mathbf{f}}_0(\Theta) = \begin{cases} \dfrac{F(\bar{x} + \frac{\delta x}{2}) - F(\bar{x} - \frac{\delta x}{2})}{\delta x} & \delta x \neq 0 \\ \dfrac{f(\bar{x}^+) + f(\bar{x}^-)}{2} & \delta x = 0 \end{cases} \tag{41}$$

$$\bar{\mathbf{f}}_1(\Theta) = \begin{cases} \dfrac{12}{\delta x}\left(\dfrac{F(x_1) + F(x_0)}{2} - \overline{\mathbf{F}}_0(\Theta)\right) & \delta x \neq 0 \\ 0 & \delta x = 0 \end{cases} \tag{42}$$

where $x_1 = \bar{x} + \delta x/2$, $x_0 = \bar{x} - \delta x/2$.

Proof. See Appendix A. □

Note that for a scalar (or separable) potential F, using $f = \nabla F$, and $\bar{x} = (x_0 + x_1)/2$, $\delta x = x_1 - x_0$ in property 5.1 yields the *Average Discrete Gradient* from [8] (this is also an instance of anti-derivative anti-aliasing)

$$\overline{\nabla} F(x_0, x_1) := \bar{\mathbf{f}}_0(\Theta). \tag{43}$$

Additional results for linear gradients are given in appendix B.

5.2. Newton iteration

For each time step Ω, let Θ denote the unknown parameters of a local state $\boldsymbol{x}_\Theta \in (\mathbb{P}^K(\Omega))^m$ we look for a zero $\overline{\mathbf{N}}(\Theta^\star) = 0$ of

$$\overline{\mathbf{N}}(\Theta) := \left[\left\langle \ell_i \,\middle|\, N(\boldsymbol{x}_\Theta)\right\rangle / \|\ell_i\|^2\right]_{i=0\ldots K} \tag{44}$$

using Newton iteration (line search is not used in this paper)

$$\Theta_{\kappa+1} = \Theta_\kappa + \Delta\Theta_\kappa, \quad \Delta\Theta_\kappa = -\overline{\mathbf{N}}'(\Theta_\kappa)^{-1}\overline{\mathbf{N}}(\Theta_\kappa). \tag{45}$$

A detailed convergence analysis for the general case is out of the scope this paper and is left for future work. Please refer to [20] for more details. When $\overline{\mathbf{N}}$ is only semi-smooth which is the case of the ideal and hard clipping diodes, special care should be taken to ensure convergence using semi-smooth Newton methods [21].

It should be noted that in piecewise constant spaces ($k = 0$), algebraic constraints simplifies to $\overline{\mathbf{V}}(\mathbf{s}) = V(s)$, $\overline{\mathbf{I}}(\mathbf{s}) = I(s)$, and one can compute the Jacobian from the derivative V', I'. For affine trajectories ($k = 1$) one should use the results from properties 5.1 and the following property from [8] to compute the coefficients and the Jacobian.

Property 5.2. Given a potential $F \in C^2(\mathbb{R}, \mathbb{R})$, and its discrete gradient $\overline{\nabla} F(x_0, x_1)$ defined in Equation (43), the derivative of the discrete gradient with respect to x_1 is

$$\frac{\partial \overline{\nabla} F}{\partial x_1} = \begin{cases} \dfrac{\nabla F(x_1) - \overline{\nabla} F(x_0, x_1)}{x_1 - x_0} & x_0 \neq x_1 \\ \dfrac{1}{2}\dfrac{\partial^2 F}{\partial x^2}(x_0) & x_0 = x_1 \end{cases}. \tag{46}$$

6. EXAMPLES

6.1. Diode Clipper

We consider the diode clipper circuit shown in Figure 5. This circuit which is dissipatively stiff because of the diode unbounded Lipschitz constant is commonly used to benchmark numerical schemes. In this paper the nonlinear resistor D is considered abstract and will be substituted by anti-parallel Shockley and hard clipping diode models from Section 4.

Over each time step it is parametrized by the vector of Legendre coefficients $\Theta = (\mathbf{i}_S, \mathbf{i}_C, \mathbf{v}_R, \mathbf{x}_D) \in (\mathbb{R}^{K+1})^4$, for $K \in \{0, 1\}$ and the functional state $\boldsymbol{x}_\Theta = [i_S, i_C, v_R, x_D]^T \in (\mathbb{P}^K(\Omega))^4$ such that each element $v \in \mathbb{P}^K(\Omega)$ is of the form $v(t_0 + h\tau) = \sum_{n=0}^{K} \ell_n(\tau)\mathbf{v}[n]$. The projected Dirac structure (where $\mathbf{1}$ is the identity on $\mathbb{R}^{K+1}$) is then given by the operator

$$\overline{\mathbf{N}} = \begin{bmatrix} \mathbf{i}_S \\ \mathbf{i}_C \\ \mathbf{v}_R \\ \overline{\mathbf{V}}_D(\mathbf{x}_D) \end{bmatrix} - \begin{bmatrix} . & . & -\mathbf{1} & \mathbf{0} \\ . & . & \mathbf{1} & -\mathbf{1} \\ \mathbf{1} & -\mathbf{1} & . & . \\ \mathbf{0} & \mathbf{1} & . & . \end{bmatrix} \begin{bmatrix} \overline{\mathbf{V}}_S \\ \overline{\mathbf{V}}_C(\mathbf{i}_C) \\ \overline{\mathbf{I}}_R(\mathbf{v}_R) \\ \overline{\mathbf{I}}_D(\mathbf{x}_D) \end{bmatrix}. \tag{47}$$

Results are shown on Figure 2 for an input $v_S = V\sin(2\pi f_0 t)$ with high input gain $V = 10^4$, fundamental frequency $f_0 = 500$ Hz, $R = 1$ kΩ , $C = 10\mu$F, $I_S = 100$ fA, $R_0 = 0.1\Omega$, sampling frequency $f_s = 96$ kHz. Anti-parallel Shockley diodes with arc length converge on average in 2 iterations and 4 times reduction of the worst-case iteration count (Newton tolerance $\epsilon_r = 10^{-5}$), Hard clipping diodes exhibit convergence in one iteration most of the time (2 when switching) even for $\epsilon_r = 10^{-10}$.

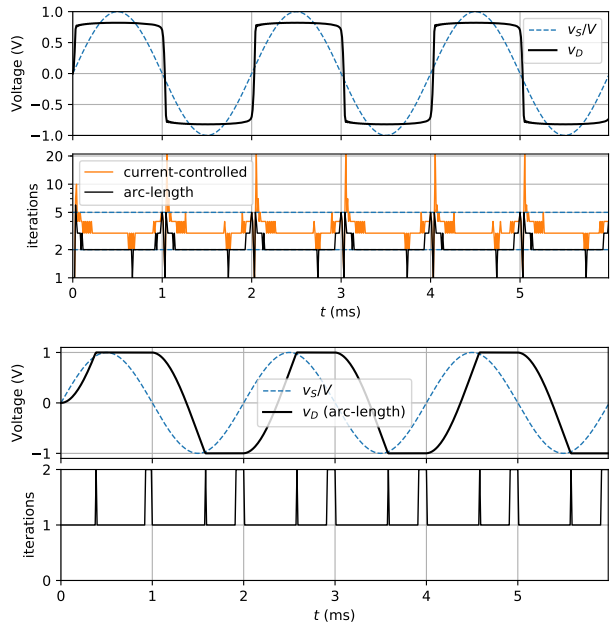


Figure 2: *Diode clipper: anti-parallel Shockley diodes (top) with $V = 10^4$ to emphasize Newton iteration differences, Hard clipping diodes (bottom) $V = 10^2$ to see dynamic and saturation.*

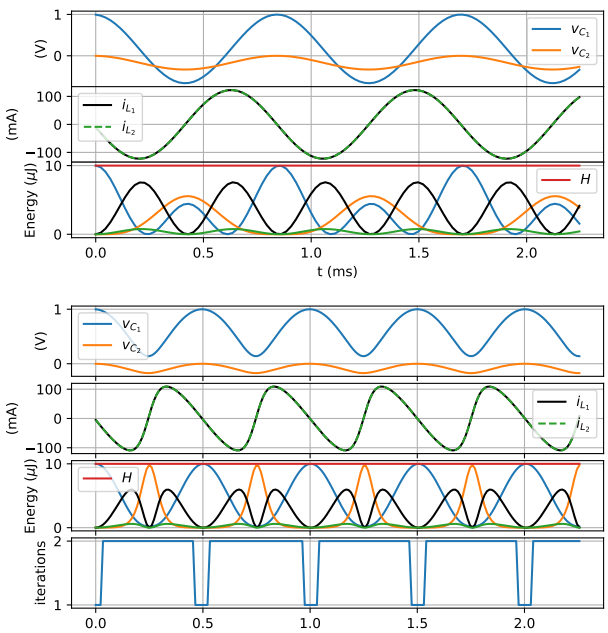


Figure 3: *Conservative LCLC circuit: Linear (top) and Nonlinear (bottom). Notice the periodicity change and conserved energy.*

6.2. LCLC circuit

We study here an LCLC circuit (shown on Figure 6) chosen to demonstrate the proposed method when the circuit is conservative, nonlinear and contains topological constraints (parallel capacitors, serial inductors ...). Here the circuit contains two inductors with the implicit topological constraint $i_{L_1} = i_{L_2}$.

In traditional solvers, such constraints usually needs to be detected and eliminated before proceeding to simulation. A possible approach is the use of equivalent macro components (see [22] [23]). In contrast, the proposed approach doesn't require such a preprocessing step, and keeps the modularity and sparsity of the component-based description. To demonstrate energy conservation, the capacitor C_2 is chosen first with a linear law $V_{C_2}(q) = q/C_2$ and an hardening nonlinearity $V_{C_2}(q) = V_\alpha \sinh(\frac{q}{C_2 V_\alpha})$ with $V_\alpha = 1/30$ (V).

Using the vector of Legendre coefficients as unknown $\Theta = (\mathbf{i}_{C_1}, \mathbf{v}_{L_1}, \mathbf{i}_{C_2}, \mathbf{v}_{L_2})$, we have the projected Dirac structure operator

$$\overline{\mathbf{N}} = \begin{bmatrix} \mathbf{i}_{C1} \\ \overline{\mathbf{I}}_{L1}(\mathbf{v}_{L1}) \\ \mathbf{i}_{C2} \\ \mathbf{v}_{L2} \end{bmatrix} - \begin{bmatrix} . & . & . & \mathbf{1} \\ . & . & . & \mathbf{1} \\ . & . & . & \mathbf{1} \\ -\mathbf{1} & -\mathbf{1} & -\mathbf{1} & . \end{bmatrix} \begin{bmatrix} \overline{\mathbf{V}}_{C1}(\mathbf{i}_{C1}) \\ \mathbf{v}_{L1} \\ \overline{\mathbf{V}}_{C2}(\mathbf{i}_{C2}) \\ \overline{\mathbf{I}}_{L2}(\mathbf{v}_{L2}) \end{bmatrix} \quad (48)$$

Simulation results are shown for the implicit and nonlinear LCLC circuit on Figure 3 for $f_s = 88.2$ kHz, $C_1 = 20\mu$F, $C_2 = 100\mu$F, $L_1 = 1$mH, $L_2 = 100\mu$H, zero initial conditions and $v_{C_1}(0) = 1$V. We observe that both the algebraic constraint $i_{L_1} = i_{L_2}$ and the conservation of total energy H are respected. Convergence is reached in 1 iteration for the linear case and between 1 and 2 iterations for the nonlinear one (relative tolerance $\epsilon_r = 10^{-5}$).

7. CONCLUSIONS

A new power-balanced, fully implicit component oriented method has been presented with a functional time-discretization. Its main strengths (not necessarily unique to this method) are: a) it retains the topological sparsity and modularity of the network based description, b) it is power-balanced and energy-conserving (including nonlinear components), c) it can deal with implicit topological constraints (capacitor loops, inductor cutsets) without the need of manual substitution of equivalent components, d) it can deal with implicit components including switching components, e) it uses finite-dimensional subspace projection as a unifying discretization tool common to ODE, PDE and DAE. f) Newton iteration converges faster using arc-length description of algebraic components with unbounded Lipschitz constants,

Regarding perspectives, a detailed convergence study of the Newton iteration is needed (such as the one in [24]), but has been postponed for future work. Using different and higher order functional approximation spaces is also an obvious perspective provided the projections can be computed exactly and efficiently. In particular, from a generalized sampling theory viewpoint, it would be interesting to perform a comparative analysis of implementation cost and convergence rate (to the true solution) between functional projection and oversampling.

8. REFERENCES

[1] G. Hachtel, R. Brayton, and F. Gustavson, "The sparse tableau approach to network analysis and design," *IEEE Transactions on Circuit Theory*, vol. 18, no. 1, pp. 101–113, 1971.

[2] E. Celledoni and E. H. Høiseth, "Energy-preserving and passivity-consistent numerical discretization of port-

Hamiltonian systems," *arXiv preprint arXiv:1706.08621*, 2017.

[3] I. Y. Dorfman, "Dirac structures of integrable evolution equations," *Physics Letters A*, vol. 125, no. 5, pp. 240–246, 1987.

[4] T. Courant and A. Weinstein, "Beyond Poisson structures," *Action hamiltoniennes de groupes. Troisieme théoreme de Lie*, vol. 27, pp. 39–49, 1988.

[5] P. Kotyczka, B. Maschke, and L. Lefèvre, "Weak form of Stokes–Dirac structures and geometric discretization of port-Hamiltonian systems," *Journal of Computational Physics*, vol. 361, pp. 442–476, 2018.

[6] A. van der Schaft and D. Jeltsema, "Port-Hamiltonian systems theory: An introductory overview," *Foundations and Trends in Systems and Control*, vol. 1, no. 2-3, pp. 173–378, 2014.

[7] A. Falaize and T. Hélie, "Passive guaranteed simulation of analog audio circuits: A port-Hamiltonian approach," *Applied Sciences*, vol. 6, no. 10, 2016.

[8] R. Muller and T. Hélie, "Power-balanced modelling of circuits as skew gradient systems," in *Proc. 21th Conf. Digital Audio Effects*, 2018.

[9] V. Mehrmann and R. Morandin, "Structure-preserving discretization for port-Hamiltonian descriptor systems," *arXiv preprint arXiv:1903.10451*, 2019.

[10] M. Holters and U. Zölzer, "A generalized method for the derivation of non-linear state-space models from circuit schematics," in *2015 23rd European Signal Processing Conference (EUSIPCO)*, pp. 1073–1077, 2015.

[11] C.-W. Ho, A. Ruehli, and P. Brennan, "The modified nodal approach to network analysis," *IEEE Transactions on Circuits and Systems*, vol. 22, no. 6, pp. 504–509, 1975.

[12] T. Serafini and P. Zamboni, "State variable changes to avoid non computational issues," *Available on-line at http://www. simulanalog. org/statevariable. pdf (checked April 15, 2020)*.

[13] K. J. Werner, M. J. Olsen, M. Rest, and J. Parker, "Generalizing root variable choice in wave digital filters with grouped nonlinearities," in *Proc. 20th Int. Conf. Digit. Audio Effects, Edinburgh, UK*, 2017.

[14] P. C. Breedveld, "A systematic method to derive bond graph models," in *Proc. of the 2nd European Simulation Congress*, pp. 38–44, 1986.

[15] V. Acary, O. Bonnefon, and B. Brogliato, *Nonsmooth modeling and simulation for switched circuits*, vol. 69. Springer Science & Business Media, 2010.

[16] L. O. Chua, C. A. Desoer, and E. S. Kuh, *Linear and nonlinear circuits*. 1987.

[17] B. Jacob and H. J. Zwart, *Linear port-Hamiltonian systems on infinite-dimensional spaces*, vol. 223. Springer Science & Business Media, 2012.

[18] P. L. Dragotti, M. Vetterli, and T. Blu, "Sampling moments and reconstructing signals of finite rate of innovation: Shannon meets strang–fix," *IEEE Transactions on Signal Processing*, vol. 55, no. 5, pp. 1741–1757, 2007.

[19] W. Tang and Y. Sun, "Time finite element methods: A unified framework for numerical discretizations of ODEs," *Applied Mathematics and Computation*, vol. 219, no. 4, pp. 2158–2179, 2012.

[20] P. Deuflhard, *Newton methods for nonlinear problems: Affine invariance and adaptive algorithms*, vol. 35. Springer, 2011.

[21] M. Hintermüller, "Semismooth Newton methods and applications," tech. rep., Department of Mathematics, Humboldt-University of Berlin, 2010.

[22] J. Najnudel, T. Hélie, H. Boutin, D. Roze, T. Maniguet, and S. Vaiedelich, "Analog circuits and port-Hamiltonian realizability issues: A resolution method for simulations via equivalent components," in *Audio Engineering Society Convention 145*, 2018.

[23] J. Najnudel, T. Hélie, and D. Roze, "Simulation of the ondes martenot ribbon-controlled oscillator using energy-balanced modeling of nonlinear time-varying electronic components," *Journal of the Audio Engineering Society*, vol. 67, no. 12, pp. 961–971, 2019.

[24] F. Fontana and E. Bozzo, "Newton–raphson solution of nonlinear delay-free loop filter networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 10, pp. 1590–1600, 2019.

A. PROOF OF PROPERTY 5.1

The proof of Equation (41) is available in [8] and is not reproduced here. To prove its extension to semi-continuous functions, using left and right Taylor series expansion one finds

$$\begin{aligned}\lim_{\delta x\to 0}\overline{\mathbf{f}}_0(\mathbf{x}) &= \lim_{\delta x\to 0}\frac{F(\bar{x}+\frac{\delta x}{2})-F(\bar{x}-\frac{\delta x}{2})}{\delta x}\\ &= \lim_{\delta x\to 0}\frac{f(\bar{x}+\frac{\delta x}{2})\frac{\delta x}{2}+f(\bar{x}-\frac{\delta x}{2})\frac{\delta x}{2}+\mathcal{O}(|\delta x|^2)}{\delta x}\\ &= \frac{f(\bar{x}^+)+f(\bar{x}^-)}{2}.\end{aligned}$$

For the second coefficient, one finds $\|\ell_1\|^2 = 1/12$ and using integration by parts, one gets the recursive relation

$$\begin{aligned}\overline{\mathbf{f}}_1(\mathbf{x}) &= \int_0^1 \ell_1(\tau)f(x(\tau))\,\mathrm{d}\tau = \frac{1}{\delta x}\int_0^1 \ell_1(\tau)\,(F\circ x)'\,(\tau)\,\mathrm{d}\tau\\ &= \frac{1}{\delta x}\left(\left[\ell_1(\tau)(F\circ x)(\tau)\right]_0^1 - \int_0^1 (F\circ x)\,(\tau)\,\mathrm{d}\tau\right)\\ &= \frac{1}{\delta x}\left(\frac{F(x_1)+F(x_0)}{2} - \overline{\mathbf{F}}_0(\mathbf{x})\right).\end{aligned}$$

Finally, when $\delta x = 0$, one finds

$$\overline{\mathbf{f}}_1(\mathbf{x}) = \int_0^1 \ell_1(\tau)f(\bar{x})\,\mathrm{d}\tau = f(\bar{x})\int_0^1 \ell_1(\tau)\,\mathrm{d}\tau = 0.$$

B. LINEAR DIFFERENTIAL COMPONENTS

When $\nabla H(\boldsymbol{x}) = \mathbf{W}\boldsymbol{x}$ with state $\dot{\boldsymbol{x}}(\tau)\in\mathbb{P}^1$ and coefficients $\dot{\mathbf{x}}$ expressed in the *orthonormal* Legendre basis $\{L_i\}$, the projected gradient is

$$\overline{\nabla}H(\mathbf{x}_0,\dot{\mathbf{x}}) = \mathbf{W}\otimes\left(\left[\begin{smallmatrix}\mathbf{x}_0\\ 0\end{smallmatrix}\right] + h\left[\begin{smallmatrix}1/2 & -\sqrt{3}/6\\ \sqrt{3}/6 & 0\end{smallmatrix}\right]\dot{\mathbf{x}}\right).$$

For $\dot{\boldsymbol{x}}(\tau)\in\mathbb{P}^0$ it reduces to the midpoint integration rule

$$\overline{\nabla}H(\mathbf{x}_0,\dot{\mathbf{x}}) = \mathbf{W}\left(\mathbf{x}_0 + \frac{h}{2}\dot{\mathbf{x}}\right).$$

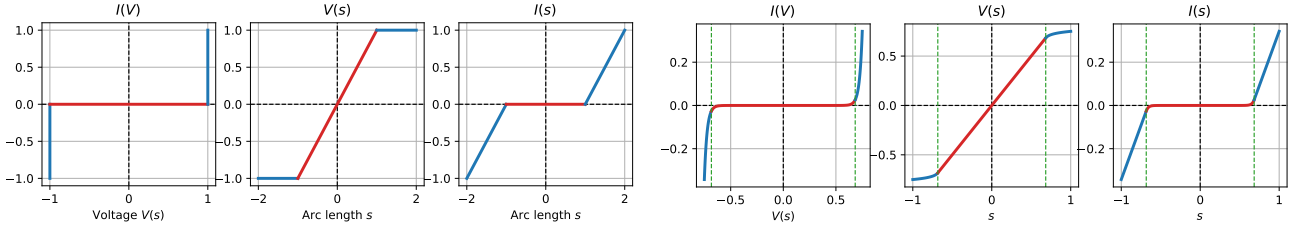
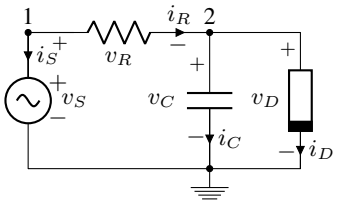
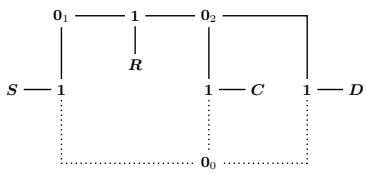


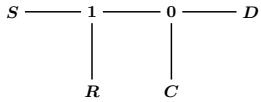
Figure 4: *(Pseudo) Arc-length parametrization of hard clipping resistor and anti-parallel Shockley diodes.*



(a) Diode Clipper circuit



(b) Equivalent Bondgraph



(c) Reduced Bondgraph

D	0	1	2
S	-1	1	0
R	0	1	-1
C	-1	0	1
D	-1	0	1

(d) (Reduced) Incidence Matrix

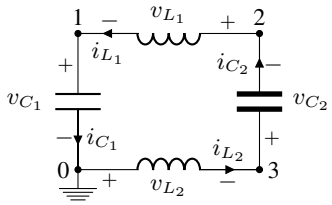
	e_1	e_2	i_S	i_R	i_C	i_D
$i_1 = 0$	.	.	-1	-1	0	0
$i_2 = 0$	.	.	0	1	-1	-1
v_S	1	0	.	.	.	.
v_R	1	-1	.	.	.	.
v_C	0	1	.	.	.	.
v_D	0	1	.	.	.	.

(e) Kirchhoff Dirac structure

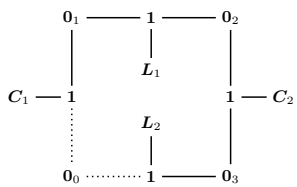
	v_S	v_C	i_R	i_D
i_S	.	.	-1	0
i_C	.	.	1	-1
v_R	1	-1	.	.
v_D	0	1	.	.

(f) Reduced Hybrid Dirac structure

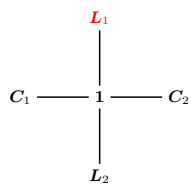
Figure 5: *Diode Clipper circuit: From the schematic (a) Kirchhoff laws immediately yield the bond-graph (b) which can be reduced to the bond-graph c). Using the Graph incidence matrix d), one obtains the Kirchhoff-Dirac structure e). Elimination of the node voltages yields the reduced Dirac structure f).*



(a) LCLC circuit



(b) Equivalent Bondgraph



(c) Reduced Bondgraph

D	0	1	2	3
C_1	-1	1	0	0
L_1	0	-1	1	0
C_2	0	0	-1	1
L_2	1	0	0	-1

(d) (Reduced) Incidence Matrix

	e_1	e_2	e_3	i_S	i_R	i_C	i_D
$i_1 = 0$	.	.	.	-1	1	0	0
$i_2 = 0$	.	.	.	0	-1	1	0
$i_3 = 0$	.	.	.	0	0	-1	1
v_{C_1}	1	0	0	.	.	.	.
v_{L_1}	-1	1	0	.	.	.	.
v_{C_2}	0	-1	1	.	.	.	.
v_{L_2}	0	0	-1	.	.	.	.

(e) Kirchhoff Dirac structure

	v_{C_1}	v_{L_1}	v_{C_2}	i_{L_2}
i_{C_1}	.	.	.	1
i_{L_1}	.	.	.	1
i_{C_2}	.	.	.	1
v_{L_2}	-1	-1	-1	.

(f) Reduced Hybrid Dirac structure

Figure 6: *Conservative LCLC circuit (a single cell of a transmission line). There is an apparent* computational causality *conflict shown in red on subfigure f): the loop current can either be controlled by L_1 or L_2 but not by both. The circuit has thus an implicit constraint $I_{L_1}(\phi_1) = I_{L_2}(\phi_2)$. The inductor L_1 is said to have a* differential causality *since $v_{L_1} = \dot{\phi}_1$, whereas C_1, C_2, L_2 are said to have an* integral causality.

References

[AB90] M. Alexander and D. F. Bowers, "SPICE-compatible op. amp. macro-models," *Analog Devices, Application Note, AN-138*, 1990.

[AB03a] J. S. Abel and D. P. Berners, "Discrete-time shelf filter design for analog modeling," *Journal of the Audio Engineering Society*, october 2003.

[AB03b] ——, "On peak-detecting and RMS feedback and feedforward compressors," *Journal of the Audio Engineering Society*, october 2003.

[AB08] V. Acary and B. Brogliato, *Numerical methods for nonsmooth dynamical systems: applications in mechanics and electronics.* Springer Science & Business Media, 2008.

[Aba14] L. L. Abath, "Circuit theory via algebraic topology," Master's thesis, Universidade Federal de Pernambuco, 2014.

[ABI19] P. Amodio, L. Brugnano, and F. Iavernaro, "A note on the continuous-stage Runge–Kutta (–Nyström) formulation of Hamiltonian Boundary Value Methods (HBVMs)," *Applied Mathematics and Computation*, vol. 363, p. 124634, 2019.

[AC12] J. P. Aubin and A. Cellina, *Differential inclusions: set-valued maps and viability theory.* Springer Science & Business Media, 2012, vol. 264.

[Alb20] D. Albertini, "Antiderivative antialiasing in nonlinear wave digital filters," 2020.

[AMH10] A. H. Al-Mohy and N. J. Higham, "A new scaling and squaring algorithm for the matrix exponential," *SIAM Journal on Matrix Analysis and Applications*, vol. 31, no. 3, pp. 970–989, 2010.

[AOI07] J.-J. Aucouturier, Y. Ogai, and T. Ikegami, "Making a robot dance to music using chaotic itinerancy in a network of fitzhugh-nagumo neurons," in *International Conference on Neural Information Processing.* Springer, 2007, pp. 647–656.

[Art06] R. Arthan, "A minimalist construction of the geometric algebra," *arXiv preprint math/0607190*, 2006.

[Aub11] J. P. Aubin, *Applied functional analysis.* John Wiley & Sons, 2011, vol. 47.

[BAC06] S. Bilbao, K. Arcas, and A. Chaigne, "A physical model for plate reverberation," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 5, 2006, pp. V–V.

[BB93] M. Borri and C. Bottasso, "A general framework for interpreting time finite element formulations," *Computational Mechanics*, vol. 13, no. 3, pp. 133–142, 1993.

[BC19] J. Bensoam and P. Carré, "Geometric numerical methods with Lie groups," in *International Conference on Geometric Science of Information.* Springer, 2019, pp. 75–84.

[BCD84] S. Boyd, L. O. Chua, and C. A. Desoer, " Analytical Foundations of Volterra Series," *IMA Journal of Mathematical Control and Information*, vol. 1, no. 3, pp. 243–282, 09 1984.

[BCL99] H. Brezis, P. G. Ciarlet, and J. L. Lions, *Analyse fonctionnelle: théorie et applications.* Dunod Paris, 1999, vol. 91.

[BDPR00] G. Borin, G. De Poli, and D. Rocchesso, "Elimination of delay-free loops in discrete-time models of nonlinear acoustic systems," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 5, pp. 597–605, 2000.

[Bel68] V. Belevitch, "Classical network theory," *Holden-day*, vol. 7, 1968.

[BEPV17] S. Bilbao, F. Esqueda, J. Parker, and V. Välimäki, "Antiderivative antialiasing for memoryless nonlinearities," *IEEE Signal Processing Letters*, vol. 24, no. 7, pp. 1049–1053, 2017.

[Ber70] S. Bergman, *The kernel function and conformal mapping.* American Mathematical Soc., 1970, vol. 5.

[BEV17] S. Bilbao, F. Esqueda, and V. Valimaki, "Antiderivative antialiasing, Lagrange interpolation and spectral flatness," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA).* IEEE, 2017, pp. 141–145.

[BFCI14] L. Brugnano, G. Frasca Caccia, and F. Iavernaro, "Hamiltonian boundary value methods (HBVMs) and their efficient implementation," *Math. Eng. Sci. Aerosp*, vol. 5, pp. 343–411, 2014.

[BG08] J. C. Butcher and N. Goodwin, *Numerical methods for ordinary differential equations.* Wiley Online Library, 2008, vol. 2.

[BHR18] D. Bouvier, T. Hélie, and D. Roze, "Homophase signals separation for volterra series identification," in *2018 IEEE Conference on Decision and Control (CDC).* IEEE, 2018, pp. 3854–3861.

[Bil04] S. Bilbao, *Wave and scattering methods for numerical simulation.* John Wiley & Sons, 2004.

[Bil05] ——, "Conservative numerical methods for nonlinear strings," *The Journal of the Acoustical Society of America*, vol. 118, no. 5, pp. 3316–3327, 2005.

[Bil08] ——, "A family of conservative finite difference schemes for the dynamical von karman plate equations," *Numerical Methods for Partial Differential Equations: An International Journal*, vol. 24, no. 1, pp. 193–216, 2008.

[Bil09] ——, *Numerical sound synthesis.* John Wiley & Sons, 2009.

[BM64a] R. Brayton and J. Moser, "A theory of nonlinear networks. i," *Quarterly of Applied Mathematics*, vol. 22, no. 1, pp. 1–33, 1964.

[BM64b] ——, "A theory of nonlinear networks. ii," *Quarterly of applied mathematics*, vol. 22, no. 2, pp. 81–104, 1964.

[BMBM18] A. M. Badlyan, B. Maschke, C. Beattie, and V. Mehrmann, "Open physical systems: from GENERIC to port-Hamiltonian systems," *arXiv preprint arXiv:1804.04064*, 2018.

[BMS20] A. Bernardini, P. Maffezzoni, and A. Sarti, "Vector wave digital filters and their application to circuits with two-port elements," *IEEE Transactions on Circuits and Systems I: Regular Papers*, pp. 1–14, 2020.

[Bog12] J. Bognár, *Indefinite inner product spaces.* Springer Science & Business Media, 2012, vol. 78.

[Boh88] D. A. Bohn, "Operator adjustable equalizers: An overview," in *Audio Engineering Society Conference: 6th International Conference: Sound Reinforcement.* Audio Engineering Society, 1988.

[Bor09] W. Borutzky, *Bond graph methodology: development and analysis of multidisciplinary dynamic system models.* Springer Science & Business Media, 2009.

[BOS05] H. Berland, B. Owren, and B. Skaflestad, "B-series and order conditions for exponential integrators," *SIAM Journal on Numerical Analysis*, vol. 43, no. 4, pp. 1715–1727, 2005.

[Bot97] C. L. Bottasso, "A new look at finite elements in time: a variational interpretation of Runge–Kutta methods," 1997.

[Bou18] D. Bouvier, "Identification de systèmes non linéaires représentés en séries de volterra: applications aux systèmes sonores," Ph.D. dissertation, Sorbonne Université/Université Pierre et Marie Curie-Paris VI, 2018.

[Boy01] J. P. Boyd, *Chebyshev and Fourier spectral methods.* Courier Corporation, 2001.

[BP10] S. Bilbao and J. Parker, "A virtual model of spring reverberation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 4, pp. 799–808, 2010.

[BPCS74] G. R. Boyle, D. O. Pederson, B. M. Cohn, and J. E. Solomon, "Macromodeling of integrated circuit operational amplifiers," *IEEE Journal of Solid-State Circuits*, vol. 9, no. 6, pp. 353–364, 1974.

[Bre85] P. C. Breedveld, "Multibond graph elements in physical systems theory," *Journal of the Franklin Institute*, vol. 319, no. 1-2, pp. 1–36, 1985.

[Bre86] ——, "A systematic method to derive bond graph models," in *Second European Simulation Congress, Antwerp, Belgium*, 1986.

[Bro99a] J. F. Broenink, "20-sim software for hierarchical bond-graph/block-diagram models," *Simulation Practice and Theory*, vol. 7, no. 5-6, pp. 481–492, 1999.

[Bro99b] ——, "Introduction to physical systems modelling with bond graphs," *SiE whitebook on simulation methodologies*, vol. 31, 1999.

[BS00] P. Betsch and P. Steinmann, "Inherently energy conserving time finite elements for classical mechanics," *Journal of Computational Physics*, vol. 160, no. 1, pp. 88–116, 2000.

[BS16] A. Bernardini and A. Sarti, "Dynamic adaptation of instantaneous nonlinear bipoles in wave digital networks," in *2016 24th European Signal Processing Conference (EUSIPCO).* IEEE, 2016, pp. 1038–1042.

[BS17] ——, "Biparametric wave digital filters," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 64, no. 7, pp. 1826–1838, 2017.

[BTC83] S. Boyd, Y. Tang, and L. Chua, "Measuring volterra kernels," *IEEE Transactions on Circuits and Systems*, vol. 30, no. 8, pp. 571–577, 1983.

[BTI09] L. Brugnano, D. Trigiante, and F. Iavernaro, "Analysis of Hamiltonian boundary value methods (HBVMs) for the numerical solution of polynomial Hamiltonian dynamical systems," Tech. Rep., 2009.

[BTU04] T. Blu, P. Thévenaz, and M. Unser, "Linear interpolation revitalized," *IEEE Transactions on Image Processing*, vol. 13, no. 5, pp. 710–719, 2004.

[BU99] T. Blu and M. Unser, "Quantitative Fourier analysis of approximation techniques. I. Interpolators and projectors," *IEEE Transactions on signal processing*, vol. 47, no. 10, pp. 2783–2795, 1999.

[But30] S. Butterworth, "On the theory of filter amplifiers," *Wireless Engineer*, vol. 7, no. 6, pp. 536–541, 1930.

[But72] J. C. Butcher, "An algebraic theory of integration methods," *Mathematics of Computation*, vol. 26, no. 117, pp. 79–106, 1972.

[But10] J. Butcher, "Trees, B-series and exponential integrators," *IMA journal of numerical analysis*, vol. 30, no. 1, pp. 131–140, 2010.

[BVS20] A. Bernardini, A. E. Vergani, and A. Sarti, "Wave digital modeling of nonlinear 3-terminal devices for virtual analog applications," *Circuits, Systems, and Signal Processing*, pp. 1–31, 2020.

[BW17] Ó. Bogason and K. J. Werner, "Modeling circuits with operational transconductance amplifiers using wave digital filters," in *Proc. 20th Int. Conf. Digital Audio Effects, Edinburgh, UK*, 2017, pp. 130–137.

[Car64] H. Carlin, "Singular network elements," *IEEE Transactions on circuit theory*, vol. 11, no. 1, pp. 67–72, 1964.

[Car67] ——, "On the existence of a scattering representation for passive networks," *IEEE Transactions on Circuit Theory*, vol. 14, no. 4, pp. 418–419, 1967.

[Car20] A. Carson, "Aliasing reduction in virtual analogue modelling," Master's thesis, 09 2020.

[CB01] B. Carter and T. R. Brown, *Handbook of operational amplifier applications.* Texas Instruments Dallas, Tex, USA, 2001.

[CB17] P. Carré and J. Bensoam, "Intégrateurs multisymplectiques par action d'un groupe de Lie: test de méthodes numériques HPC sur des systemes completement intégrables." 2017.

[CB19] ——, "Geometric numerical methods for mechanics," in *Congrès Français de Mécanique*, 2019.

[CC76] L. O. Chua and L.-K. Chen, "Diakoptic and generalized hybrid analysis," *IEEE Transactions on Circuits and Systems*, vol. 23, no. 12, pp. 694–705, 1976.

[CDK87] L. O. Chua, C. A. Desoer, and E. S. Kuh, *Linear and nonlinear circuits*, 1987.

[Cel] E. Celledoni, "Energy preservation (dissipation) in numerical ODEs and PDEs."

[Cel03] . O. B. Celledoni, E., "Lie group methods for rigid body dynamics and time integration on manifolds," *Computer Methods in Applied Mechanics and Engineering*, 2003.

[CGM+12] E. Celledoni, V. Grimm, R. I. McLachlan, D. I. McLaren, D. O'Neale, B. Owren, and G. Quispel, "Preserving energy resp. dissipation in numerical PDEs using the "average vector field" method," *Journal of Computational Physics*, vol. 231, no. 20, pp. 6770–6789, 2012.

[CH17] E. Celledoni and E. H. Høiseth, "Energy-preserving and passivity-consistent numerical discretization of port-Hamiltonian systems," *arXiv preprint arXiv:1706.08621*, 2017.

[Cha99] F. Chaplais, "The Strang and Fix conditions," 1999.

[Che51] C. Cherry, "Some general theorems for non-linear systems possessing reactance," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 42, no. 333, pp. 1161–1177, 1951.

[Chr16] O. Christensen, *An introduction to frames and Riesz bases.* Springer, 2016.

[Chu75] L. O. Chua, "Computer-aided analysis of electronic circuits," *Algorithms and computational techniques*, 1975.

[CI01] E. Celledoni and A. Iserles, "Methods for the approximation of the matrix exponential in a Lie-algebraic setting," *IMA Journal of Numerical Analysis*, vol. 21, no. 2, pp. 463–488, 2001.

[Cie13] J. L. Cieśliński, "Locally exact modifications of numerical schemes," *Computers & Mathematics with Applications*, vol. 65, no. 12, pp. 1920–1938, 2013.

[Cie14] ——, "Improving the accuracy of the AVF method," *Journal of Computational and Applied Mathematics*, vol. 259, pp. 233–243, 2014.

[CMM+09] E. Celledoni, R. I. McLachlan, D. I. McLaren, B. Owren, G. R. W. Quispel, and W. M. Wright, "Energy-preserving runge-kutta methods," *ESAIM: Mathematical Modelling and Numerical Analysis*, vol. 43, no. 4, pp. 645–649, 2009.

[CMOQ10] E. Celledoni, R. I. McLachlan, B. Owren, and G. Quispel, "Energy-preserving integrators and the structure of B-series," *Foundations of Computational Mathematics*, vol. 10, no. 6, pp. 673–693, 2010.

[COCR09] M. Conti, S. Orcioni, M. Caldari, and F. Ripa, "Real time implementation of fuzz-face electric guitar effect," in *Intelligent Technical Systems*. Springer, 2009, pp. 89–100.

[Coh12] I. Cohen, "Modélisation, analyse et identification de circuits non linéaires: application aux amplificateurs guitare à lampes pour la simulation en temps réel," Ph.D. dissertation, Paris 6, 2012.

[Col08] N. Collins, "Errant sound synthesis," in *ICMC*, 2008.

[COS14] E. Celledoni, B. Owren, and Y. Sun, "The minimal stage, energy preserving Runge–Kutta method for polynomial Hamiltonian systems is the averaged vector field method," *Mathematics of Computation*, vol. 83, no. 288, pp. 1689–1700, 2014.

[Cou90] T. J. Courant, "Dirac manifolds," *Transactions of the American Mathematical Society*, vol. 319, no. 2, pp. 631–661, 1990.

[CR10] J. L. Cieśliński and B. Ratkiewicz, "Improving the accuracy of the discrete gradient method in the one-dimensional case," *Physical Review E*, vol. 81, no. 1, p. 016704, 2010.

[CR11] ——, "Energy-preserving numerical schemes of high accuracy for one-dimensional hamiltonian systems," *Journal of Physics A: Mathematical and Theoretical*, vol. 44, no. 15, p. 155206, 2011.

[CvW15a] V. Chatziioannou and M. van Walstijn, "Discrete-time conserved quantities for damped oscillators," *Proc. Third Vienna Talk on Music Acoustics*, pp. 135–139, 2015.

[CvW15b] ——, "Energy conserving schemes for the simulation of musical instrument contact dynamics," *Journal of Sound and Vibration*, vol. 339, pp. 262–279, 2015.

[CZ12] R. F. Curtain and H. Zwart, *An introduction to infinite-dimensional linear systems theory*. Springer Science & Business Media, 2012, vol. 21.

[D'A14] S. D'Angelo, "Virtual analog modeling of nonlinear musical circuits," Ph.D. dissertation, Aalto University, 2014.

[DBT96] G. Di Battista and R. Tamassia, "On-line maintenance of triconnected components with SPQR-trees," *Algorithmica*, vol. 15, no. 4, pp. 302–318, 1996.

[Deo17] N. Deo, *Graph theory with applications to engineering and computer science*. Courier Dover Publications, 2017.

[Deu87] P. Deuflhard, "Uniqueness theorems for stiff ODE initial value problems." *Proceedings of the 13th Biennial Conference on Numerical Analysis 1989*, pp. 74–205, 1987.

[Deu11] ——, *Newton methods for nonlinear problems: affine invariance and adaptive algorithms.* Springer Science & Business Media, 2011, vol. 35.

[Dev19] A. Devices, "LT1366 datasheet," https://www.analog.com/en/products/lt1366.html, 2019, online; accessed: 2019-03-22.

[DGL+03] C. Doran, S. R. Gullans, A. Lasenby, J. Lasenby, and W. Fitzgerald, *Geometric algebra for physicists.* Cambridge University Press, 2003.

[DHR19] B. Damien, T. Hélie, and D. Roze, "Phase-based order separation for Volterra series identification," *International Journal of Control*, pp. 1–11, Nov. 2019.

[DHSVA93] C. Doran, D. Hestenes, F. Sommen, and N. Van Acker, "Lie groups as spin groups," *Journal of Mathematical Physics*, vol. 34, no. 8, pp. 3642–3669, 1993.

[DHZ10] K. Dempwolf, M. Holters, and U. Zölzer, "Discretization of parametric analog circuits for real-time simulations," in *Proceedings of the 13th international conference on digital audio effects (DAFx-10)*, 2010.

[DHZ11] ——, "A triode model for guitar amplifier simulation with individual parameter fitting," in *Audio Engineering Society Convention 131.* Audio Engineering Society, 2011.

[dLVR13] I. G. de La Vega and R. Riaza, "Hybrid analysis of nonlinear circuits: DAE models with indices zero and one," *Circuits, Systems, and Signal Processing*, vol. 32, no. 5, pp. 2065–2095, 2013.

[DMSB09] V. Duindam, A. Macchelli, S. Stramigioli, and H. Bruyninckx, *Modeling and control of complex physical systems: the port-Hamiltonian approach.* Springer Science & Business Media, 2009.

[DSS09] G. De Sanctis and A. Sarti, "Virtual analog modeling in the wave-digital domain," *IEEE transactions on audio, speech, and language processing*, vol. 18, no. 4, pp. 715–727, 2009.

[DVB07] P. L. Dragotti, M. Vetterli, and T. Blu, "Sampling moments and reconstructing signals of finite rate of innovation: Shannon meets Strang–Fix," *IEEE Transactions on signal processing*, vol. 55, no. 5, pp. 1741–1757, 2007.

[DZ11a] K. Dempwolf and U. Zölzer, "Discrete state-space model of the fuzz-face," in *Proceedings of Forum Acusticum, Aalborg, Denmark*, 2011.

[DZ11b] ——, "A physically-motivated triode model for circuit simulations," *Proc. Digital Audio Effects (DAFx-11), Paris, France*, 2011.

[Eck02] D. S. Eck, "Real-time musical beat induction with spiking neural networks," Citeseer, Tech. Rep., 2002.

[EFHZ14] F. Eichas, M. Fink, M. Holters, and U. Zölzer, "Physical modeling of the mxr phase 90 guitar effect pedal." in *DAFx*, 2014, pp. 153–158.

[EGZ17] F. Eichas, E. Gerat, and U. Zölzer, "Virtual analog modeling of dynamic range compression systems," in *Audio Engineering Society Convention 142.* Audio Engineering Society, 2017.

[Ehl69] B. L. Ehle, "On Padé approximations to the exponential function and A-stable methods for the numerical solution of initial value problems," Ph.D. dissertation, University of Waterloo Waterloo, Ontario, 1969.

[Ele20a] ElectroSmash, "Fuzz face analysis," https://www.electrosmash.com/fuzz-face, 2020, [Online; accessed 23-November-2020].

[Ele20b] ——, "Tube screamer analysis," https://www.electrosmash.com/tube-screamer-analysis, 2020, [Online; accessed 23-November-2020].

[EMVDS06] D. Eberard, B. Maschke, and A. Van Der Schaft, "Energy-conserving formulation of RLC-circuits with linear resistors," in *Proc. 17th International Symposium on Mathematical Theory of Networks and systems*, 2006.

[EPPB17a] F. Esqueda, H. Pöntynen, J. D. Parker, and S. Bilbao, "Virtual analog models of the Lockhart and Serge wavefolders," *Applied Sciences*, vol. 7, no. 12, p. 1328, 2017.

[EPPB17b] ——, "Virtual analog models of the Lockhart and Serge wavefolders," *Applied Sciences*, vol. 7, no. 12, p. 1328, 2017.

[EST00] D. Estévez Schwarz and C. Tischendorf, "Structural analysis of electric circuits and consequences for MNA," *International Journal of Circuit Theory and Applications*, vol. 28, no. 2, pp. 131–162, 2000.

[EZ16] F. Eichas and U. Zölzer, "Black-box modeling of distortion circuits with block-oriented models," in *Proc. Int. Conf. Digital Audio Effects (DAFx-16), Brno, Czech Republic*, 2016.

[EZ18] ——, "Virtual analog modeling of guitar amplifiers with Wiener–Hammerstein models," in *44th Annual Convention on Acoustics (DAGA 2018)*, 2018.

[FAC63] R. M. Fano, R. B. Adler, and L. J. Chu, *Electromagnetic fields, energy, and forces*. Taylor & Francis, 1963.

[Fal16] A. Falaize, "Modélisation, simulation, génération de code et correction de systèmes multi-physiques audios: Approche par réseau de composants et formulation Hamiltonienne à ports," Ph.D. dissertation, 2016.

[Far12] R. T. Farouki, "The Bernstein polynomial basis: A centennial retrospective," *Computer Aided Geometric Design*, vol. 29, no. 6, pp. 379–419, 2012.

[Fet86] A. Fettweis, "Wave digital filters: Theory and practice," *Proceedings of the IEEE*, vol. 74, no. 2, pp. 270–327, 1986.

[FH13] A. Falaize and T. Hélie, "Simulation of an analog circuit of a wah pedal: a port-Hamiltonian approach," in *Audio Engineering Society Convention 135*. Audio Engineering Society, 2013.

[FH16a] ——, "Passive guaranteed simulation of analog audio circuits: A port-Hamiltonian approach," *Applied Sciences*, vol. 6, no. 10, p. 273, 2016.

[FH16b] ——, "PyPHS: Passive modeling and simulation in python," 2016.

[FH17] ——, "Passive simulation of the nonlinear port-Hamiltonian modeling of a rhodes piano," *Journal of Sound and Vibration*, vol. 390, pp. 289–309, 2017.

[FH20] ——, "Passive modelling of the electrodynamic loudspeaker: from the Thiele–Small model to nonlinear port-Hamiltonian systems," *Acta Acustica*, vol. 4, no. 1, p. 1, 2020.

[Fit55] R. FitzHugh, "Mathematical models of threshold phenomena in the nerve membrane," *The bulletin of mathematical biophysics*, vol. 17, no. 4, pp. 257–278, 1955.

[FOO05] D. Franken, J. Ochs, and K. Ochs, "Generation of wave digital structures for networks containing multiport elements," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 52, no. 3, pp. 586–596, 2005.

[FS69] G. Fix and G. Strang, "Fourier analysis of the finite element method in Ritz–Galerkin theory," *Studies in Applied mathematics*, vol. 48, no. 3, pp. 265–273, 1969.

[Gau81] W. Gautschi, "A survey of Gauss-Christoffel quadrature formulae," in *E.B. Christoffel*. Springer, 1981, pp. 72–147.

[GBJR20] M. Günther, A. Bartel, B. Jacob, and T. Reis, "Dynamic iteration schemes and port-Hamiltonian formulation in coupled DAE circuit simulation," *arXiv preprint arXiv:2004.12951*, 2020.

[GE13] D. J. Gillespie and D. P. Ellis, "Modeling nonlinear circuits with linearized dynamical models via kernel regression," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 2013, pp. 1–4.

[Gea71] C. Gear, "Simultaneous numerical solution of differential-algebraic equations," *IEEE transactions on circuit theory*, vol. 18, no. 1, pp. 89–95, 1971.

[GEPP18] G. Gormond, F. Esqueda, H. Pöntynen, and J. Parker, "Waveshaping with Norton amplifiers: modeling the Serge triple waveshaper," in *International Conference on Digital Audio Effects*, 2018, pp. 288–295.

[GEZ17] E. Gerat, F. Eichas, and U. Zölzer, "Virtual analog modeling of a UREI 1176ln dynamic range control system," *Journal of the Audio Angineering Society*, october 2017.

[GHVdSR20] H. Gernandt, F. Haller, A. Van der Schaft, and T. Reis, "Port-Hamiltonian formulation of nonlinear electrical circuits," *arXiv preprint arXiv:2004.10821*, 2020.

[GLD93] S. Gull, A. Lasenby, and C. Doran, "Imaginary numbers are not real, the geometric algebra of spacetime," *Foundations of Physics*, vol. 23, no. 9, pp. 1175–1201, 1993.

[GMR12] D. Giannoulis, M. Massberg, and J. D. Reiss, "Digital dynamic range compressor design - a tutorial and analysis," *Journal of the audio engineering society*, vol. 60, no. 6, pp. 399–408, june 2012.

[Gou20] V. Goudard, "Représentation et contrôle dans le design interactif des instruments de musique numériques," Ph.D. dissertation, Sorbonne Université, 2020.

[GP70] H. K. Gummel and H. Poon, "An integral charge control model of bipolar transistors," *Bell System Technical Journal*, vol. 49, no. 5, pp. 827–852, 1970.

[Gua11] M. Gualtieri, "Generalized complex geometry," *Annals of mathematics*, pp. 75–123, 2011.

[GVdSBM03] G. Golo, A. Van der Schaft, P. Breedveld, and B. Maschke, "Hamiltonian formulation of bond graphs," *Nonlinear and hybrid systems in automotive control*, pp. 351–372, 2003.

[Hai10] E. Hairer, "Energy-preserving variant of collocation methods," *Journal of Numerical Analysis, Industrial and Applied Mathematics*, vol. 5, pp. 73–84, 2010.

[Hai11] ——, "Solving differential equations on manifolds," *Lecture Notes, Université de Genève*, 2011.

[Ham07] E. Hammerich, "A generalized sampling theorem for frequency localized signals," *arXiv preprint arXiv:0707.0285*, 2007.

[Haw01] S. Hawking, *The universe in a nutshell*. Bantam, 2001.

[HBG71] G. Hachtel, R. Brayton, and F. Gustavson, "The sparse tableau approach to network analysis and design," *IEEE Transactions on circuit theory*, vol. 18, no. 1, pp. 101–113, 1971.

[HCB05] T. J. Hughes, J. A. Cottrell, and Y. Bazilevs, "Isogeometric analysis: CAD, finite elements, NURBS, exact geometry and mesh refinement," *Computer methods in applied mechanics and engineering*, vol. 194, no. 39-41, pp. 4135–4195, 2005.

[HDF+10] M. Heinrich, T. Dahms, V. Flunkert, S. W. Teitsworth, and E. Schöll, "Symmetry-breaking transitions in networks of nonlinear circuit elements," *New Journal of Physics*, vol. 12, no. 11, p. 113030, 2010.

[HDZ11] M. Holters, K. Dempwolf, and U. Zölzer, "A digital emulation of the boss sd-1 super overdrive pedal based on physical modeling," in *Audio Engineering Society Convention 131*. Audio Engineering Society, 2011.

[Hél09] T. Hélie, "Volterra series and state transformation for real-time simulations of audio circuits including saturations: Application to the Moog ladder filter," *IEEE transactions on audio, speech, and language processing*, vol. 18, no. 4, pp. 747–759, 2009.

[Hél11] ——, "Lyapunov stability analysis of the Moog ladder filter and dissipativity aspects in numerical solutions," in *Proceedings of the 14th International Conference on Digital Audio Effects DAFx-11, Paris, France*, 2011, pp. 19–23.

[Hes86] D. Hestenes, "A unified language for mathematics and physics," in *Clifford algebras and their applications in mathematical physics*. Springer, 1986, pp. 1–23.

[Hes93] ——, "Hamiltonian mechanics with geometric calculus," in *Spinors, Twistors, Clifford Algebras and Quantum Deformations*. Springer, 1993, pp. 203–214.

[Hes11] ——, "The shape of differential geometry in geometric calculus," in *Guide to Geometric Algebra in Practice*. Springer, 2011, pp. 393–410.

[Hes14] ——, "Tutorial on geometric calculus," *Advances in Applied Clifford Algebras*, vol. 24, no. 2, pp. 257–273, 2014.

[HHVW17] B. Holmes, M. Holters, and M. Van Walstijn, "Comparison of germanium bipolar junction transistor models for real-time circuit simulation," 2017.

[Hit01] E. M. Hitzer, "Antisymmetric matrices are real bivectors," *Mem. Fac. Eng. Fukui Univ.*, vol. 49, no. 2, 2001.

[HL14] E. Hairer and C. Lubich, "Energy-diminishing integration of gradient systems," *IMA Journal of Numerical Analysis*, vol. 34, no. 2, pp. 452–461, 2014.

[HLR06] E. Hairer, C. Lubich, and M. Roche, *The numerical solution of differential-algebraic systems by Runge-Kutta methods*. Springer, 2006, vol. 1409.

[HLW06] E. Hairer, C. Lubich, and G. Wanner, *Geometric numerical integration: structure-preserving algorithms for ordinary differential equations*. Springer Science & Business Media, 2006, vol. 31.

[HNW93] E. Hairer, S. P. Nørsett, and G. Wanner, *Solving ordinary differential equations I. Nonstiff problems*. Springer Series in Computational Mathematics, 1993.

[HO10] M. Hochbruck and A. Ostermann, "Exponential integrators." *Acta Numer.*, vol. 19, no. May, pp. 209–286, 2010.

[Hol16] M. Holmes, Ben et Van Walstijn, "Physical model parameter optimisation for calibrated emulation of the dallas rangemaster treble booster guitar pedal," in *Proc. Int. Conf. Digital Audio Effects (DAFx-16), Brno, Czech Republic*, 2016, pp. 47–54.

[Hol19] B. Holmes, "Guitar effects: pedal emulation and identification," Ph.D. dissertation, Queen's University Belfast, 2019.

[Hol20] M. Holters, "Antiderivative antialiasing for stateful systems," *Applied Sciences*, vol. 10, no. 1, p. 20, 2020.

[HP18] M. Holters and J. D. Parker, "A combined model for a bucket brigade device and its input and output filters," in *Proc. Int. Conf. Digital Audio Effects (DAFx-18), Aveiro, Portugal*, 2018.

[HRB75] C.-W. Ho, A. Ruehli, and P. Brennan, "The modified nodal approach to network analysis," *IEEE Transactions on circuits and systems*, vol. 22, no. 6, pp. 504–509, 1975.

[Hul92] G. M. Hulbert, "Time finite element methods for structural dynamics," *International Journal for Numerical Methods in Engineering*, vol. 33, no. 2, pp. 307–331, 1992.

[Huo04] A. Huovilainen, "Non-linear digital implementation of the Moog ladder filter," in *Proceedings of the International Conference on Digital Audio Effects (DAFx-04)*, 2004, pp. 61–64.

[Huo05] ——, "Enhanced digital models for analog modulation effects," in *Proc. Int. Conf. Digital Audio Effects (DAFx-05), Madrid, Spain*, 2005, pp. 155–160.

[HW96] E. Hairer and G. Wanner, *Solving ordinary differential equations II: Stiff and Differential-Algebraic Problems.* Springer series in computational mathematics, 1996.

[HZ11] M. Holters and U. Zölzer, "Physical modelling of a wah-wah effect pedal as a case study for application of the nodal dk method to circuits with variable parts," *Proc. Digital Audio Effects (DAFx-11), Paris, France*, 2011.

[HZ15] ——, "A generalized method for the derivation of non-linear state-space models from circuit schematics," in *2015 23rd European Signal Processing Conference (EUSIPCO).* IEEE, 2015, pp. 1073–1077.

[HZ16] ——, "Circuit simulation with inductors and transformers based on the Jiles-Atherton model of magnetization," in *19th International Conference on Digital Audio Effects (DAFx-16), Brno, Czech Republic*, 2016.

[IMKNZ00] A. Iserles, H. Z. Munthe-Kaas, S. P. Nørsett, and A. Zanna, "Lie-group methods," *Acta numerica*, vol. 9, pp. 215–365, 2000.

[Ise09] A. Iserles, *A first course in the numerical analysis of differential equations.* Cambridge university press, 2009, no. 44.

[Ise11] ——, "A fast and simple algorithm for the computation of legendre coefficients," *Numerische Mathematik*, vol. 117, no. 3, pp. 529–553, 2011.

[IT10] S. Iwata and M. Takamatsu, "Index minimization of differential-algebraic equations in hybrid analysis for circuit simulation," *Mathematical programming*, vol. 121, no. 1, pp. 105–121, 2010.

[ITT12] S. Iwata, M. Takamatsu, and C. Tischendorf, "Tractability index of hybrid equations for circuit simulation," *Mathematics of Computation*, vol. 81, no. 278, pp. 923–939, 2012.

[JDT+17] M. Jossic, V. Denis, O. Thomas, A. Mamou-Mani, B. Chomette, and D. Roze, "Active control of chinese gongs," *The Journal of the Acoustical Society of America*, vol. 141, no. 5, pp. 3619–3620, 2017.

[JL93] R.-Q. Jia and J. Lei, "A new version of the Strang–Fix conditions," *Journal of approximation theory*, vol. 74, no. 2, pp. 221–225, 1993.

[JRH+17] M. Jossic, D. Roze, T. Hélie, B. Chomette, and A. Mamou-Mani, "Energy shaping of a softening duffing oscillator using the formalism of port-Hamiltonian systems," in *20th International Conference on Digital Audio Effects (DAFx-17)*, 2017.

[JS03] D. Jeltsema and J. M. Scherpen, "A dual relation between port-Hamiltonian systems and the Brayton–Moser equations for nonlinear switched RLC circuits," *Automatica*, vol. 39, no. 6, pp. 969–979, 2003.

[JZ12] B. Jacob and H. J. Zwart, *Linear port-Hamiltonian systems on infinite-dimensional spaces.* Springer Science & Business Media, 2012, vol. 223.

[KG02] H. K. Khalil and J. W. Grizzle, *Nonlinear systems.* Prentice hall Upper Saddle River, NJ, 2002, vol. 3.

[KL19] P. Kotyczka and L. Lefèvre, "Discrete-time port-Hamiltonian systems: A definition based on symplectic integration," *Systems & Control Letters*, 2019.

[KM06] P. Kunkel and V. Mehrmann, *Differential-algebraic equations: analysis and numerical solution.* European Mathematical Society, 2006, vol. 2.

[KML18] P. Kotyczka, B. Maschke, and L. Lefèvre, "Weak form of Stokes–Dirac structures and geometric discretization of port-Hamiltonian systems," *Journal of Computational Physics*, vol. 361, pp. 442–476, 2018.

[KR65] E. S. Kuh and R. A. Rohrer, "The state-variable approach to network analysis," *Proceedings of the IEEE*, vol. 53, no. 7, pp. 672–686, 1965.

[KR68] D. Karnopp and R. C. Rosenberg, "Analysis and simulation of multiport systems: the bond graph approach to physical system dynamics," 1968.

[Kru56] J. B. Kruskal, "On the shortest spanning subtree of a graph and the traveling salesman problem," *Proceedings of the American Mathematical society*, vol. 7, no. 1, pp. 48–50, 1956.

[KZ17] S. Kraft and U. Zölzer, "LP-BLIT: Bandlimited impulse train synthesis of lowpass-filtered waveforms," 2017.

[LC12] C. Lederman and J.-L. Cambier, "Time-parallel solutions to ordinary differential equations on GPUs with a new functional optimization approach related to the Sobolev gradient method," Air force research lab Edwards AFB CA rocket propulsion dir, Tech. Rep., 2012.

[LH20] T. Lebrun and T. Hélie, "Correction of the doppler distortion generated by a vibrating baffled piston," *Acta Acustica*, vol. 4, no. 1, p. 2, 2020.

[LHF15] N. Lopes, T. Hélie, and A. Falaize, "Explicit second-order accurate method for the passive guaranteed simulation of port-Hamiltonian systems," *IFAC-PapersOnLine*, vol. 48, no. 13, pp. 223–228, 2015.

[Lig91] W. A. Light, "Recent developments in the Strang-Fix theory for approximation orders," in *Curves and surfaces.* Elsevier, 1991, pp. 285–292.

[LO13] V. T. Luan and A. Ostermann, "Exponential B-series: The stiff case," *SIAM Journal on Numerical Analysis*, vol. 51, no. 6, pp. 3431–3445, 2013.

[Lop16] N. Lopes, "Approche passive pour la modélisation, la simulation et l'étude d'un banc de test robotisé pour les instruments de type cuivre." Ph.D. dissertation, 2016.

[LWH+20] T. Lebrun, M. Wijnand, T. Hélie, D. Roze, and B. d'Andréa Novel, "Electroacoustic absorbers based on passive finite-time control of loudspeakers: a numerical investigation," in *Nonlinear Dynamics and Control.* Springer, 2020, pp. 23–31.

[Mac02] A. MacDonald, "An elementary construction of the geometric algebra," *Advances in applied Clifford algebras*, vol. 12, no. 1, pp. 1–6, 2002.

[Mac10] ——, *Linear and geometric algebra*, 2010.

[Mac12a] J. Macák, "Real-time digital simulation of guitar amplifiers as audio effects," Ph.D. dissertation, Ph. D. thesis, Brno University of Technology, Brno, 2012.

[Mac12b] A. MacDonald, *Vector and geometric calculus*, 2012.

[Mac16] J. Macák, "Simulation of analog flanger effect using bbd circuit," in *Proc. Int. Conf. Digital Audio Effects (DAFx-16), Brno, Czech Republic*, 2016, pp. 5–9.

[Mac17] A. MacDonald, "A survey of geometric algebra and geometric calculus," *Advances in Applied Clifford Algebras*, vol. 27, no. 1, pp. 853–891, 2017.

[Mak10] J. Maks, "A spinor approach to port-hamiltonian systems," in *Proceedings of the 19th International Symposium on Mathematical Theory of Networks and Systems–MTNS*, vol. 5, no. 9, 2010.

[Mar65] G. Martinelli, "On the nullor," *Proceedings of the IEEE*, vol. 53, no. 3, pp. 332–332, 1965.

[MB16] Y. Miyatake and J. C. Butcher, "A characterization of energy-preserving methods and the construction of parallel integrators for hamiltonian systems," *SIAM Journal on Numerical Analysis*, vol. 54, no. 3, pp. 1993–2013, 2016.

[Meh12] V. Mehrmann, "Index concepts for differential-algebraic equations," *Encyclopedia of Applied and Computational Mathematics*, vol. 1, pp. 676–681, 2012.

[MFS00] W. Marquis-Favre and S. Scavarda, "Alternative causality assignment procedures in bond graph language," in *ASME International Mechanical Engineering Congress and Exposition, Dynamic Systems and Control Division, Symposium on Automated Modeling*, vol. 2, 2000, pp. 817–824.

[MFS02] ——, "Alternative causality assignment procedures in bond graph for mechanical systems," *J. Dyn. Sys., Meas., Control*, vol. 124, no. 3, pp. 457–463, 2002.

[MH17] R. Müller and T. Hélie, "Trajectory anti-aliasing on guaranteed-passive simulation of nonlinear physical systems," in *Proc. of the 20th Int. Conference on Digital Audio Effects*, 2017.

[MH18] ——, "Power-balanced modelling of circuits as skew gradient systems," in *Proc. of the 21st Int. Conference on Digital Audio Effects*, 2018.

[MH19] ——, "A minimal passive model of the operational amplifier: Application to Sallen–Key analog filters," in *Proc. of the 22nd Int. Conference on Digital Audio Effects*, 2019.

[MH20] ——, "Fully-implicit algebro-differential parametrization of circuits," in *Proc. of the 23rd Int. Conference on Digital Audio Effects*, 2020.

[Mil51] W. Millar, "Some general theorems for non-linear systems possessing resistance," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 42, no. 333, pp. 1150–1160, 1951.

[MM10] G. Meinsma and L. Mirkin, "Sampling from a system-theoretic viewpoint: Part i-concepts and tools," *IEEE transactions on signal processing*, vol. 58, no. 7, pp. 3578–3590, 2010.

[MMMKV17] R. I. McLachlan, K. Modin, H. Munthe-Kaas, and O. Verdier, "Butcher series: a story of rooted trees and numerical methods for evolution equations," *Asia Pacific Mathematics Newsletter*, vol. 7, no. 1, pp. I–II, 2017.

[Mor86] P. J. Morrison, "A paradigm for joined hamiltonian and dissipative systems," *Physica D: Nonlinear Phenomena*, vol. 18, no. 1-3, pp. 410–419, 1986.

[MQR99] R. I. McLachlan, G. R. W. Quispel, and N. Robidoux, "Geometric integration using discrete gradients," *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, vol. 357, no. 1754, pp. 1021–1045, 1999.

[MQSW07] H. Ma, D. Qi, R. Song, and T. Wang, "The complete orthogonal V-system and its applications," *Communications on Pure & Applied Analysis*, vol. 6, no. 3, p. 853, 2007.

[MRBR20] M. A. Martínez Ramírez, E. Benetos, and J. D. Reiss, "Deep learning for black-box modeling of audio effects," *Applied Sciences*, vol. 10, no. 2, 2020.

[MV92] B. Maschke and A. Van der Schaft, "Port-controlled Hamiltonian systems: Modelling origins and system theoretic properties," *IFAC Proceedings Volumes*, vol. 25, no. 13, pp. 359–365, 1992, 2nd IFAC Symposium on Nonlinear Control Systems Design 1992, Bordeaux, France, 24-26 June.

[MVL78] C. Moler and C. Van Loan, "Nineteen dubious ways to compute the exponential of a matrix," *SIAM review*, vol. 20, no. 4, pp. 801–836, 1978.

[MVL03] ——, "Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later," *SIAM review*, vol. 45, no. 1, pp. 3–49, 2003.

[Nag75] L. W. Nagel, "Spice2: A computer program to simulate semiconductor circuits," *Ph. D. dissertation, University of California at Berkeley*, 1975.

[NAY62] J. Nagumo, S. Arimoto, and S. Yoshizawa, "An active pulse transmission line simulating nerve axon," *Proceedings of the IRE*, vol. 50, no. 10, pp. 2061–2070, 1962.

[Neu09] J. Neuberger, *Sobolev gradients and differential equations.* Springer Science & Business Media, 2009.

[Ng06] S. S. K. K. Ng, *Tunnel Devices.* John Wiley & Sons, Ltd, 2006, pp. 415–465.

[NH14] D. Nehab and H. Hoppe, *A Fresh Look at Generalized Sampling.* Now Foundations and Trends, 2014.

[NHB+18] J. Najnudel, T. Hélie, H. Boutin, D. Roze, T. Maniguet, and S. Vaiedelich, "Analog circuits and port-Hamiltonian realizability issues: a resolution method for simulations via equivalent components," in *Audio Engineering Society Convention 145.* Audio Engineering Society, 2018.

[NHRB20] J. Najnudel, T. Hélie, D. Roze, and H. Boutin, "Simulation of an ondes Martenot circuit," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 28, pp. 2651–2660, Sep. 2020.

[NMHR20] J. Najnudel, R. Müller, T. Hélie, and D. Roze, "A power-balanced dynamic model of ferromagnetic coils," in *23rd International Conference on Digital Audio Effects (eDAFx-20)*, Vienne, Austria, Sep. 2020.

[Nør74] S. P. Nørsett, "One-step methods of Hermite type for numerical integration of stiff systems," *BIT Numerical Mathematics*, vol. 14, no. 1, pp. 63–77, 1974.

[Obr40] N. Obreshkov, *Neue quadraturformeln.* Verlag der Akademie der Wissenschaften, in Kommission bei W. de Gruyter, 1940.

[Olv00] P. J. Olver, *Applications of Lie groups to differential equations.* Springer Science & Business Media, 2000, vol. 107.

[OU80] L. Odess and H. Ur, "Nullor equivalent networks of nonideal operational amplifiers and voltage-controlled sources," *IEEE Transactions on Circuits and Systems*, vol. 27, no. 3, pp. 231–235, 1980.

[OZ92] B. Owren and M. Zennaro, "Derivation of efficient, continuous, explicit Runge–Kutta methods," *SIAM journal on scientific and statistical computing*, vol. 13, no. 6, pp. 1488–1501, 1992.

[Pay61] H. M. Paynter, *Analysis and design of engineering systems.* MIT press, 1961.

[PB19] F. E. Parker, Julian and A. Bergner, "Modelling of nonlinear state-space systems using a deep neural network," in *Proceedings of the 23rd International Conference on Digital Audio Effects (DAFx-19), Birmingham, UK*, 2019.

[Pd13] J. Parker and S. d'Angelo, "A digital model of the Buchla lowpass-gate," in *Proc. Int. Conf. Digital Audio Effects (DAFx-13), Maynooth, Ireland*, 2013, pp. 278–285.

[PdPV12] R. C. Paiva, S. d'Angelo, J. Pakarinen, and V. Valimaki, "Emulation of operational amplifiers and diodes in audio distortion circuits," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 59, no. 10, pp. 688–692, 2012.

[Pir13] W. Pirkle, "Modeling the Korg35 lowpass and highpass filters," in *Audio Engineering Society Convention 135.* Audio Engineering Society, 2013.

[Pri57] R. C. Prim, "Shortest connection networks and some generalizations," *The Bell System Technical Journal*, vol. 36, no. 6, pp. 1389–1401, 1957.

[PS02] J. M. Pozo and G. Sobczyk, "Geometric algebra in linear algebra and geometry," *Acta Applicandae Mathematica*, vol. 71, no. 3, pp. 207–244, 2002.

[PV11] J. Pekonen and V. Välimäki, "The brief history of virtual analog synthesis," in *Proc. 6th Forum Acusticum. Aalborg, Denmark: European Acoustics Association*, 2011, pp. 461–466.

[PY] J. Pakarinen and D. T. Yeh, "A review of digital techniques for modeling vacuum-tube guitar amplifiers."

[PZLB16] J. Parker, V. Zavalishin, and E. Le Bivic, "Reducing the aliasing of nonlinear waveshaping using continuous-time convolution," in *Proc. Int. Conf. Digital Audio Effects (DAFx-16), Brno, Czech Republic*, 2016, pp. 137–144.

[QM08] G. Quispel and D. I. McLaren, "A new class of energy-preserving numerical integration methods," *Journal of Physics A: Mathematical and Theoretical*, vol. 41, no. 4, p. 045206, 2008.

[RB16] E. K. Ryu and S. Boyd, "Primer on monotone operator methods," *Appl. Comput. Math*, vol. 15, no. 1, pp. 3–43, 2016.

[RCA63] RCA, "Tunnel diode manual," RCA, Tech. Rep., 1963.

[Rei91] S. Reich, "On an existence and uniqueness theory for nonlinear differential-algebraic equations," *Circuits, Systems and Signal Processing*, vol. 10, no. 3, pp. 343–359, 1991.

[Rhe90] W. Rheinboldt, "On the theory and numerics of differential-algebraic equations," Pittsburgh university, Institute for computational mathematics and applications, Tech. Rep., 1990.

[RS10] C. Raffel and J. Smith, "Practical modeling of bucket-brigade device circuits," in *Proc. 13th Int. Conf. Digital Audio Effects (DAFx-10), Graz, Austria*, 2010.

[SB13] J. Stoer and R. Bulirsch, *Introduction to numerical analysis.* Springer Science & Business Media, 2013, vol. 12.

[SBM] J. Snyder, A. Bhatia, and M. Mulshine, "Neuron-modeled audio synthesis," in *Proceedings of the International Conference on New Interfaces for Musical Expression*, pp. 394–397.

[SF11] G. Strang and G. Fix, "A Fourier analysis of the finite element variational method," in *Constructive aspects of functional analysis.* Springer, 2011, pp. 793–840.

[SH11] S. Sarkka and A. Huovilainen, "Accurate discretization of analog audio filters with application to parametric equalizer design," *IEEE transactions on audio, speech, and language processing*, vol. 19, no. 8, pp. 2486–2493, 2011.

[Sha49] C. E. Shannon, "Communication in the presence of noise," *Proceedings of the IRE*, vol. 37, no. 1, pp. 10–21, 1949.

[Sha79] H. S. Shapiro, "Stefan bergman's theory of doubly-orthogonal functions. an operator-theoretic approach," in *Proceedings of the Royal Irish Academy. Section A: Mathematical and Physical Sciences.* JSTOR, 1979, pp. 49–58.

[Sho49] W. Shockley, "The theory of p-n junctions in semiconductors and p-n junction transistors," *Bell System Technical Journal*, vol. 28, no. 3, pp. 435–489, 1949.

[SHV19] F. Silva, T. Hélie, and W. Victor, "Port-Hamiltonian Representation of Dynamical Systems. Application to Self-Sustained Oscillations in the Vocal Apparatus," in *7th Int. Conf. on Nonlinear Vibrations, Localization and Energy Transfer*, vol. 160, Marseille, France, Jun. 2019.

[SK55] R. P. Sallen and E. L. Key, "A practical method of designing RC active filters," *IRE Transactions on Circuit Theory*, vol. 2, no. 1, pp. 74–85, 1955.

[SL19] X. Shen and M. Leok, "Geometric exponential integrators," *Journal of Computational Physics*, vol. 382, pp. 27–42, 2019.

[Sle12] P. Slepian, *Mathematical foundations of network analysis.* Springer Science & Business Media, 2012, vol. 16.

[Sma00] S. Smale, "On the mathematical foundations of electrical circuit theory," in *The Collected Papers of Stephen Smale: Volume 2.* World Scientific, 2000, pp. 951–968.

[Sob08] G. Sobczyk, "Geometric matrix algebra," *Linear Algebra and its Applications*, vol. 429, no. 5-6, pp. 1163–1173, 2008.

[Sob20] ——, "Periodic table of geometric numbers," *arXiv preprint arXiv:2003.07159*, 2020.

[SP61] D. Slepian and H. O. Pollak, "Prolate spheroidal wave functions, fourier analysis and uncertainty-I," *Bell System Technical Journal*, vol. 40, no. 1, pp. 43–63, 1961.

[SS96] T. Stilson and J. Smith, "Analyzing the Moog vcf with considerations for digital implementation," in *Proceedings of the 1996 International Computer Music Conference, Hong Kong, Computer Music Association*, 1996.

[SS98] A. S. Sedra and K. C. Smith, *Microelectronic circuits.* New York: Oxford University Press, 1998.

[SSvdSF05] S. Stramigioli, C. Secchi, A. J. van der Schaft, and C. Fantuzzi, "Sampled data systems passivity and discrete port-hamiltonian systems," *IEEE Transactions on Robotics*, vol. 21, no. 4, pp. 574–587, 2005.

[Ste74] N. Steiner, "Voltage-tunable active filter features, low, high and bandpass modes," in *Electronic design 25, December 6*, 1974.

[Sti05] T. S. Stilson, "Efficiently-variable non-oversampled algorithms in virtual-analog music synthesis: A root-locus perspective." 2005.

[Sti06] T. Stinchcombe, "A study of the Korg MS10 & MS20 filters," http://www.timstinchcombe.co.uk/synth/MS20_study.pdf, 2006, online; accessed: 2019-03-22.

[Sun15] H. S. Sundklakk, "A library for computing with trees and B-series," Master's thesis, NTNU, 2015.

[SVDSMM02] S. Stramigioli, A. Van Der Schaft, B. Maschke, and C. Melchiorri, "Geometric scattering in robotic telemanipulation," *IEEE Transactions on Robotics and Automation*, vol. 18, no. 4, pp. 588–596, 2002.

[SW06] R. Schaback and H. Wendland, "Kernel techniques: From machine learning to meshless methods," *Acta numerica*, vol. 15, p. 543, 2006.

[Tan18] W. Tang, "A note on continuous-stage Runge–Kutta methods," *Applied Mathematics and Computation*, vol. 339, pp. 231–241, 2018.

[Tar12] U. Tarik, "Analyseur de circuit électronique analogique audio, et génération automatique de code pour la simulation temps-réel," IRCAM–Orosys, Tech. Rep., 2012.

[TB99] P. G. Thomsen and C. Bendtsen, "Numerical solution of differential algebraic equations," *Technical University of Denmark DK-2800 Lyngby Denmark May*, vol. 6, 1999.

[Tel52] B. Tellegen, "A general network theorem, with applications," *Philips Res Rep*, vol. 7, pp. 256–269, 1952.

[Tel66] ——, "On nullators and norators," *IEEE Transactions on circuit theory*, vol. 13, no. 4, pp. 466–469, 1966.

[THB14] A. Torin, B. Hamilton, and S. Bilbao, "An energy conserving finite difference scheme for the simulation of collisions in snare drums." in *DAFx*, 2014, pp. 145–152.

[Thi] J. Thickstun, "The Volterra operator," https://homes.cs.washington.edu/~thickstn/docs/volterra.pdf, accessed: 2020-09-01.

[TI10] M. Takamatsu and S. Iwata, "Index characterization of differential–algebraic equations in hybrid analysis for circuit simulation," *International Journal of Circuit Theory and Applications*, vol. 38, no. 4, pp. 419–440, 2010.

[Tis98] C. Tischendorf, "Topological index-calculation of DAEs in circuit simulation," *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik*, vol. 78, no. S3, pp. 1103–1104, 1998.

[Tre09] S. Trenn, *Distributional differential algebraic equations*. Technische Universität Ilmenau, 2009.

[TS12] W. Tang and Y. Sun, "Time finite element methods: a unified framework for numerical discretizations of ODEs," *Applied Mathematics and Computation*, vol. 219, no. 4, pp. 2158–2179, 2012.

[TSC17] W. Tang, Y. Sun, and W. Cai, "Discontinuous galerkin methods for hamiltonian ODEs and PDEs," *Journal of Computational Physics*, vol. 330, pp. 340–364, 2017.

[TSE+15] D. C. Thomas, M. A. Scott, J. A. Evans, K. Tew, and E. J. Evans, "Bézier projection: a unified approach for local projection and quadrature-free refinement and coarsening of NURBS and T-splines with particular application to isogeometric design and analysis," *Computer Methods in Applied Mechanics and Engineering*, vol. 284, pp. 55–105, 2015.

[UAE93a] M. Unser, A. Aldroubi, and M. Eden, "B-spline signal processing. I. theory," *IEEE transactions on signal processing*, vol. 41, no. 2, pp. 821–833, 1993.

[UAE93b] ——, "B-spline signal processing. II. efficiency design and applications," *IEEE transactions on signal processing*, vol. 41, no. 2, pp. 834–848, 1993.

[UB05] M. Unser and T. Blu, "Cardinal exponential splines: part I-theory and filtering algorithms," *IEEE Transactions on Signal Processing*, vol. 53, no. 4, pp. 1425–1438, 2005.

[Uns96] M. Unser, "Approximation power of biorthogonal wavelet expansions," *IEEE Transactions on Signal Processing*, vol. 44, no. 3, pp. 519–527, 1996.

[Uns00] ——, "Sampling-50 years after Shannon," *Proceedings of the IEEE*, vol. 88, no. 4, pp. 569–587, 2000.

[Uns05] ——, "Cardinal exponential splines: part II-think analog, act digital," *IEEE Transactions on Signal Processing*, vol. 53, no. 4, pp. 1439–1449, 2005.

[UVI21] UVI, "Falcon," https://www.uvi.net/falcon.html, 2021, online; accessed: 2021-04-15.

[VBS+11] V. Välimäki, S. Bilbao, J. Smith, J. Abel, J. Pakarinen, and D. Berners, "Virtual analog effects," in *DAFX: Digital Audio Effects*. Wiley Online Library, 2011, pp. 473–522.

[VBS17] M. Verasani, A. Bernardini, and A. Sarti, "Modeling Sallen-Key audio filters in the wave digital domain," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 431–435.

[VD95] J. Van Dijk, "On the role of bond graph causality in modelling mechatronic systems." 1995.

[VdS06] A. Van der Schaft, "Port-Hamiltonian systems: an introductory survey," 2006.

[VdS17] ——, *L2-gain and passivity techniques in nonlinear control*. Springer, 2017, vol. 3.

[VDSJ14] A. Van Der Schaft and D. Jeltsema, "Port-Hamiltonian systems theory: An introductory overview," *Foundations and Trends in Systems and Control*, vol. 1, no. 2-3, pp. 173–378, 2014.

[VdSM11] A. Van der Schaft and B. Maschke, "Discrete conservation laws and port-Hamiltonian systems on graphs and complexes," *submitted for publication*, 2011.

[VdSM13] ——, "Port-Hamiltonian systems on graphs," *SIAM Journal on Control and Optimization*, vol. 51, no. 2, pp. 906–937, 2013.

[VdSM18] ——, "Generalized port-Hamiltonian DAE systems," *Systems & Control Letters*, vol. 121, pp. 31–37, 2018.

[VFSZ10] V. Valimaki, F. Fontana, J. O. Smith, and U. Zölzer, "Introduction to the special issue on virtual analog audio effects and musical instruments," *IEEE transactions on audio, speech, and language processing*, vol. 18, no. 4, pp. 713–714, 2010.

[VGO+20] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright *et al.*, "Scipy 1.0: fundamental algorithms for scientific computing in python," *Nature methods*, vol. 17, no. 3, pp. 261–272, 2020.

[VH06] V. Välimäki and A. Huovilainen, "Oscillator and filter algorithms for virtual analog synthesis," *Computer Music Journal*, vol. 30, no. 2, pp. 19–31, 2006.

[VL88] P. P. Vaidyanathan and V. C. Liu, "Classical sampling theorems in the context of multirate and polyphase digital filter bank structures," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 9, pp. 1480–1495, 1988.

[Vla94] A. Vladimirescu, *The SPICE book*. Wiley New York, 1994.

[VLM12] N. M. T. VU, L. Lefevre, and B. Maschke, "Port-Hamiltonian formulation for systems of conservation laws: application to plasma dynamics in tokamak reactors," *IFAC Proceedings Volumes*, vol. 45, no. 19, pp. 108–113, 2012.

[VMB02] M. Vetterli, P. Marziliano, and T. Blu, "Sampling signals with finite rate of innovation," *IEEE transactions on Signal Processing*, vol. 50, no. 6, pp. 1417–1428, 2002.

[WBK02] T. Wong, P. Bigras, and K. Khayati, "Causality assignment using multi-objective evolutionary algorithms," in *IEEE International Conference on Systems, Man and Cybernetics*, vol. 4. IEEE, 2002, pp. 6–pp.

[WBSS18] K. J. Werner, A. Bernardini, J. O. Smith, and A. Sarti, "Modeling circuits with arbitrary topologies and active linear multiports using wave digital filters," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 65, no. 12, pp. 4233–4246, 2018.

[WC77] J. L. Wyatt and L. O. Chua, "A theory of nonenergic n-ports," *International Journal of Circuit Theory and Applications*, vol. 5, no. 2, pp. 181–208, 1977.

[WdNF+19] M. Wijnand, B. d'Andréa Novel, B. Fabre, T. Hélie, L. Rosier, and D. Roze, "Active control of the axisymmetric vibration modes of a tom-tom drum," in *2019 IEEE 58th Conference on Decision and Control (CDC)*. IEEE, 2019, pp. 6887–6892.

[WdNHR18] M. Wijnand, B. d'Andréa Novel, T. Hélie, and D. Roze, "Contrôle des vibrations d'un oscillateur passif: stabilisation en temps fini et par remodelage d'énergie," in *14ème Congrès Français d'Acoustique*, 2018.

[WDR+16] K. J. Werner, W. R. Dunkel, M. Rest, M. J. Olsen, and J. O. Smith, "Wave digital filter modeling of circuits with operational amplifiers," in *2016 24th European Signal Processing Conference (EUSIPCO)*. IEEE, 2016, pp. 1033–1037.

[WDV19] A. Wright, E.-P. Damskägg, and V. Välimäki, "Real-time black-box modelling with recurrent neural networks," in *Proceedings of the 23rd International Conference on Digital Audio Effects (DAFx-19), Birmingham, UK*, 2019.

[Wei83] A. Weinstein, "Symplectic geometry," *Proceedings of Symposia in Pure Mathematics*, vol. 39, 1983.

[Wer16] K. J. Werner, "Virtual analog modeling of audio circuitry using wave digital filters," Ph.D. dissertation, Stanford University, 2016.

[WHN78] G. Wanner, E. Hairer, and S. P. Nørsett, "Order stars and stability theorems," *BIT Numerical Mathematics*, vol. 18, no. 4, pp. 475–489, 1978.

[WHS19] V. Wetzel, T. Hélie, and F. Silva, "Power balanced time-varying lumped parameter model of a vocal tract: modelling and simulation," in *26th International Conference on Sound and Vibration*, 2019.

[WHS20] ——, "Power-balanced modelling of the vocal tract: a recast of the classical lumped-parameter model," in *Forum Acusticum 2020 (e-Forum Acusticum)*, 2020.

[WNSA15] K. J. Werner, V. Nangia, J. O. Smith, and J. S. Abel, "A general and explicit formulation for wave digital filters with multiple/multiport nonlinearities and complicated topologies," in *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2015, pp. 1–5.

[WW18] X. Wu and B. Wang, "Exponential average-vector-field integrator for conservative or dissipative systems," in *Recent Developments in Structure-Preserving Algorithms for Oscillatory Differential Equations*. Springer, 2018, pp. 29–53.

[WX12] H. Wang and S. Xiang, "On the convergence rates of Legendre approximation," *Mathematics of Computation*, vol. 81, no. 278, pp. 861–877, 2012.

[YAS10] D. T. Yeh, J. S. Abel, and J. O. Smith, "Automated physical modeling of nonlinear audio circuits for real-time audio effects-part i: Theoretical development," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 4, pp. 728–737, 2010.

[ZM88] G. Zhong and J. E. Marsden, "Lie-Poisson Hamilton-Jacobi theory and Lie-Poisson integrators," *Physics Letters A*, vol. 133, no. 3, pp. 134–139, 1988.

[ZRM09] R. K. Zia, E. F. Redish, and S. R. McKay, "Making sense of the Legendre transform," *American Journal of Physics*, vol. 77, no. 7, pp. 614–622, 2009.