Horizon 2020

Reduced Order Modelling, Simulation and Optimization of Coupled systems

Reports about 8 selected benchmark cases of model hierarchies

Deliverable number: D5.1

Version 3.0

Funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 765374

Project Acronym: ROMSOC
Project Full Title: Reduced Order Modelling, Simulation and Optimization of Coupled systems
Call: H2020-MSCA-ITN-2017
Topic: Innovative Training Network
Type of Action: European Industrial Doctorates
Grant Number: 765374

Editors:	Andrés Prieto, Peregrina Quintela, ITMATI
Deliverable nature:	Report (R)
Dissemination level:	Public (PU)
Contractual Delivery Date:	01/08/2018
Actual Delivery Date	30/09/2018
Number of pages:	114
Keywords:	Benchmarks, Model hierarchies
Authors:	Naomi Auer, WIAS Berlin Patricia Barral, ITMATI-USC Jean-David Benamou, INRIA Andreas Baermann, FAU Andres Binder, MathConsult Daniel Fernández, Microflown Technologies Michele Girfoglio, SISSA Lena Hauberg-Lotte, U-HB Michael Hintermüller, WIAS Berlin Wilbert Ijzerman, Signify Onkar Jadhav, MathConsult Karl Knall, MathTec Peter Maass, U-HB Gianfranco Marconi, Danieli Marco Martinolli, MOX, PoliMi Volker Mehrmann, TU Berlin Pier Paolo Monticone, CorWave SA Umberto Emil Morelli, ITMATI Ashwin Nayak, ITMATI Luc Polverelli, CorWave Inc. Andrés Prieto, ITMATI-UDC Peregrina Quintela, ITMATI-USC Ronny Ramlau, Industrial Mathematics Institute JKU Conte Riccardo, Danieli Gianluigi Rozza, SISSA Giorgi Rukhaia, INRIA Nirav Shah, SISSA Bernadett Stadler, Industrial Mathematics Institute JKU Jonasz Staszek, FAU Christian Vergara, MOX, PoliMi
Peer review:	Andrés Prieto, ITMATI-UDC Peregrina Quintela, ITMATI-USC

Abstract

Based on the multitude of industrial applications, benchmarks for model hierarchies will be created that will form a basis for the interdisciplinary research and for the training programme. These will be equipped with publically available data and will be used for training in modelling, model testing, reduced order modelling, error estimation, efficiency optimization in algorithmic approaches, and testing of the generated MSO/MOR software. The present document includes the description about the selection of (at least) eight benchmark cases of model hierarchies. The present document has been structured in three main parts to distinguish those contributions which are focused on coupling methods, model order reduction methods, and optimization methods.

Disclaimer & acknowledgment
This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 765374.

This document reflects the views of the author(s) and does not necessarily reflect the views or policy of the European Commission. The REA cannot be held responsible for any use that may be made of the information this document contains.

Contents

List of Acronyms

ITMATI	Technological Institute of Industrial Mathematics
USC	University of Santiago de Compostela
UDC	University of A Coruña
JKU	Johannes Kepler University Linz
RICAM	Johann Radon Institute for Computational and Applied Mathematics
AAO	Austrian Adaptive Optics Team
AO	Adaptive Optics
SCAO	Single Conjugated AO
LTAO	Laser Tomography AO
MCAO	Multi Conjugated AO
MOAO	Multi Object AO
CuRe	Cummulative Reconstructor
CuReD	Cummulative Reconstructor with Domain Decomposition
MVM	Matrix Vector Multiplication
PSF	Point Spread Function
OTF	Optical Transfer Function
WFS	Wavefront Sensor
GS	Guided Starts
NGS	Natural Guided Starts
LGS	Laser Guided Starts
CCD	Carbonate Compensation Depth
ELT	Extremly Large Telescope
FEWHA	Finite Element Wavelet Hybrid Algorithm
FWHM	Full Width at Half Maximum
PML	Perfectly Matched Layer
FEM	Finite Element Method
SEM	Scanning Electron Microscope
MPI	magnetic particle imaging
SPIO	superparamagnetic iron oxide nanoparticles
DL	deep learning
NN	neural networks
CT	computer tomography
MRI	magnetic resonance imaging
PET	position emission tomography
FFP	field free point
FFL	field free line
FOV	field-of-view
ADMM	alternating direction method of multipliers
LISTA	Learning Fast Approximations of Sparse Coding
CNN	convolutional neural network

LSTM	long-short-term-memory network
MDF	MPI data format
RNN	recurrent neural network
BC	Boundary Condition
CC	Continuous Casting
RANS	Reynolds Averaged Navier-Stokes equations
ROM	Reduced Order Modeling
SEN	Submerged Entry Nozzle
SISSA	Scuola Internazionale Superiore di Studi Avanzati
FEM	Finite Element Method
FSI	Fluid-Structure Interaction
LIFEV	LIbrary of Finite Elements V
LVADs	Left Ventricular Assist Devices
NS	Navier-Stokes Equations
PDEs	Partial Differential Equations
X-FEM	Extended Finite Element Method
FEM	Finite Element Method
RANS	Reynolds-averaged Navier-Stokes

Part I.

Coupling methods

1. Benchmark cases in development of coupled models for a sound intensity probe

Ashwin Nayak[1], Andrés Prieto[1], Daniel Fernández[2]

[1] *Instituto Tecnolóxico de Matemática Industrial, Universidade da Coruña*

[2] *Microflown Technologies*

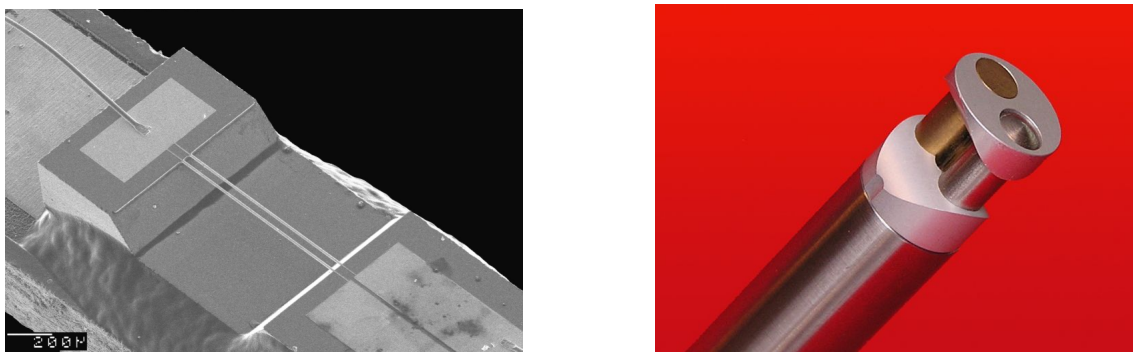
Abstract. We present the progress towards the development of a mathematical model and simulation for better understanding of the physical processes behind a sound intensity PU probe. The particle velocity sensor and the sound pressure microphone are mounted on two closely spaced cylinders which are protected with a porous windscreen. We evaluate the possibilities of using different porous enclosures around the probe to improve its sensitivity and response to air flow through the means of a mathematical model. A thorough physical model can help optimize the design criteria and enhance the signal quality. This preliminary report discusses the impact of the windscreen on the sensor's outputs by simulating their physical behavior in multiple stages. Benchamark stages are identified along the development to highlight progress and initial approach is explained. The response is measured by simulating a simplified model and the gain is computed and compared.

Keywords: Scattering, aeroacoustics, porous materials.

1.1. Introduction

The Microflown PU probes are acoustical sensors distinguished by their ability to measure fluctuations in particle velocity and pressure fields simultaneously and are of considerable importance in sound measurement. They have been used to measure sound-intensity, acoustic impedance, pressure and have been adopted for a variety of applications like sound localization, acoustic quantification and noise control. In the past 24 years since its invention at University of Twente unto present, it is said to be the only sensor to measure particle velocity directly off the acoustic field.

The sensor works on the principal of differential heating. The microscopic transducer (Fig1.1a) consists of a two tiny, resistive strips of platinum which measure the acoustic particle velocity. In the presence of particle velocity, the temperature distribution around the resistors is asymmetrically altered, and a difference in temperature occurs between the two wires, as the upstream wire is cooled more than the downstream wire by the acoustic airflow. The resulting resistance difference translates into variations in a signal which captures the particle velocity of a single point of a sound field in an arbitrary direction.



(a) **(b)**

Figure 1.1: (a) SEM image of the Microflown's particle velocity transducer, and (b) the PU Probe housing the particle velocity and pressure sensors

The Microflown sensor being used in different environmental conditions, is observed to be sensitive to flow conditions. Since the transducer is mainly a thermal device, the probe wires have to be heated for proper operation. Previous models of the microflown transducers [1], [2] indicate that at large flow velocities, the metal wires, having a finite heat capacity, cools down and the signal can distort considerably or die down unexpectedly. Due to this physical limitation, the sensor is usually used with a windscreen when measurements are taken in an environment which has a significant air flow. The company offers multiple windscreens in various sizes and materials, from cylindrical porous covers to metallic meshes.

Even while the flow is sufficiently blocked, the low frequency variations in the flow may still be large enough to cause problems when amplified. Studies indicate that windscreens on the sensor may induce significant package gain and phase error in the particle velocity measurements [3],[4]. If the wind is sufficiently blocked, the low frequency variations in the flow may still be large enough to cause problems when amplified. Also at higher flow velocities, any introduction of physical obstacle may cause turbulence which effectively translates into acoustic noise. Moreover, since the self noise induced by the flow is used as reference from which the noise is measured, a dynamic calibration maybe required in such cases. Hence, different wind-screening strategies should be considered with the aim of mitigating such effects on the probe measurements.

The objective of this project is to study the effects of introducing a porous media layer over a sound intensity PU probe to prevent noise inhibitions whist a broadband frequency response. A computational model which simulates the behaviour of the sensor can highlight the physical characteristics and help optimize the design for improving the measurement. This report shows the approach towards development of such a model and preliminary results in this regard by trying to measure the impact of the windscreen of the sensor on its measurement.

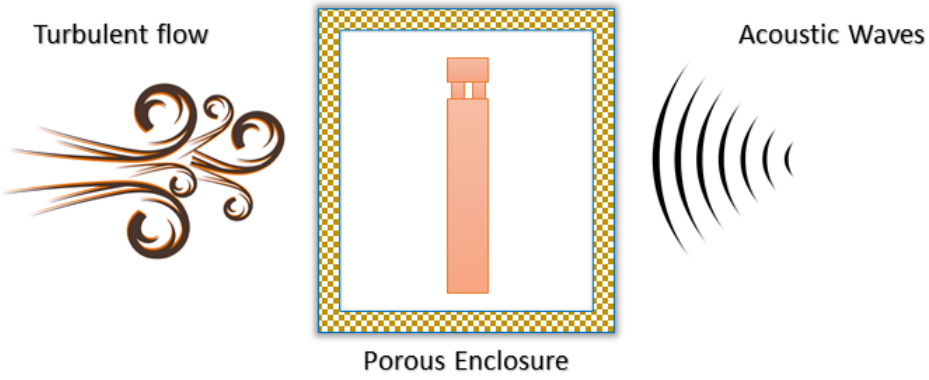


Figure 1.2: Schematic of the target coupled problem

1.2. Model development

The physical behaviour of the acoustical sensor enclosed by a porous material in flow requires an astute understanding of sound propagation over the porous structure in different flow conditions. The material properties and geometry of the porous windscreen can either dampen or accentuate either the magnitude or the phase of the measured signal. Additionally, turbulence in the acoustic fluid can give rise to noise at various frequency ranges. In order to predict these responses accurately it is necessary to model the coupling between the various physical models of flow and structure.

We consider the development of a computational model to simulate such coupled behaviour in 6 stages as shown in Fig 1.3. Each stage is represented by a schematic diagram indicating the coupling required. For simplicity, we begin the model development with a spherical geometry in free-field conditions with fluid at rest and an incident acoustic wave. The model plainly describes the effect of placing a sphere in an acoustic field. This is easily quantified by measuring the signal characteristics close to the sphere with the known analytic expression thereby giving us the gain or loss in signal. The model is further developed to incorporate the coupling with the

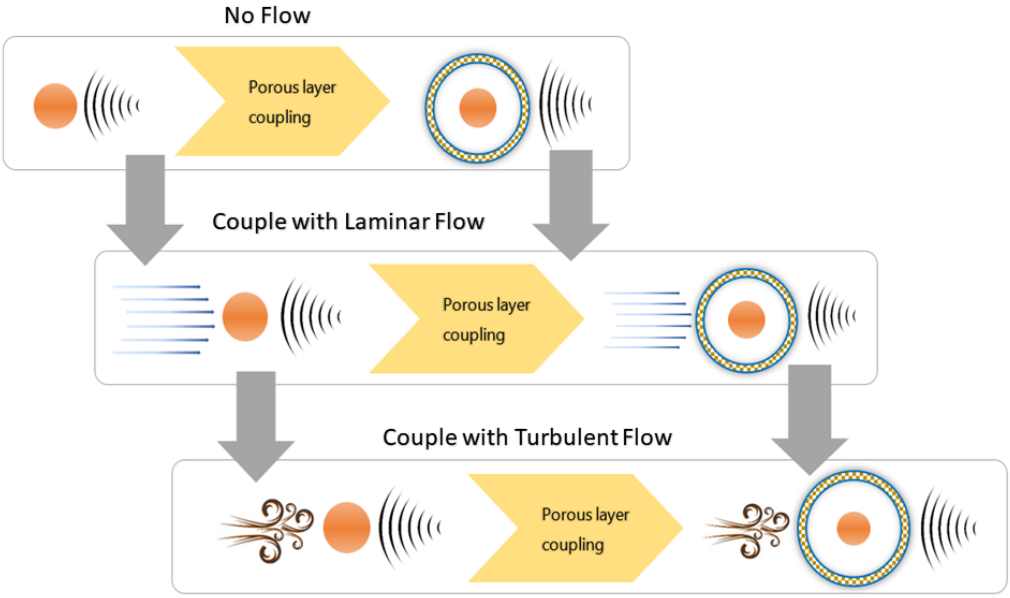


Figure 1.3: Schematic of the model development hierarchy for the different couplings. The orange blob represents a simplified geometry, acoustic waves are represented as concentric wavy arcs, porous windscreen with the checkered layer, laminar flow as straight blue arrows and turbulent flow with the swirls

porous material to help understand the acoustic behaviour in presence of such an windscreen.

To study the effects of flow on the acoustic fluid, we then start with coupling the initial model with the fluid assumed to be laminar and non-viscous. This model is also later evolved to incorporate a porous housing. Turbulent behaviour is incorporated at last after a thorough understanding of the flow characteristics in the laminar regime. Finally, including a porous windscreen within it, we arrive at the target model to study the behaviour of the all these coupling events.

1.3. Benchmark Cases

The previous section indicated the various stages of development of the mathematical model towards modeling the porous windscreen on the sound intensity PU probe. We discuss here, the application of each of the model stages on the probe which can help validate the model at the various stages. Each such application can be considered a benchmark case to assess the project development.

Each stage of the mathematical model developed can be tested and validated by comparing the numerical results obtained to experimental data obtained from the company. Table 1.1 lists each of the six cases along with the schematic representation. Once the mathematical model is developed for a simplified geometry in no-flow conditions, the same model may be applied to the probe housing to study the theoretical effect of its presence in the measurement. The numerical results obtained can be verified experimentally to validate the results. The model is then developed to couple with the material properties of a porous windscreen, to arrive at the predicted difference in gain of the particle velocity signal. This can again, be corroborated by reproducing the experiment in a laboratory setup.

Adding flow conditions to the initial model will require re-examining the basic conservation principles and arrive at a different mathematical model. As before, it can be used to obtain an estimate of difference gain and phase response of the signal and be validated with experiments. At each stage, details of the model, the results obtained from the model and the data obtained from experiments are enumerated and compared. We begin with the easier laminar flow models at lower Reynolds number. Once validated, we proceed to couple it with the material response of the porous windscreen of the sensor. The results are then validated with reproducing the

conditions in laboratory. Subsequently, turbulent models are developed and again be verified by experiments. The methodology is developed over the course of the project and presently, we begin by modelling the simplest case of no-flow conditions with no porous windscreen.

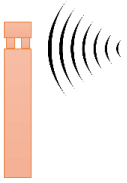
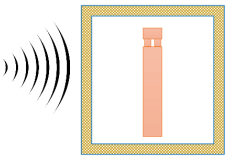
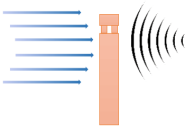
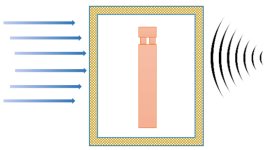
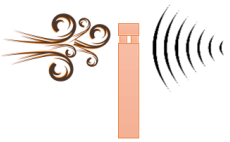
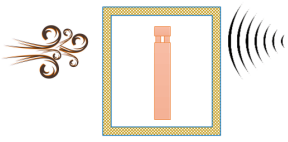
Case	Description	Schematic Representation
1	Mathematical model of acoustic fluidwith fluid at rest	
2	Coupled model with porous windscreen and fluid at rest	
3	Mathematical model of acoustic field in laminar flow	
4	Coupled model with porous windscreen in laminar flow	
5	Mathematical model of acoustic field in turbulent flow	
6	Coupled model with porous windscreen in turbulent flow	

Table 1.1: Benchmark Cases

1.4. Case 1: Fluid at rest

1.4.1. Computational Model

To study the effects of the windscreen on the probe measurement, we study the simple scattering problem on the PU probe in free-field conditions. The acoustic fluid around the field is assumed homogeneous, non-viscous, compressible, and at isentropic regime. A plane wave is assumed to impinge on the probe from the x-direction. The pressure field of the fluid is $P(\mathbf{x}, t)$, the density field is $\rho(\mathbf{x}, t)$ and the velocity vector field is $\mathbf{U}(\mathbf{x}, t)$. With these assumptions, to obtain the equations governing the acoustic wave propagation, we consider the conservation equations of mass and momentum:

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{U}) = 0, \tag{1.1}$$

$$\rho \frac{\partial \mathbf{U}}{\partial t} + \nabla P = 0. \tag{1.2}$$

We also need constitutive assumptions which characterize the relation between state variables in the medium. With the assumption of acoustic fluid being compressible, we obtain the linear relation

$$P = c^2 \rho, \tag{1.3}$$

where the material constant c is the speed of sound in the fluid. Using (1.1), (1.2) and (1.3), we obtain the equation governing the acoustic wave propagation as,

$$\frac{1}{c^2} \frac{\partial^2 P}{\partial t^2} = \Delta P. \tag{1.4}$$

1.4.1.1. Helmholtz problem

The time-harmonic assumption suggests that a sound signal can partitioned over the frequency bands and the pressure field for a given frequency ω may be expressed as,

$$P(\mathbf{x}, t) = p(\mathbf{x}) \exp(-i\omega t). \tag{1.5}$$

Using this, the wave equation (1.4) becomes,

$$-\Delta p - k^2 p = 0, \tag{1.6}$$

known as the *Helmholtz equation*. The constant $k = \omega/c$ is called the acoustic wave number. While this equation describes the propagation of pressure perturbations within the acoustic field, it needs to closed by prescribing the right boundary conditions.

1.4.1.1.1. Boundary conditions In order to setup a scattering problem, it is necessary to introduce the plane wave as a boundary condition. A plane wave is a solution of the Helmholtz equation of a particular form:

$$p_{\text{inc}}(\mathbf{x}) = A \exp(-i\mathbf{k} \cdot \mathbf{x}), \tag{1.7}$$

where, $\mathbf{k}$ represents the wave number vector. This vector has the magnitude of the wave number k and the direction along the propagation of the plane wave. It is necessary to impose this on the structure boundary as a coupling condition with the fluid. Since, the probe body is considered rigid, Neumann boundary conditions are assumed on its boundary, i.e,

$$\frac{\partial p}{\partial \mathbf{n}} = -\frac{\partial p_{\text{inc}}}{\partial \mathbf{n}}. \tag{1.8}$$

Here, p_{inc} is the pressure of the incident plane wave which hits the structure and $\mathbf{n}$ represents unit vectors on the surface.

Since the study is at free-field conditions, the scattering problem also requires that the acoustic field vanish at points further away from the domain of interest. This can be written as the Sommerfeld boundary condition in radial coordinates i.e,

$$\lim_{r\to\infty} r\left(ikp - \frac{\partial p}{\partial r}\right) = 0, \tag{1.9}$$

with $r = \|\mathbf{x}\|$ being the radial distance from the origin. While this non-reflective boundary condition is theoretically suitable, it is particularly hard to implement in analytical calculations [5, 6]. In a practical scenario, the domain needs to be truncated at some finite distance and this would introduce spurious reflections off the boundary.

This necessitates that we introduce an artificial boundary around the truncated domain and prescribe 'absorbing' boundary conditions that incorporate (exactly or approximately) the far-field behaviour into the model. The Perfectly Matched Layer is one such technique in this regard which mimics free-field conditions on the fluid domain.

1.4.1.2. The PML method

The Perfectly Matched Layer (PML) method is a recently developed technique as a kind of absorbing boundary condition on an artificial boundary around the domain of interest which equivalently replaces the Sommerfeld condition. The idea, inspired from electromagnetic simulations, is specifically to introduce an exterior layer at the boundary in such a way that all plane waves are totally absorbed. This allows for an efficient approach towards utilizing the techniques for bounded-domain models of wave propagation on unbounded domains. For an arbitrary incident plane wave on the exterior layer, no reflection occurs and the transmitted waves are designed to vanish at infinity [6].

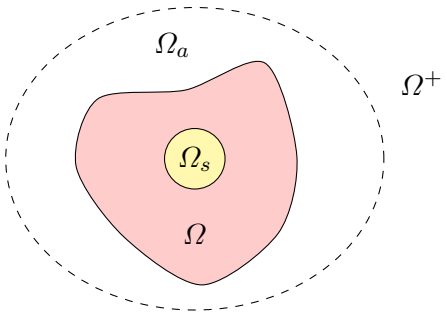


Figure 1.4: Schematic of Perfectly Matched Layers

Consider a structural domain Ω_s which is to be surrounded by seemingly, free-field conditions. Figure 1.4 shows the domain of interest, Ω^+, truncated at some finite region and surrounded by an artificial boundary layer, Ω_a. To obtain the partial differential equations associated with the fluid in Ω_a, a formal complex-valued change of variables is introduced in the Helmholtz equation. The partial derivatives in space along each direction, i.e $\partial/\partial x_j$ are scaled by complex-valued factors $1/\gamma_j$, with,

$$\gamma_j = 1 - i\frac{\sigma_j}{\omega} \quad \text{for} \quad j = 1, 2, 3. \tag{1.10}$$

The functions σ_j are called the absorption functions of the PML. In our case we assume them to be piecewise

constant given as,

$$\sigma_j = \begin{cases} \sigma_0 & \text{if} \quad x_j \in \Omega_a, \\ 0 & \text{if} \quad x_j \in \Omega. \end{cases} \tag{1.11}$$

The values σ_0 for the PML layer is hence, dependent on the frequency of the acoustic field. Hence, an optimal value needs to be arrived at for each frequency. With these change of variables, the PML governing equation can be written for the domain $\Omega \cup \Omega_a$ as follows:

$$-\nabla \cdot (C \nabla p) - k^2 M p = 0, \tag{1.12}$$

where

$$C = \text{diag}\Big(\frac{\gamma_2\gamma_3}{\gamma_1}, \frac{\gamma_3\gamma_1}{\gamma_2}, \frac{\gamma_1\gamma_2}{\gamma_3}\Big) \qquad \text{and} \qquad M = \gamma_1\gamma_2\gamma_3.$$

Note that within Ω, the equation (1.12) reduces to the Helmholtz equation. To close the problem, it is also necessary to mention the boundary condition of this PDE (1.13). The conditions on the exterior of the artificial boundary $\partial\Omega_a$ have not jump discontinuities (the jump of a quantity is denoted between brackets):

$$\begin{aligned} \frac{\partial p}{\partial n} &= g \quad \text{on} \quad \partial\Omega_s, \\ p &= 0 \quad \text{on} \quad \partial\Omega, \\ [p] &= 0 \quad \text{on} \quad \partial\Omega_a, \\ \Big[\Big(\frac{1}{\gamma_x}\frac{\partial p}{\partial x}, \frac{1}{\gamma_y}\frac{\partial p}{\partial y}, \frac{1}{\gamma_z}\frac{\partial p}{\partial z}\Big) \cdot \mathbf{n}\Big] &= 0 \quad \text{on} \quad \partial\Omega_a. \end{aligned} \tag{1.13}$$

1.4.1.3. Finite Element Method

The Finite Element Method (FEM) enables solving a problem in infinite-functional space approximately by means of a finite-dimensional discrete space. The model described below describes this method applied for acoustic problems in the unbounded domains with a perfectly matched layer (PML). First, the PDE which is to be solved in the domain is framed as a variational problem (also known as the weak formulation). For this, we multiply equation (1.12) by the complex test function φ, write all the complex terms explicitly and utilize the Green's identity to obtain,

$$\int_{\Omega\cup\Omega_a} C\nabla p \cdot \nabla\varphi dV - \int_{\Omega\cup\Omega_a} k^2 M p\varphi dV = \int_{\partial\Omega_s} g\varphi d\gamma. \tag{1.14}$$

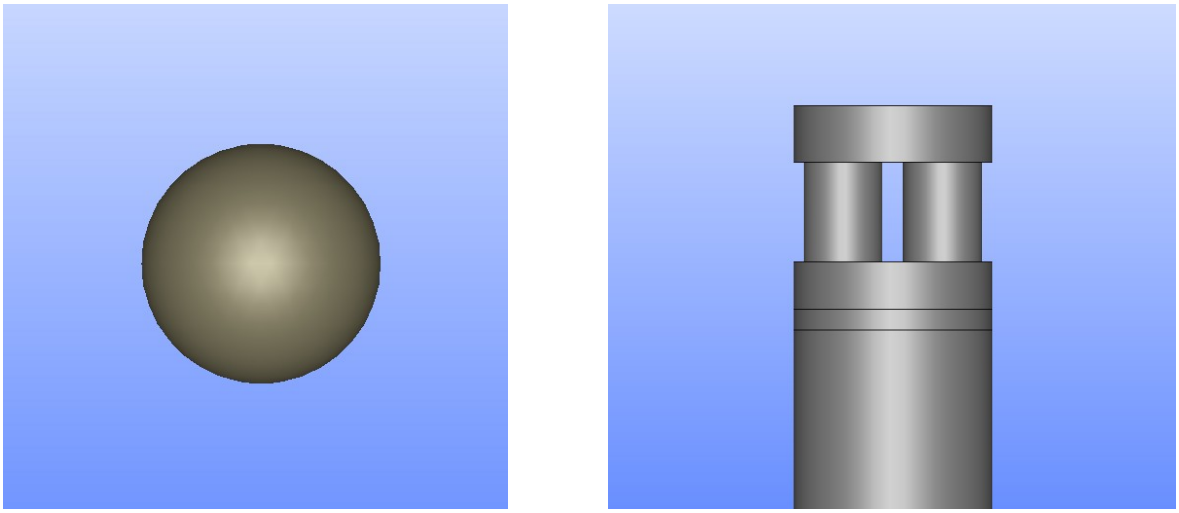
Note that the boundary conditions appear as terms on left hand side of the above equation. This complex equation is more conveniently expressed as the following variation problem: Find $p \in V(\Omega \cup \Omega_a)$ such that $p = 0$ on $\partial\Omega_a$ and it holds

$$\mathcal{A}(p, \varphi) = \mathcal{L}(\varphi), \tag{1.15}$$

for all $\varphi \in V(\Omega \cup \Omega_a)$ with $\varphi = 0$ on $\partial\Omega_a$. In the equation written above $\mathcal{A}$ and $\mathcal{L}$ denote the left and right hand sides of equation (1.14), respectively. This is the general form of the variational problem with $\mathcal{A}$ and $\mathcal{L}$ being the sesquilinear and linear functionals. Now, we approximate the infinite dimensional functional space V, with a n-dimensional space V_h and re-write (1.15) in terms of the basis functions (ψ's) of this space. We obtain two equations from equating the real and imaginary parts as,

$$\sum_{j=1}^{n}\sum_{m=1}^{n} \mathcal{A}(\psi_j, \psi_l)\mu_j = \mathcal{L}(\psi_l) \qquad \text{for } l = 1, 2..., n. \tag{1.16}$$

Equation (1.16) is solved for the linear coefficients μ_i's to obtain the approximate solution.



(a) (b)

Figure 1.5: CAD models of (a) a sphere (b) simplified Microflown PU Probe. The models were created using SALOME.

1.4.2. Code implementation

1.4.2.1. CAD Model and Meshing

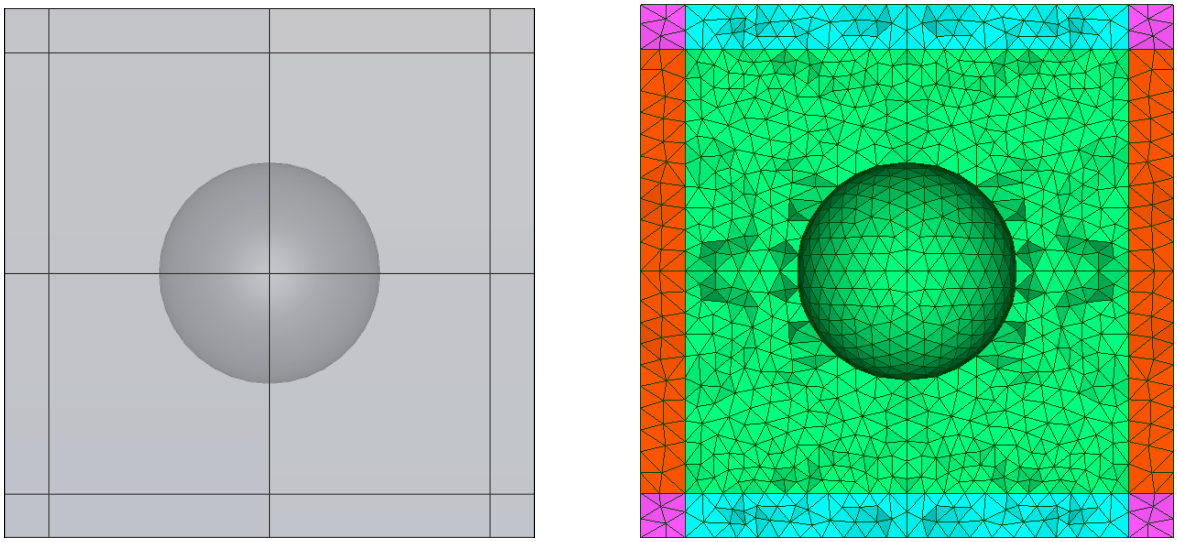
We start by modelling the geometry of the test subject so as to form a mesh representing the physical domain. This is achieved by using the modelling software, SALOME [7]. SALOME is an easy-to-use open-source CAD modelling and meshing software and offers interfaces to various simulation solvers. It also supports PYTHON scripts for modelling and meshing the subject geometry enabling dynamic changes if required.

A sphere of radius placed in air is utilized as our initial test subject to validate the code. The domain is modeled as a rectangle of sufficient width around it. The geometric dimensions of the PU probe are obtained from the partner company, Microflown Technologies and is utilized to make a CAD model for testing. The Microflown is initially considered in its simplified design which is considered to be sufficient for the preliminary study to compute gain of the probe. Both the models are shown modeled in SALOME in Figure 1.5. Since we are interested in solving the Helmholtz problem for the unbounded domain, we utilize the computational technique of replacing it with a truncated finite region around the probe and surround it with the PML as shown in Figure 1.7a. The geometry is then meshed using the NETGEN algorithms available in the SALOME's mesh module. Figure 1.6b and 1.7b shows the cross-section of the tetrahedral mesh color-codeded to distinguish the PML layers. Each colored domain represents markers to indicate varying C and M in equation (1.12).

1.4.2.2. Finite Element Solver

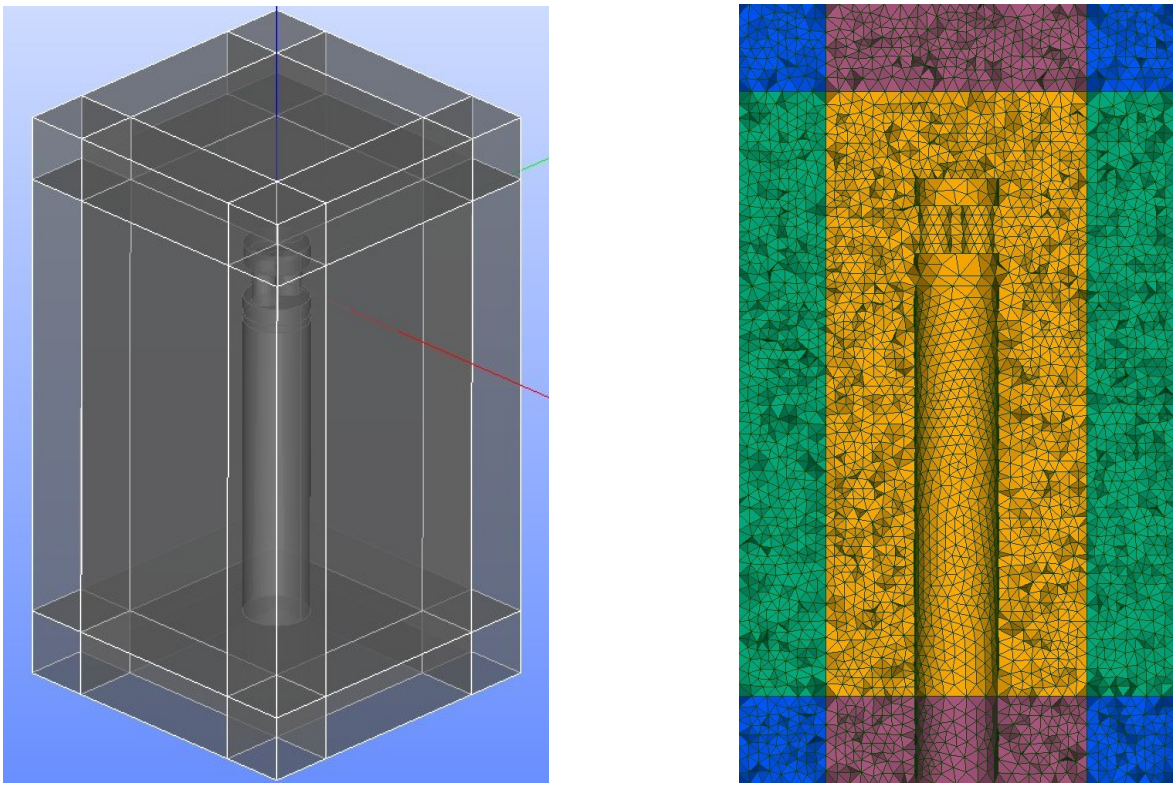
Now that a mesh is constructed, we need a finite element solver to solve the equations (1.12) with boundary conditions (1.13) on this domain. The solver of choice is FEniCS [8], a popular open-source computing platform for solving partial differential equations (PDEs) using the finite element method. FeniCS offers containers and data-structures such as meshes, function spaces, finite-element assembly and mesh refinement and modules to generate finite element basis functions and expressing variational problems, to implement the solver in mathematical concepts which makes it easy to use. Moreover, it offers interfacing with various linear algebra solvers and data structures such as PETSc and supports programming in C++ and PYTHON[9].

PYTHON scripts are are in development to solve the Helmholtz equation with the PML, developing over the codebase by Coral Real [10]. It utilizes the FEniCS libraries to assign boundary conditions, define Lagrange P1 basis functions at nodes, assemble the equations like (1.16) and finally solves them using the PETSc solver.



(a) **(b)**

Figure 1.6: (a) CAD image of the PML around the sphere, and
(b) Cross-section of the tetrahedral mesh volumes



(a) **(b)**

Figure 1.7: (a) CAD image of the PML around the probe geometry, and
(b) Cross-section of the tetrahedral mesh volumes

1.4.3. Results

1.4.3.1. Computing the PML absorption coefficients

The code is initially validated by considering a sphere in free-field conditions. By choosing the boundary conditions as a solution of a monopole source, an analytical solution is available to be compared. The value of the absorption function of the PML is optimized over a set of frequencies and spline interpolation is used to compute the values in between. Figure 1.8 shows the values of σ_0 obtained over a range of frequencies.

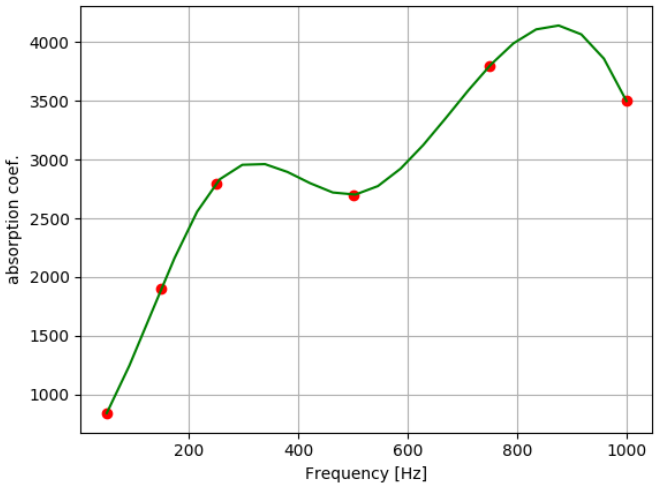


Figure 1.8: Values of PML absorption coefficient σ_0 for various frequencies

1.4.3.2. Validation

The pressure and velocity signals at the surface of sphere is then compared with the analytical solution to compute the accuracy of the implementation. Figure 1.9 shows the agreement of the analytical solution of the velocity field with the one computed by solving using the Finite Element method. The difference is mainly caused by numerical discretization and also due to the approximation with PML method. The results show good agreement with the analytical solution (within $\pm$ 2dB) as shown in Fig 1.10. The solver is hence validated and can be utilized to study the gain of the pressure and velocity signals due to the sensor windscreen.

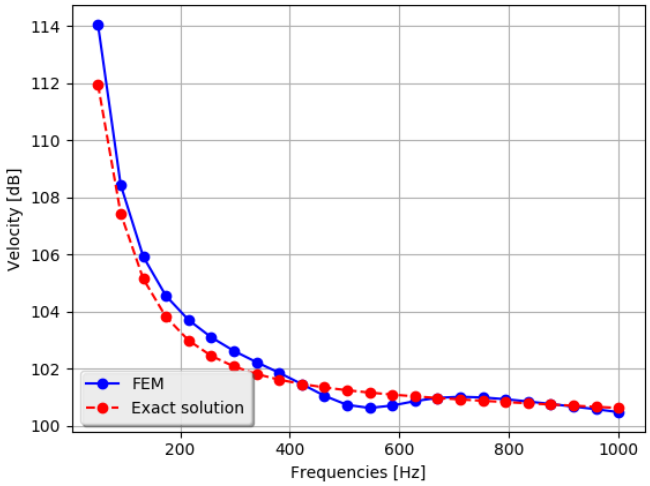


Figure 1.9: Validation of the code by comparing the velocity field at with the exact solution

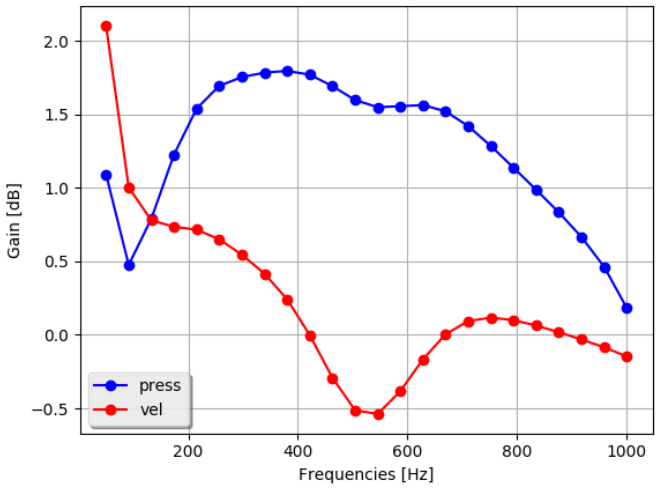


Figure 1.10: Gain in the pressure and particle velocity signals

1.5. Conclusion

In improving the sensitivity of the Microflown sensor, this project works towards exploring the use of porous material windscreen as windscreen rather than traditional rigid packaging. In this regard, it is necessary to measure the effect those windscreens can cause to an acoustic field. This project highlights the staggered development of a mathematical model to enable simulation of such effects. It highlights the various physics which needs to be coupled to obtain an accurate model. We suggest six stages to develop such a model and list six benchmark cases to validate the models at those stages with experiments. An initial study is performed on developing a mathematical model to study the loss of sensitivity due to the probe windscreen at no flow conditions. The solver to implement this model is in preliminary stages and the initial results show good agreement with analytical solution. Further development to the model, the solver and results will updated in future versions of the report.

Bibliography

[1] V. Svetovoy and I. Winter, "Model of the μ-flown microphone," *Sensors and Actuators A: Physical*, vol. 86, no. 3, pp. 171–181, Nov 2000. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924424700004544

[2] J. W. van Honschoten, G. J. M. Krijnen, V. B. Svetovoy, H.-E. de Bree, and M. C. Elwenspoek, "Analytic model of a two-wire thermal sensor for flow and sound measurements," *Journal of Micromechanics and Microengineering*, vol. 14, no. 11, p. 1468, 2004. [Online]. Available: http://stacks.iop.org/0960-1317/14/i=11/a=006

[3] R. Raangs, "Exploring the use of the microflown," Ph.D. dissertation, University of Twente, 12 2005.

[4] H.-E. de Bree, *The Microflown E-Book*. Microflown Technologies, Arnhem, 2007. [Online]. Available: http://www.microflown.com/library/books/the-microflown-e-book.htm

[5] A. Pierce, *Acoustics: An Introduction to Its Physical Principles and Applications*. Acoustical Society of America, 1989.

[6] F. Ihlenburg, *Finite element analysis of acoustic scattering*. Springer-Verlag New York, 1998, vol. 132. [Online]. Available: http://dx.doi.org/10.1007/b98828

[7] "SALOME: An Open-source Integration Platform for Simulation." [Online]. Available: http://salome-platform.org/

[8] M. S. Alnæs, J. Blechta, J. Hake, A. Johansson, B. Kehlet, A. Logg, C. Richardson, J. Ring, M. E. Rognes, and G. N. Wells, "The fenics project version 1.5," *Archive of Numerical Software*, vol. 3, no. 100, 2015.

[9] H. P. Langtangen and A. Logg, *Solving PDEs in Python*. Springer International Publishing, 2016. [Online]. Available: http://link.springer.com/10.1007/978-3-319-52462-7

[10] C. R. Llamas, "Optimization of an acoustic intensity P-U probe design using numerical methods," Master's thesis, University of A Coruña, 2018.

2. Benchmark case of model hierarchies: experimental-based validation of FSI simulations in blood pumps

Marco Martinolli[1], Christian Vergara[1], Pier Paolo Monticone[2], Luc Polverelli[2]

[1] *MOX, Dipartimento di Matematica, PoliMi*

[2] *CorWave Inc.*

Abstract. Unlike the rotary blood pumps currently available in the VAD market, the progressive wave pump developed in CorWave SA has the ability to provide true pulsatile blood flow, exerting low shear stress on red cells and thus reducing the risks associated with blood trauma and internal bleeding. The innovative pumping technology of the device relies on the mutual interaction between the blood flow and an oscillating polymer membrane, that is mathematically described by a Fluid-Structure Interaction hierarchical model. The benchmark consists of the validation of such model against experimental real data provided by CorWave, related to blood flow, pressure, membrane displacement and force. In particular, the validation approach will focus on a comparative analysis of the results obtained via simulations and via experiments, by looking at the pump characteristic PQ curves and at the recorded membrane position in time.

Keywords: Fluid-structure interaction, blood pumps, model validation.

2.1. Introduction

2.1.1. Introduction to Blood Pumps

Blood pumps are medical devices that are implanted in the chest of patients who have gone through episodes of heart failure or cardiac surgery. Belonging to the class of Left Ventricular Assist Devices (LVADs), their function is to take over the blood pumping capability of the heart by pumping mechanically part of the blood from the left ventricle into the ascending aorta. Specifically, the device is surgically implanted below the heart and connected to the apex of the left ventricle (i.e. its bottom part) to collect the blood from the ventricular chamber; then the blood enters in the pump and it is automatically ejected into the aorta via a flexible woven cannula. The blood pump is powered by external batteries via an electrical cable that passes through the patient's skin and transmits the input current to the pump engine. In addition, the motor functioning and the internal blood dynamics are controlled from the outside by means of inner sensors. The overall structure of an implantable blood pump is shown in Figure 2.1.

Blood pumps help to suppress the symptoms of heart failure and recover regular blood flow and they are able to maintain patients alive for several years (in the best cases for up to ten years). They can be used in a wide spectrum of treatments, depending on the health state of the patient (age, hypertension, smoking, cancer, etc.) and on the severity of the heart disease. In fact, blood pump treatments can be employed either on patients whose regular heart function may be recovered (bridge to recovery), patients that are eligible for transplant (bridge to transplantation), or for patients who are not transplant candidates (destination therapy). However, current rotary pumps available on the market involve also high risks of blood trauma, gastro-intestinal bleeding and consequent re-hospitalizations, because the high inertia of the continuous flow exposes the blood to high levels of shear stress and to red cell damage. Patient selection and implantation time are the most crucial factors that determine the success of LVAD therapy.

2.1.2. The Progressive Wave Pump developed in CorWave SA

CorWave SA is a biotech start-up founded in 2011 in Paris and that nowadays counts more than 50 employees, coming from bioengineering, electronic engineering, technical sciences and marketing. Its industrial activity

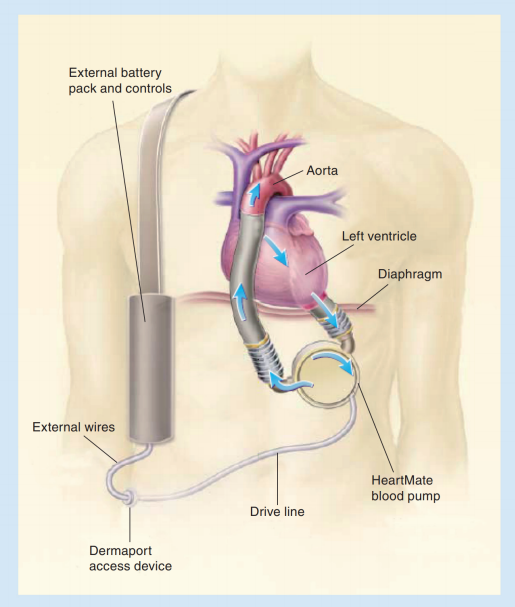


Figure 2.1: General components of an intracorporeal left ventricular assist device, specifically the HeartMate model. Figure taken from reference paper [1].

consists in the development of new technologies for cardiac support for patients who suffered from heart failure. In particular, the core project concerns the development of a new generation of implantable blood pumps, named progressive wave pumps, which employ a new pumping mechanism based on the interaction of the blood flow with an undulating elastic membrane (Figure 2.2). The relevance of this tecnology is that the membrane oscillations can be tuned in order to generate an output flow with the same pulsation of the native heart. Compared to the continuous flow rotary pumps currently available on the market, the progressive wave pump presents lower scales of inner velocity and shear stress and consequently should lead to a reduced risk of blood trauma, clotting and internal bleeding. As a result, the blood pump developed in CorWave is a promising product, because it presents important advantages in terms of performance, size and costs, with respect to the competitors active in the fast-growing market of acute cardiac assist devices.

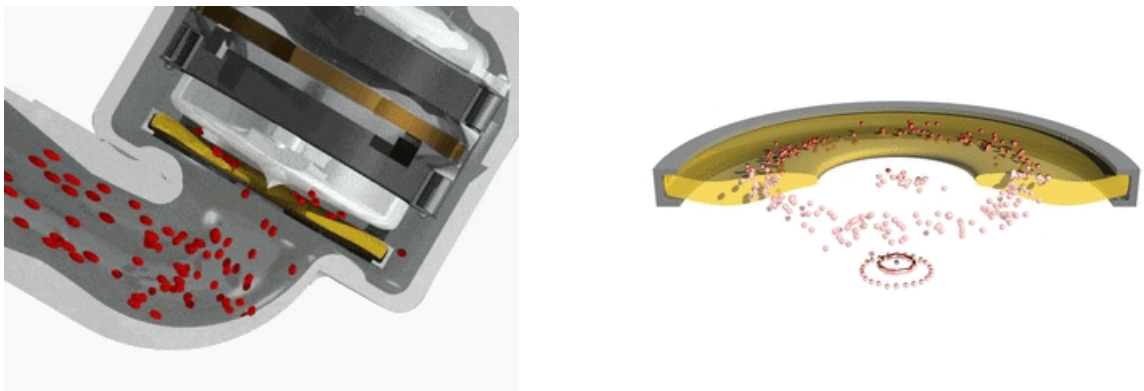


Figure 2.2: Left: Progressive wave blood pump developed in CorWave SA. Right: Detail of the blood cells gently pushed towards the inner orifice of the yellow undulating membrane. Figures included under permission of CorWave SA.

2.2. Literature and Modern Techniques

2.2.1. State-of-the-art in Blood Pumps

Heart disease is the major cause of death in the Western world and recent statistics highlighted that specifically in Europe the 47% of deaths is due to cardiovascular diseases [2]. In case of acute heart failure, heart transplantation remains the most successfull long-term treatment, showing a post-transplant survival rate of 68% at the first year and a half-life of 12 years conditional on surviving at least one year [3]. However, the limited availability of healthy transplantable hearts and the uneligibility of certain patients to enter in the waiting list have required an alternative to heart transplantation. Hence, in the last decade, mechanical circulatory support systems have become a life-saving option for many patients, being a valid treatment option as bridge to transplantation or destination therapy. In addition, blood pumps proved to bring important short-term benefits as bridge to recovery, when applied on patients who suffered from reversible cases of heart disease.
Being a such promising field of application, in the last decades there has been a continuosly increasing interest in research on VAD development, that has brought to several achievements in terms of size of the device (and whether it is implantable or not), its hydraulic efficiency and the risk of adverse events. In particular, Alba and Delgado [1] identified three generations or types of blood pumps:

1. Pulsatile volume displacement pumps, like pneumatic and implantable Thoratec pVAD and iVAD (Thoratec Laboratories Corp., CA, USA), HeartMate XVE (Thoratec Corp.) or LionHeart LVD2000 (Arrow International, PA, USA), which are equipped with inflow and outflow valves that allow to direct the blood dynamics inside the pump chamber in order to obtain a pulsatile output flow (see Figure 2.3, left). Even though they achieve a volume flow rate of up to 10 L/min, most of these pumps are rarely used due to the high internal inertia and the lower pulsation of the one of the native heart. In addition, these pumps are generally characterized by large volume - that makes implantation difficult and increase the risk of infections -, high incidence of malfunction and high working noise.
2. Axial continuous flow rotary pumps, like HeartMate II (Thoratec Inc.) and Jarvik 2000 (Jarvik Heart, Inc., NY, USA), which have an internal impeller suspended in the blood flow path by contact bearings, that imparts tangential velocity and kinetic energy to the blood. With respect to the pulsatile displacement pumps, they do not include inner valves and they eject blood flow in a continuous regime. Furthemore, they are characterized by smaller size, lower energy requirements and higher durability (Figure 2.3, right). On the other side, this second type of devices presents important limitations in hemolysis and thrombosis (due to the high gradients of velocity of the rotors), ventricular suction and pump stoppage. Thrombus formation can be prevented via anticoagulation, but increasing the risk of gastro-intestinal bleeding.
3. Centrifrugal continuous flow rotary pumps, which differ from the axial ones because they accelerate blood circumferentially and the rotor is suspended in the blood flow path with non-contact bearing design (Figure 2.4), using either hydrodynamic levitation like VentrAssist (Ventracor Ltd., Sydney, Australia), or magnetic levitation, like DuraHeart (Terumo, Inc., MI, USA) or an hybrid system, like HVAD (HeartWare Corp., FL, USA). Thus, by avoiding any mechanical contact, they reduce energy losses by friction or heat generation, improving the durability and the reliability of the device. Nonetheless, centrifrugal pumps have larger size than axial ones (especially in case of magnetically levitated impeller) and they maintain the risk of ventricular suction, blood trauma and bleeding complications because of the high speed of the rotor.

Therefore, displacement blood pumps have a disadvantageous size, while rotary pumps expose the patient to high risks of hemolysis because of their rotor speed. Thus, in the last years, a new prototype of pulsatile blood pump has been projected with the intention to overcome the drawbacks of previous LVAD models by providing physiological pulsatile blood flow with a minimum risk of blood trauma. This new generation of LVADs, currently under development in CorWave SA, employs an innovative pumping mechanism that relies on the interaction between blood and an elastic polymer membrane placed between two housing walls (or flanges). Specifically, this type of pumps works as follows: a magnetic or mechanical actuator excites

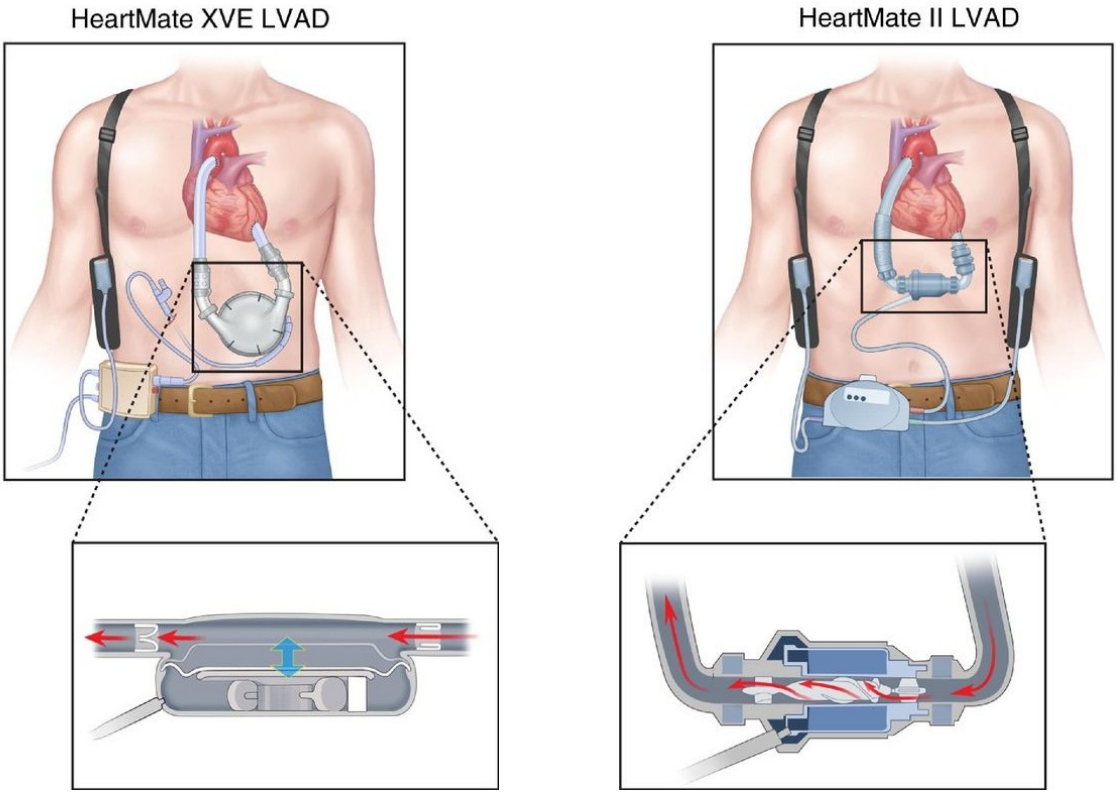


Figure 2.3: Direct comparison of the size and the mechanisms of action of a volume displacement (HeartMate XVE LVAD, left) and an axial flow rotary pump (HeartMate II LVAD, right). The displacement pump employs an electromagnetic pusher plate to cyclically change the volume of the pump chamber. The axial-flow pump uses a spinning impeller to continuously pushes the blood along a central shaft. Figure comes from Wilson et al., in the Journal of the American College of Cardiology [4].

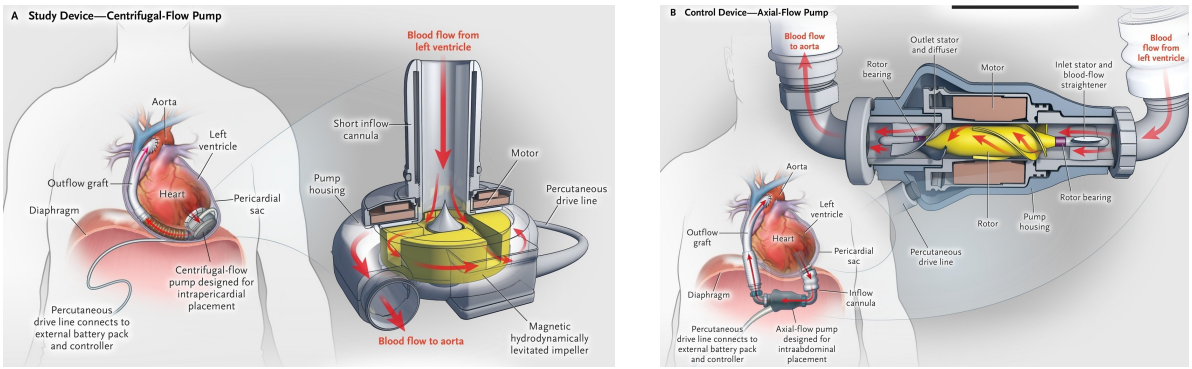


Figure 2.4: Comparison of the structure and the mechanics indide centrifrugal-flow (A) and axial-flow (B) rotary pumps. The centrifrugal pump pushes blood circumferentialy (see red lines) using a magnetically and hydrodynamically levited impeller; while the axial pump employs a spinning rotor mechanically sustained by contact bearing that moves the blood in axial direction. The images are taken from Rogers et al., in The New England Journal of Medicine [4].

the deformable membrane so that it start oscillating with a given amplitude and frequency and the resulting undulations generate progressive ways in the fluid-structure coupled system that produce an effective pulsatile pumping action. Hence, such pumps are called progressive wave pumps.
Compared with other LVAD models, progressive wave pumps are thought to provoke less trauma on the blood

cells because there are not contact bearings or critical mechanical parts like rotors or valves. Furthermore, the imposed oscillations can be optimized to easily realize a pulse-modulated form with similar pulsation of the one of the native heart.
Progressive wave pumps can be realized with two different designs: discoidal wave pump and tubular wave pump. In the former case, the actuator imparts transversal oscillations on the external circumference of the membrane disc, so that the generated waving motion pushes gently the blood cells from the perifery into the inner orifice of the membrane. This is the design of the wave pump developed in CorWave. In the latter case, the overall functioning is analogous, but the membrane has tubular shape and the direction of excitation is not radial (Figure 2.5). Although tubular wave pumps have not been yet realized, they are thought to be potentially better than discoidal ones, because of their axial pumping action and their limited space consumption [5].

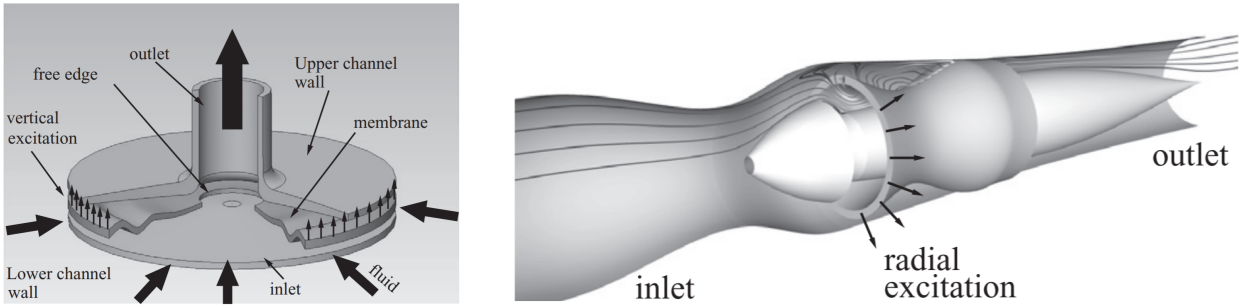


Figure 2.5: Discoidal (left) and tubular (right) design of progressive wave pumps. Illustrations taken from Perschall et al, in the Jouranl of Artificial Organs [5].

2.2.2. Computational Simulations as Support for LVAD Development

In order to reduce the risks underneath the usage of blood pumps, computational simulations are frequently used to reproduce the inner dynamics and investigate the causes of blood trauma [6, 7, 8]. In particular, such simulations are used to predict critical flow patterns within the pump - that may be difficult to detect experimentally -, where the blood flow develops turbolent dynamics and recirculation vortices with high inertia and stagnation time of the blood cells. In fact, these are typical factors that cause energy losses and favor adverse events, like hemolysis and thrombosis. Thus, the LVAD development team can take advantage of the valuable insights coming from simulation results to predict undesired dynamics and prevent them by properly adjusting the design of the pump.
In CorWave, they use commercial software packages like Comsol, ANSYS Fluent and ADINA, to simulate the fluid-structure interaction inside the progressive wave pump and test the effect of geometrical features (e.g. flange angle, membrane diameter, inflow area, etc.) on the global dynamics. Taking advantage of the cilindrical symmetry of the device, they perform 2D axis-symmetric simulations, in order to save computational time. However, experimental evidences highlighted that the vibrational modes of the elastic membrane do not always show a cylindrical symmetric behavior, but that there are not neglectable strain variations along the angular coordinate. Thus, in the benchmark discussed in section 2.4, the simulations will be carried out in the complete three-dimensional space, to capture the full membrane displacement.

2.3. Mathematical Modeling of the FSI problem

The dynamics inside a progressive wave pump are driven by the interaction between the blood flow and the elastic membrane. This mutual interaction can be mathematically described in the context of the Fluid-Structure Interaction (FSI) modelling, providing proper coupling conditions at the fluid-structure interface.

2.3.1. The Computational Domain

Before detailing to the mathematical formulation of the FSI problem, we need to properly define the geometry of the pump considered for the simulations.
The structure of the pump chamber is essentialy divided in two parts (Figure 2.6):

1. a magnetic or mechanic actuator (red rectangle), which converts sinusoidal input currents into mechanical oscillations of the membrane disc. The transmission of the displacement to the membrane is done via a system of pins and supports of titanium - from now on this structure is called titanium holder - which ensures that the motion is uniquely in the normal direction to the membrane plan. Thus, it transforms electric power into mechanical power.
2. the pump head (green rectangle), where the blood flow interacts with the oscillating membrane before being ejected into the aorta. Consequently, in this passage the mechanical power previously provided by the actuator is transformed into hydraulic power.

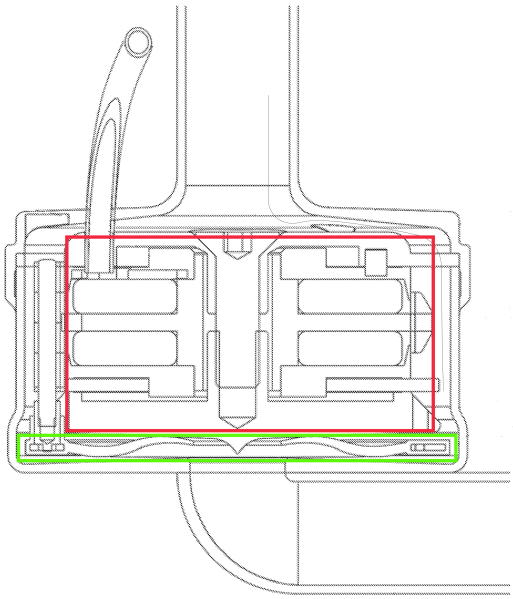


Figure 2.6: Structure of a prototype of progressive wave pump developed in CorWave. The actuator (red rectangle) imposes oscillating motion to the membrane, that interacts with the blood flow in the pump head (green rectangle). Modified version of a figure taken from CorWave licence No. WO/2017/178959, *Implantable pump system having an ondulating membrane*, 19.10.2017.

For our simulations we are mainly interested in the dynamics occuring inside the pump head, where the fluid-structure interaction between the blood and the membrane leads to a pulsatile output blood flow. As a result, in this benchmark we restrict the computational domain to the sole pump head and, as we will see in the next sections, we model the mechanical effect of the actuator using proper boundary conditions applied on the membrane disc.

The geometry of the pump head is sketched in Figure 2.7. The domain Ω of the pump head is the union of the fluid domain $\Omega_t^f = \Omega^f(t)$ and the structure domain $\Omega_t^s = \Omega^s(t)$, which are time-dependent because they both change in time due to the motion imposed by the actuator and the mutual interaction between the blood fluid and the membrane structure. As a consequence, the fluid-structure interface $\Sigma_t = \Sigma(t)$ changes in time accordingly. Thus, we have $\Omega = \Omega_t^f \cup \Omega_t^s$ and $\Sigma_t = \Omega_t^f \cap \Omega_t^s, \quad \forall t \geq 0$.

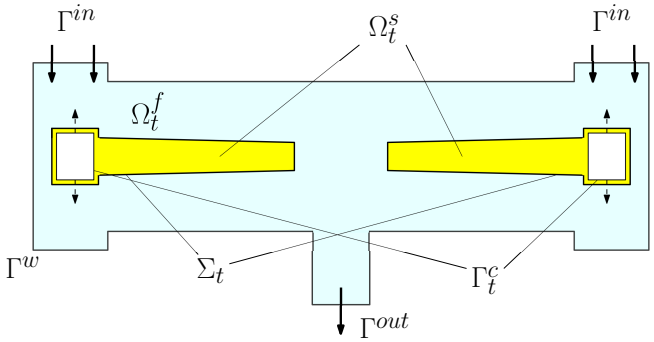


Figure 2.7: Sketch of the pump head domain.

The boundary of the pump housing $\partial\Omega$ is composed by the following subportions:

- the inlet boundary Γ^{in}, where the blood fluid enters in the computational domain;
- the output boundary Γ^{out}, where the blood is ejected from the pump chamber into the outflow cannula;
- the boundary of the inner titanium cavity Γ_t^c, that corresponds to the border of the space occupied by the titanium holder inside the membrane and that transmits to the former the oscillating motion prescribed by the actuator;
- the wall boundary Γ^w, which consists of all the remaining parts of the pump housing, including flanges and steps; i.e. we have $\Gamma^w = \partial\Omega \setminus (\Gamma^{in} \cup \Gamma^{out} \cup \Gamma_t^c)$.

The position of the boundary of titanium cavity Γ_t^c changes in time, oscillating along the vertical direction z together with the membrane, whereas the frontier of the domain $\partial\Omega$ is fixed in time.
Notice that, with respect to the description of the pump geometry in the previous paragraph, we omit the titanium holder structure in the computational domain. This is justified by observing that its movement is purely rigid. Thus, since the movement is determined by the actuator, its effect on the membrane can be modeled by means of a Dirichlet boundary condition directly on the boundary Γ_t^c.

2.3.2. Formulation of the FSI Problem

In the context of Fluid-Structure Interaction problems, the dynamics are represented by a system of Partial Differential Equations (PDEs) which describe separately the behaviour of the fluid and of the structure in the respective domains, while coupling conditions define their interaction at the interface.
Thus, FSI problems are instances of hiearchical or multilevel model, which have two basic components: the fluid dynamics model and the structure model (Figure 2.8). On one side, the blood fluid is modeled using Navier-Stokes Equations (NS), which correspond to the conservation laws of mass and momentum for an incompressible viscous newtonian fluid. On the other side the membrane motion is mathematically represented through the elasto-dynamics equations. Finally, the two subproblems are coupled at the interface between blood and membrane to impose the continuity of velocity and of traction forces. This is indicated in the literature as physical coupling of the system. Furthermore, the problem presents an additional coupling condition, due to the mutual dependancy of the domains. In fact, the structure displacement does not only determine the position of the structure domain Ω_t^s, but it determines also the configuration of the fluid domain Ω_t^f by means of a perfect adherence condition between the two subdomains holding at the interface (geometrical coupling).

Specifically, the FSI problem reads as follows: for each time $t > 0$, find fluid velocity and pressure $(\mathbf{u}(t), p(t))$:

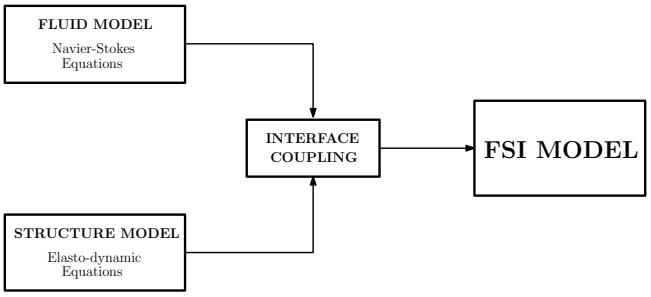


Figure 2.8: Multi-level representation of the Fluid-Structure Interaction building blocks.

$\Omega_t^f \times \Omega_t^f \to \mathbf{R}^3 \times \mathbf{R}$ and membrane displacement $\hat{\mathbf{d}}(t) : \widehat{\Omega}^s \to \mathbf{R}^3$, such that:

$$\begin{cases} \rho_f \left(\partial_t \mathbf{u} + \mathbf{u} \cdot \nabla \mathbf{u}\right) - \nabla \cdot \mathbf{T}^f(\mathbf{u}, p) = \mathbf{0} & \text{in } \Omega_t^f, \quad (2.1a) \\ \nabla \cdot \mathbf{u} = 0 & \text{in } \Omega_t^f, \quad (2.1b) \\ \rho_s \partial_{tt} \widehat{\mathbf{d}} - \nabla \cdot \widehat{\mathbf{T}}^s(\widehat{\mathbf{d}}) = \mathbf{0} & \text{in } \hat{\Omega}^s, \quad (2.1c) \\ \mathbf{u} = \partial_t \mathbf{d} & \text{on } \Sigma_t, \quad (2.1d) \\ \mathbf{T}^f(\mathbf{u}, p)\, \mathbf{n}^f = -\mathbf{T}^s(\mathbf{d})\, \mathbf{n}^s & \text{on } \Sigma_t \quad (2.1e) \end{cases}$$

where Ω_t^f depends through the geometrical coupling on the displacement $\boldsymbol{d}$, i.e. $\Omega_t^f = \Omega_t^f(\boldsymbol{d})$. Here, ρ_f and ρ_s are the mass densities of fluid and structure, $\mathbf{n}^f$ and $\mathbf{n}^s$ are the external normal vectors from fluid and structure domains, satisfying $\mathbf{n}^f = -\mathbf{n}^s = \mathbf{n}$, $\mathbf{T}^f(\mathbf{u}, p) = -p\mathbf{I} + 2\mu_f \mathbf{D}(\mathbf{u})$ is the Cauchy stress tensor for a viscous newtonian fluid with viscosity μ_f and symmetric operator $\mathbf{D}(\mathbf{w}) = \frac{1}{2}(\nabla \mathbf{w} + \nabla \mathbf{w}^T)$, and $\widehat{\mathbf{T}}^s(\hat{\mathbf{d}})$ is the first Piola-Kirkhhoff tensor of the structure, such that $\widehat{\mathbf{T}}^s(\widehat{\mathbf{d}}) = J\, \mathbf{T}^s(\mathbf{d}) \mathbf{F}^{-T}$ with $\mathbf{T}^s$ being the solid Cauchy stress tensor, $\mathbf{F} = \nabla \mathbf{x}$ the gradient of deformation and $J = \det \mathbf{F}$ its determinant.
In line with standard continuum mechanics theory, the fluid problem has been posed in an Eulerian framework, whereas the structure problem has been formulated in a Lagrangian framework, indicating with superscript $\widehat{\cdot}$ the quantities in the reference configuration at time 0, i.e. in $\hat{\Omega}^s = \Omega_0^s$. In case we assume that the material is linearly elastic and isotropic the solid stress tensor $\mathbf{T}^s(\mathbf{d})$ can be explicitly written using the Hooke's law as $\mathbf{T}^s(\mathbf{d}) = \lambda_s(\nabla \cdot \mathbf{d})\mathbf{I} + 2\mu_s \mathbf{D}(\mathbf{d})$ where $\lambda_s, \mu_s > 0$ are the Lamé parameters of the elastic material.

Finally, the problem is closed with proper initial conditions, e.g. imposing null fluid and structure velocity and membrane displacement, and the following boundary conditions:

$$\begin{cases} \mathbf{u} = \mathbf{u}_{in} & \text{on } \Gamma^{in}, \quad (2.2a) \\ \mathbf{T}^f(\mathbf{u}, p)\, \mathbf{n}^f = \mathbf{0} & \text{on } \Gamma^{out}, \quad (2.2b) \\ \mathbf{u} = \mathbf{0} & \text{on } \Gamma^{w}, \quad (2.2c) \\ \mathbf{d}(t) = \Phi\, \sin(2\pi\, f\, t)\, \mathbf{e}_z & \text{on } \Gamma_t^c \quad (2.2d) \end{cases}$$

Notice that the forcing terms in the dynamics of the coupled problem are given by the inlet velocity $\mathbf{u}_{in}$ and by the vibrations induced by the actuator, with predefined frequency f and amplitude (or stroke) Φ.

2.4. Benchmark Plan

The FSI modeling approach illustrated in Section 2.3 describes the mutual interaction between the blood fluid and the oscillating membrane that arise inside the progressive wave pump developed in CorWave. Therefore,

such model has to be validated against real data provided by the company, in order to understand whether the simulation results can be actually used to predict the inner dynamics and optimize the design of the pump geometry accordingly. Thus, the benchmark consists of an experimental-based validation of the FSI model, with a similar approach to the one used in [5].

2.4.1. Description of Real Data

The experimental data that can be used for model validation come from measurements of physical quantities related to the blood flow, pressure, membrane displacement and force. Specifically, CorWave can provide the following experimental data:

- measurements of the volume flow rate Q, made with an ultrasonic flowmeter clamped at the inlet of the testing pipes, with a sampling rate of 10 Hz;
- pressure data P, obtained with pressure sensors placed at the inlet and the outlet in order to measure the pressure rise ΔP inside te pump;
- based on the fact that CorWave developed a mechanical actuator that mimics the implantable pump, a strain gauge force sensor placed below the membrane can measure the force necessary to impose a sinusoidal displacement on the membrane holder;
- displacement values are available for both the active part of the membrane as well as the membrane holder. In particular the motion of the active part of the membrane is acquired with a high speed camera, while the displacement of the membrane holder is recorded through a laser.

In addition, cross data can be derived from the raw measurements described above. These include hydraulic power and efficiency, PQ and HQ curves, oscillation frequency, membrane position in time and membrane cross section versus time.
In particular, one of the main goals in CorWave is to maximize the hydraulic efficiency η_h of the pump. In fact, this parameter measures the efficiency of the device to transform the mechanical power W coming from the actuator into hydraulic power H:

$$\eta_h = \frac{H}{W} = \frac{\Delta P\, Q}{\frac{\int_a^b F\, dx}{\Delta T}} \tag{2.3}$$

where the hydraulic power H of the pump is computed as the product of the pressure difference ΔP between inlet and outlet and the ejected volume flow rate Q and the integral of F corresponds to the mechanical work of the actuator to move the membrane from position a to position b in a time interval ΔT.
Furthermore, PQ or HQ curves are standard data representations used to study the performance of the pump and to validate the FSI model. In these relationships, the pressure rise (here denoted simply as P) or the hydraulic power (H) are related with each output volume flow rate (Q). In CorWave, such curves are obtained for different frequencies f of mechanical excitation of the elastic membrane, varying in a range between 60 Hz until up to 120 Hz.
Notice that testing in CorWave is performed on a mix of water and glycerin (39% in weight) at 37 °C in order to mimic the blood density and viscosity.

2.4.2. Validation of the FSI Model

The goal of the benchmark is to validate the FSI model against the real experimental data detailed in the previous paragraph.
A first point of our validation analysis is to compare the PQ (or HQ) curves obtained by the numerical results with the experimental pump characteristic curves. In Figure 2.9, there are the results of this type of comparative study for model validation, performed by Perschall et al. [5]. In the left panel of Figure 2.9, we see the compar-

ison between the numerical results (circles) and the experimental findings (dots) for different input frequencies of membrane oscillation (80, 90, 100 and 110 Hz). Since there is a systematic discrepancy between the curves, the authors propose a linear correction function for the numerical values in order to compute modified values of the pressure (squares) that are in a better agreement with the data. In the right panel of Figure 2.9, the model is validated also across different geometries of the pump. In fact, the numerical simulations are carried out for two different geometries, the reference one (dashed line) and a slightly modified version of it (solid line), and two different input frequencies (80 and 100 Hz).
In our benchmark we want to adopt a similar validation strategy: compare the pump characteristic PQ (or HQ) curves obtained from simulations and from experiments to validate the FSI model, so that it can be used to test pump geometries with different domain and model parameters. Specifically, example of feature parameters that we are interested to vary in our simulations are:

- the angle of the pump flanges,
- the length of the vertical separation between the flanges (said gap),
- the stroke of the membrane oscillations,
- the membrane diameter,
- the inflow area,
- the outflow area.

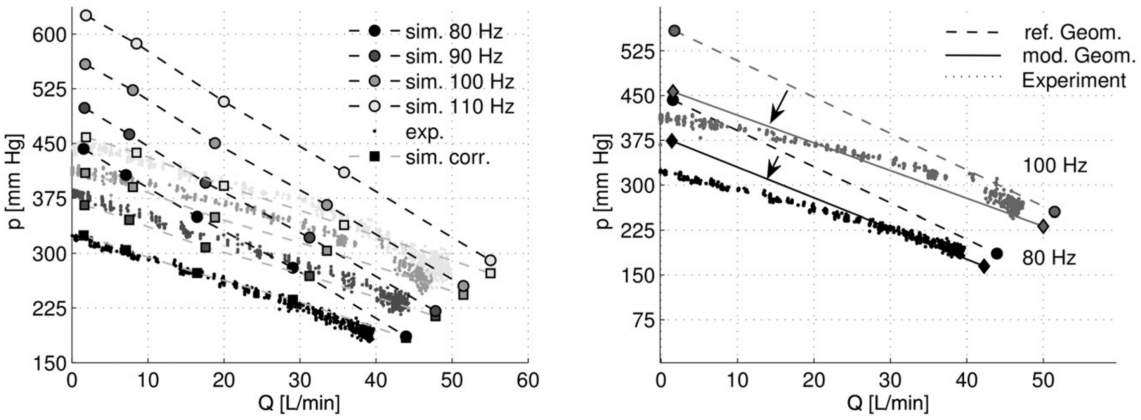


Figure 2.9: Example of validation of FSI model inside a progressive wave blood pump with discoidal structure, taken from Perschall et al. [5]. Left: Comparison of simulation results for different frequencies (sim: circles) with experimental PQ curves (exp: dots). Square symbols correspond to modified simulation results (sim. cor.), obtained by the application of the linear correction function $p_{corr} = a\,p + b\,Q$ in order to mimic better the experimental data. $a = 0.73$, $b = 1.5\,10^{-5}$ bar/(L/min). Right: Similar comparative study performed for two oscillation frequencies on two different geometries of the pump, the reference one (ref. Geom.: dashed line) and a similar one where the angle and the height of the channel are slightly changed (mod. Geom.: solid line).

A second important point of our model validation of the progressive wave pump is based on the information coming from the membrane displacement. In particular, the spatial coordinates of the membrane position in time obtained by the recordings will be compared with the simulated displacement predicted by the FSI model. Moreover, since the implemented simulations are going to be carried out in the full three-dimensional space, the validation will specifically address the three-dimensional modes of deformation, comparing the recordings with simulation results.

Finally, the benchmark will also include a comparison of the performance of our implementation with the

bi-dimensional axis-symmetric simulations performed in COMSOL currently employed in CorWave, in terms of both accuracy and computational efficiency.

The FSI simulations for the benchmark will be performed using an advanced numerical approach, named Extended Finite Element Method (X-FEM) [9, 10], that is briefly explained in the Appendix A.1. The implementation will be carried out in the LIbrary of Finite Elements V (LIFEV) [11], an academic library that is particularly suitable for cardiovascular applications and has been developed by MOX-PoliMi (Milan), CMCS-EPFL (Lausanne), ESTIME-INRIA (Paris) and E(CM)2-Emory University (Atlanta).

Bibliography

[1] A. C. Alba and D. H. Delgado, "The future is here: ventricular assist devices for the failing heart," *Expert Review of Cardiovascular Therapy*, vol. 7, no. 9, pp. 1067–1077, 2009.

[2] D. Mozaffarian, E. J. Benjamin, A. S. Go, D. K. Arnett, M. J. Blaha, M. Cushman, S. R. Das, S. de Ferranti, J. P. Després, H. J. Fullerton *et al.*, "Heart disease and stroke statistics-2016 update a report from the american heart association," *Circulation*, vol. 133, no. 4, pp. e38–e48, 2016.

[3] J. D. Christie, L. B. Edwards, A. Y. Kucheryavaya, C. Benden, F. Dobbels, R. Kirk, A. O. Rahmel, J. Stehlik, and M. I. Hertz, "The registry of the international society for heart and lung transplantation: twenty-eighth adult lung and heart-lung transplant report—2011," *The Journal of Heart and Lung Transplantation*, vol. 30, no. 10, pp. 1104–1122, 2011.

[4] J. G. Rogers, F. D. Pagani, A. J. Tatooles, G. Bhat, M. S. Slaughter, E. J. Birks, S. W. Boyce, S. S. Najjar, V. Jeevanandam, A. S. Anderson *et al.*, "Intrapericardial left ventricular assist device for advanced heart failure," *New England Journal of Medicine*, vol. 376, no. 5, pp. 451–460, 2017.

[5] M. Perschall, J. B. Drevet, T. Schenkel, and H. Oertel, "The progressive wave pump: numerical multiphysics investigation of a novel pump concept with potential to ventricular assist device application," *Artificial organs*, vol. 36, no. 9, 2012.

[6] A. Untaroiu, H. G. Wood, P. E. Allaire, A. L. Throckmorton, S. Day, S. M. Patel, P. Ellman, C. Tribble, and D. B. Olsen, "Computational design and experimental testing of a novel axial flow lvad," *ASAIO journal*, vol. 51, no. 6, pp. 702–710, 2005.

[7] S. Pirker, H. Schima, and M. Stoiber, "Numerical simulation of hemolysis: A comparison of lagrangian and eulerian modelling," *WIT Transactions on Biomedicine and Health*, vol. 8, pp. 361–370, 2005.

[8] D. Carswell, D. McBride, T. Croft, A. Slone, M. Cross, and G. Foster, "A cfd model for the prediction of haemolysis in micro axial left ventricular assist devices," *Applied Mathematical Modelling*, vol. 37, no. 6, pp. 4199 – 4207, 2013.

[9] E. Burman and M. A. Fernández, "An unfitted nitsche method for incompressible fluid–structure interaction using overlapping meshes," *Computer Methods in Applied Mechanics and Engineering*, vol. 279, pp. 497–514, 2014.

[10] S. Zonca, C. Vergara, and L. Formaggia, "An unfitted formulation for the interaction of an incompressible fluid with a thick structure via an xfem/dg approach," *SIAM Journal on Scientific Computing*, vol. 40, no. 1, pp. B59–B84, 2018.

[11] L. Bertagna, S. Deparis, L. Formaggia, D. Forti, and A. Veneziani, "The lifev library: engineering mathematics beyond the proof of concept," *arXiv preprint arXiv:1710.06596*, 2017.

Part II.

Model order reduction methods

3. Model order reduction for parametric high dimensional interest rate models in the analysis of financial risk

Andreas Binder[1], Onkar Jadhav[2], Volker Mehrmann[2]

[1] *MathConsult*

[2] *Technische Universität Berlin*

Abstract. The European Parliament has introduced regulations (No 1286/2014) on packaged retail investment and insurance products (PRIIPs). According to this regulation, PRIIP manufacturers must provide a key information document (KID) describing the risk and the possible returns of these products. The formation of a KID requires expensive valuations rising the need for efficient computations. To perform such valuations efficiently, we establish a model order reduction approach based on a proper orthogonal decomposition (POD) method. The study involves the computations of high dimensional parametric convection-diffusion reaction partial differential equations. This report provides a collection of benchmark cases that are essential for validating the developed numerical algorithms and their software implementations. The first benchmark case solves the problem of parameter calibration for financial models. Once we obtain the parameters for the desired model, the second benchmark case serves to verify the implemented model order reduction approach.

Keywords: PRIIP, model order reduction, POD-greedy approach, interest rate modeling.

3.1. Introduction

3.1.1. Industrial partner

MathConsult GmbH is a research company developing mathematics-based solutions for the producing industry and for financial institutions. It is located in Linz, Austria and headed by Dr. Andreas Binder. To analyze the financial risk of the assets and instruments, MathConsult has been developing 'UnRisk' software. UnRisk provides robust techniques for the calibration of models and the valuation of financial instruments using highly advanced numerical schemes. The analysis of financial instruments involves expensive computations, which motivates to design a cost-effective model order reduction (MOR) approach. MOR can cheaply solve costly problems with the same accuracy as that of the original system.

3.1.2. Introduction and Motivation

Packaged retail investment and insurance products (PRIIPs) are at the essence of the retail investment market. PRIIPs offer considerable benefits for retail investors which make up a market in Europe worth up to €10 trillion. However, the product information provided by financial institutions to investors can be overly complicated and contains confusing legalese. To overcome these shortcomings, the EU has introduced new regulations on PRIIPs (European Parliament Regulation (EU) No 1286/2014) [1]. According to this regulation, a PRIIP manufacturer must provide a key information document (KID) for an underlying product that is easy to read and understand. The PRIIPs include interest rate derivatives such as the interest rate cap and floor [2], interest rate swaps [3] etc.

A KID includes a section about '*what could an investor get in return?*' for the invested product which requires costly numerical simulations of financial instruments. This work evaluates interest rate derivatives based on the dynamics of the short-rate models [4]. For the simulations of short-rate models, techniques based on discretized convection-diffusion reaction partial differential equations (PDEs) are often superior [5]. We implement the finite difference method (FDM) for such simulations [6]. The FDM method has been proven to be efficient for solving the short-rate models [7, 8, 9]. The model parameters are usually calibrated based on market structures like yield curves, cap volatilities, or swaption volatilities [4]. The regulation demands to perform yield curve

simulations for at least 10,000 times. A yield curve shows the interest rates varying with respect to the 20-30 time points known as 'Tenor points'. These time points are the contract lengths of an underlying instrument. The calibration based on several thousand simulated yield curves generates a high dimensional model parameter space as a function of these tenor points. We need to solve the high dimensional model (HDM) obtained by discretizing the short-rate PDE for such a high dimensional parameter space [10]. These simulations are computationally costly and additionally, have the disadvantage of being affected by the so-called *curse of dimensionality* [11].
To avoid this problem, we establish a parametric model order reduction (MOR) approach based on the proper orthogonal decomposition (POD) method [12, 13]. The method is also known as the Karhunen-Loéve decomposition [14] or principal component analysis [15] in statistics. A combination of Galerkin projection approach and POD creates a powerful method for generating a reduced order model from the HDM that has a high dimensional space [16]. POD generates an optimally ordered orthonormal basis in the least squares sense for a given set of computational data. Further, the reduced order models are obtained by projecting a high dimensional system onto a low dimensional subspace obtained by truncating the optimal basis. The selection of the data set plays an important role and most prominently obtained by the *method of snapshots* introduced in [17]. In this method, the optimal basis is computed based on the set of state solutions. These state solutions are known as snapshots and are calculated by solving the HDM for some parameter values.

3.2. Mathematical Description

The management of interest rate risks, i.e., the control of change in future cash flows due to the fluctuations in interest rates is of great importance. Especially, the pricing of products based on the stochastic nature of the interest rate creates the necessity for mathematical models. Before introducing the stochastic differential equations (SDEs), we present some basic concepts required to construct the short-rate models.

3.2.1. Bank Account and Short-Rate

First we introduce the definition for a bank account or also called as a money-market account. When investing a certain amount in a bank account, we expect it to grow at some rate as time pass by. A money-market account represents a risk-less investment with a continuous profit at a risk-free rate.

Definition 1. ***Bank account (Money-market account).*** *Let $B(t)$ be the value of a bank account at time $t \geq 0$. We assume that the bank account evolves according to the following differential equation with $B(0) = 1$,*

$$dB(t) = B(t)r_t dt, \tag{3.1}$$

where r_t is a short-rate. Which leads

$$B(t) = \exp\left(\int_o^t r_s ds\right). \tag{3.2}$$

According to the above definition, investing a unit amount at time $t = 0$ yields the value in (3.2) at time t, and r_t is the short-rate at which the bank account grows.

3.2.2. Yield Curve

A fundamental curve that can be obtained from the market data is the zero coupon curve or also known as the yield curve. It depends on maturity dates and interest rates for an underlying instrument.

Definition 2. ***Maturity or maturity date.*** *A maturity or maturity date is the final payment date of a financial instrument, at which the principal (along with the remaining interest) is due to be paid.*

Definition 3. ***Yield curve.*** *The yield curve or a zero-copoun curve at time t is a curve showing interest rates plotted against different maturities for a similar financial instrument.*

Such a curve is also known as 'the term structure of interest rates' at time t. It is a plot of simply-compounded interest rates for all maturities T up to one year and of annually compounded rates for maturities T larger than one year. The maturity time points are also known as tenor points. We plot the yields at each tenor point T for $0 \leq T \leq T^*$ where T^* is the last maturity date. An example of such a curve is shown in Fig. 3.1.

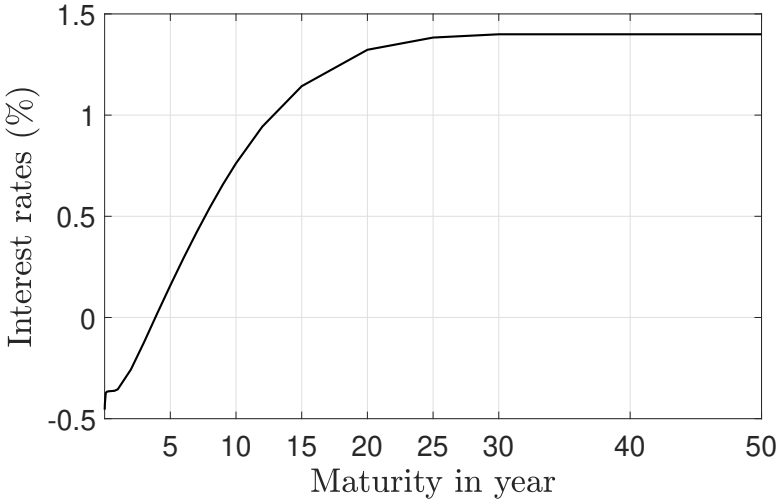


Figure 3.1: Sample yield curve

3.2.3. Options

An option is a right, but not an obligation, to purchase or sell the underlying instrument at some time $t \geq 0$ at a pre-defined price. This price is known as a *strike price*. There are two types of options: call options and put options. The call option gives a buyer the right to buy the asset at an agreed price, whereas the put option gives the right to sell. The payoff of an option is its value at the time of its exercise. In the case of a call option with a strike price K and an underlying instrument with a value V at the expiry T, the payoff for a call option C_V is given as

$$\begin{aligned} C_V &= \begin{cases} V - K, & \text{if } V > K \\ 0, & \text{if } V \leq K \end{cases} \\ C_V &= \max(V - K, 0) = (V - K)^+. \end{aligned} \tag{3.3}$$

For a call option, in the case of $V \leq K$, the asset can be purchased at a lower price than K in the market, and the buyer will not exercise the call option.
Similarly, for the put option, the payoff will be,

$$P_V = \max(K - V, 0) = (K - V)^+. \tag{3.4}$$

For a put option, in case of $V \geq K$, the asset can be sold at a higher price than K in the market.

3.2.4. Interest Rate Cap and Floor

Assume that a borrower pays a floating interest rate (e.g., quarterly payments of Euribor3M[1]) on some outstanding loan. The borrower may not be able to pay the interest payments if the Euribor3M rises substantially so, it is agreed that the interest rate shall not exceed 5 %. That means the interest rate is capped at 5 %.
Typically, the interest rate at time t_i is fixed by the interest rate at time t_{i-1}. For our example, the interest rate

[1]The Euro Interbank Offered Rate (Euribor) is a daily reference rate, published by the European Money Markets Institute, based on the averaged interest rates at which Eurozone banks offer to lend unsecured funds to other banks in the euro wholesale money market.

can be written as

$$\min(\text{Euribor3M}(t_{i-1}), 5\%).$$

Here Euribor3M is know as the reference interest rate R. The payoff $(R-K)^+$ is called as a 'caplet' with the strike price K, and is similar to the call option. For the fixed income products, the interest payments are made on scheduled dates $(t_1, \cdots, t_n)$. The collection of caplets for these single payments into one contract gives a so-called 'cap'. A cap has a discounted payoff

$$\sum_{i=1}^{n} DF(t_i)(t_i - t_{i-1})(R(t_{i-1}) - K)^+$$

Similarly, the interest rate floor is a contract in which the buyer receives payments at the end of each period in the case of the reference interest rate is below the strike price. The payoff $(K-R)^+$ gives a 'floorlet' and collection of such floorlets gives a floor. The floorlet is similar to the put option.

3.2.5. No-Arbitrage Pricing

Consider a contract which pays cash flow C_k at time $t = k$, where $k = 1, \cdots, T$. You have to pay the price p at $t = 0$ to get this contract. The no-arbitrage condition bounds this price p for the underlying contract. There are two types of no-arbitrage conditions: a weak no-arbitrage and a strong no-arbitrage.

- Weak no-arbitrage:

The weak no-arbitrage condition state that, if the cash flow C_k is non-negative for all time points $k \geq 1$, then the price p for this contract must be greater than or equal to zero. That is

$$\text{If } C_k \geq 0 \text{ for all } k \geq 1 \text{ then } p \geq 0.$$

- Strong no-arbitrage:

The strong no-arbitrage condition state that, if the cash flow C_k is non-negative for all time points $k \geq 1$, and there exist some time t_l, in future, such that $C_l > 0$, then the price p must be strictly positive.

$$\text{If } C_k \geq 0 \text{ for all } k \geq 1 \text{ and } C_l > 0 \text{ for some } l \in k \text{ then } p > 0.$$

These no-arbitrage conditions have the following importance. In a market, if there are contracts for which you get something for nothing, then prices for these contracts will be adjusted based on the arbitrage conditions such that one can not take the advantage of this unbalance.
Suppose the price $p < 0$. Under such a condition, the buyer has to pay $-p$ price to purchase the contract. In other words, we can say that the buyer will receive an amount p. Also, the seller can keep on increasing the price p as long as $p \leq 0$, and still buyers will be interested in purchasing the contract. The no-arbitrage condition avoids this problem and put a bound on the price p.

3.2.6. Short-Rate Models

When working with the interest-rate products, the study about the variability of interest rates is essential. Therefore, it is necessary to consider the interest rate as a stochastic process.
Let S be the price of a stock at the end of the nth trading day. The daily return from day n to day $n+1$ is given by, $(S_{n+1} - S_n)/S_n$. In general, it is common to work with log returns since the log return of k days can be easily computed by adding up the daily log returns

$$\log(S_k/S_0) = \log(S_1/S_0) + \cdots + \log(S_k/S_{k-1}). \tag{3.5}$$

Based on the assumption that the log returns over disjoint time intervals are stochastically independent, and are equally distributed, the central limit theorem [18] of the probability theory implies that the log returns are normally distributed [19].
So, it is necessary to define a stochastic model in a continuous time in which log returns over arbitrary time intervals are normally distributed. The Brownian motion or also known as the Wiener process provides these properties [20].

3.2.6.1. Brownian Motion

Definition 4. ***A Brownian motion.*** *A standard Brownian motion is a stochastic process $W(t)$ where $t \in \mathbb{R}$, that is, a family of random variables $W(t)$, indexed by non-negative real numbers t with the following properties:*

- *At $t = 0$, $W_0 = 0$.*
- *With probability 1, the function $W(t)$ is countinous in t.*
- *For $t \geq 0$, the increment $W(t+s) - W(s)$ is normally distributed with mean 0 and variance t, i.e.,*

$$W(t+s) - W(s) \sim N(0,t). \tag{3.6}$$

- *For all n and times $t_0 < t_1 < \cdots < t_{n-1} < t_n$, the increments $W(t_j) - W(t_{j-1})$ are stochastically independent.*

Another important property of the Brownian motion is [21]

$$(dW(t))^2 = dt. \tag{3.7}$$

Equation (3.7) says that the $(dW(t))^2$ is a deterministic quantity but not random, and has a magnitude of dt.
Based on the definition of the Brownian motion, we can establish a stochastic differential equation. Consider an ordinary differential equation (ODE)

$$\frac{dx(t)}{dt} = a(t)x(t), \tag{3.8}$$

with an initial condition $x(0) = x_0$.
When we consider ODE (3.8) with an assumption that the parameter $a(t)$ is not a deterministic but rather a stochastic parameter, we get an SDE.
The stochastic parameter $a(t)$ is given as [21]

$$a(t) = f(t) + h(t)\Xi(t), \tag{3.9}$$

where $\Xi(t)$ is a white noise process.
Thus, we get

$$\frac{dX(t)}{dt} = f(t)X(t) + h(t)X(t)\Xi(t). \tag{3.10}$$

This equation is known as Langevin equation [22]. Where $X(t)$ is a stochastic variable having the initial condition $X(0) = X_0$ with a probability one. The Langevin force, $\Xi(t) = dW(t)/dt$, is a fluctuating quantity having Gaussian distribution.
Substituting, $dW(t) = \Xi(t)dt$, in (3.10), we get

$$dX(t) = f(t)X(t)dt + h(t)X(t)dW(t) \tag{3.11}$$

In general form an SDE is given as

$$dX(t) = f(t, X(t))dt + g(t, X(t))dW(t), \tag{3.12}$$

where $f(t, X(t)) \in \mathbb{R}$, and $g(t, X(t)) \in \mathbb{R}$ are sufficiently smooth functions.

3.2.6.2. Ito's Lemma

Consider a function $\xi(x, y)$ which depend on variables x and y. According to the chain rule for total derivatives, we can write

$$d\xi = \frac{\partial \xi}{\partial x} dx + \frac{\partial \xi}{\partial y} dy. \tag{3.13}$$

However, what if we have a function ξ which depends not only on a real variable t but also on a stochastic random variable $X(t)$, i.e., $\xi = f(t, X(t))$?
Ito's lemma provides an answer to this problem. It serves as the stochastic calculus counterpart of the chain rule.

Theorem 1. ***Ito's Lemma.*** *[23] Let $\xi(X(t), t)$ be a sufficiently smooth function and let the stochastic process $X(t)$ is given by (3.12), with probability one,*

$$\begin{aligned} d\xi(X(t), t) = &\left(\frac{\partial \xi}{\partial X(t)} f(X(t), t) + \frac{\partial \xi}{\partial t} + \frac{1}{2} \frac{\partial^2 \xi}{\partial X^2(t)} g^2(X(t), t) \right) dt \\ &+ \frac{\partial \xi}{\partial X(t)} g(X(t), t) dW(t). \end{aligned} \tag{3.14}$$

Proof. We expand $d\xi$ using the second order Taylor series expansion as

$$\begin{aligned} d\xi(X(t), t) &= \xi(X(t + dt), t + dt) - \xi(X(t), t), \\ &= \frac{\partial \xi}{\partial X(t)} dX(t) + \frac{\partial \xi}{\partial t} dt + \frac{1}{2} \frac{\partial^2 \xi}{\partial X^2(t)} (dX(t))^2 \\ &+ \frac{\partial^2 \xi}{\partial X(t) \partial t} dt + \frac{1}{2} \frac{\partial^2 \xi}{\partial t^2} (dt)^2 + \cdots \end{aligned} \tag{3.15}$$

Substituting (3.12) and applying (3.7), we get

$$\begin{aligned} d\xi(X(t), t) = &\left(\frac{\partial \xi}{\partial X(t)} f(X(t), t) + \frac{\partial \xi}{\partial t} + \frac{1}{2} \frac{\partial^2 \xi}{\partial X^2(t)} g^2(X(t), t) \right) dt \\ &+ \frac{\partial \xi}{\partial X(t)} g(X(t), t) dW(t) + \mathcal{O}(dt). \end{aligned} \tag{3.16}$$

□

Now based on the Ito's lemma, we can derive a general partial differential equation for any interest rate derivative depending on the short-rate r_t. An SDE with r_t as a stochastic random variable can be written as

$$dr_t = f(t, r_t) dt + g(t, r_t) dW(t), \tag{3.17}$$

Consider a risk-neutral portfolio[2] Π_t depends on a short-rate r_t, and consists of (i) a call option on the original interest rate product with maturity T_1 and price V_1, (ii) Δ units in another product with different maturity T_2 and price V_2, and (iii) the position of $V_1 - \Delta V_2$ in the risk-less asset. Now consider the value change of the portfolio over the infinitesimal time interval:

- the call option: the change in the product price is described by $dV_1(t) = V_1(t + dt) - V_1(t)$,

[2] In finance, a portfolio is a collection of investments held by an investment company, a hedge fund, a financial institution or an individual.

- the underlying units: the value change in the second product will be dV_2, so for Δ units, the change will be ΔdV_2,
- the risk-free asset: the interest rate is paid/received so that the change in this position will be $(V_1 - \Delta V_2) r_t dt$.

Therefore, the total change in the portfolio will be

$$d\Pi_t = \Delta dV_2 + (V_1 - \Delta V_2) r dt - dV_1. \tag{3.18}$$

According to Ito's lemma, we can define dV as

$$dV = \left(\frac{\partial V}{\partial r_t} f(r_t, t) + \frac{\partial V}{\partial t} + \frac{1}{2} \frac{\partial^2 V}{\partial r_t^2} g^2(r_t, t) \right) dt + \frac{\partial V}{\partial r_t} g(r_t, t) dW(t). \tag{3.19}$$

Substituting dV_1 and dV_2 in (3.18) we get

$$\begin{aligned} d\Pi_t = {} & (V_1 - \Delta V_2) r_t dt \\ & - \left[\left(\frac{\partial V_1}{\partial r_t} f(r_t, t) + \frac{\partial V_1}{\partial t} + \frac{1}{2} \frac{\partial^2 V_1}{\partial r_t^2} g^2(r_t, t) \right) dt + \frac{\partial V_1}{\partial r_t} g(r_t, t) dW(t) \right] \\ & + \Delta \left[\left(\frac{\partial V_2}{\partial r_t} f(r_t, t) + \frac{\partial V_2}{\partial t} + \frac{1}{2} \frac{\partial^2 V_2}{\partial r_t^2} g^2(r_t, t) \right) dt + \frac{\partial V_2}{\partial r_t} g(r_t, t) dW(t) \right]. \end{aligned} \tag{3.20}$$

Choosing $\Delta = \frac{\partial V_1}{\partial r_t} / \frac{\partial V_2}{\partial r_t}$ eliminates the stochastic term dW from the above equation. Also, to avoid arbitrage opportunities, the rate of return of this portfolio must be equal to the rate of return of the risk-free asset, r_t, and remained zero due to the zero net investment requirement, i.e., $d\Pi_t = 0$.

$$\begin{aligned} 0 = {} & \left[V_1 - \left(\frac{\partial V_1}{\partial r_t} \Big/ \frac{\partial V_2}{\partial r_t} \right) V_2 \right] r_t dt \\ & - \left[\left(\frac{\partial V_1}{\partial r_t} f(r_t, t) + \frac{\partial V_1}{\partial t} + \frac{1}{2} \frac{\partial^2 V_1}{\partial r_t^2} g^2(r_t, t) \right) dt + \frac{\partial V_1}{\partial r_t} g(r_t, t) dW(t) \right] \\ & + \left(\frac{\partial V_1}{\partial r_t} \Big/ \frac{\partial V_2}{\partial r_t} \right) \left[\left(\frac{\partial V_2}{\partial r_t} f(r_t, t) + \frac{\partial V_2}{\partial t} + \frac{1}{2} \frac{\partial^2 V_2}{\partial r_t^2} g^2(r_t, t) \right) dt + \frac{\partial V_2}{\partial r_t} g(r_t, t) dW(t) \right]. \end{aligned} \tag{3.21}$$

Eliminating the stochastic term

$$\begin{aligned} & \left[V_1 - \left(\frac{\partial V_1}{\partial r_t} \Big/ \frac{\partial V_2}{\partial r_t} \right) V_2 \right] r_t dt \\ & = \left[\frac{\partial V_1}{\partial t} + \frac{1}{2} \frac{\partial^2 V_1}{\partial r_t^2} g^2(r_t, t) - \left(\frac{\partial V_1}{\partial r_t} \Big/ \frac{\partial V_2}{\partial r_t} \right) \left(\frac{\partial V_2}{\partial t} + \frac{1}{2} \frac{\partial^2 V_2}{\partial r_t^2} g^2(r_t, t) \right) \right] dt. \end{aligned} \tag{3.22}$$

Rearranging the terms, and using notation $r = r_t$, we get

$$\frac{\frac{\partial V_1}{\partial t} + \frac{1}{2} \frac{\partial^2 V_1}{\partial r^2} g^2(r, t) - r V_1}{\frac{\partial V_1}{\partial r}} = \frac{\frac{\partial V_2}{\partial t} + \frac{1}{2} \frac{\partial^2 V_2}{\partial r^2} g^2(r, t) - r V_2}{\frac{\partial V_2}{\partial r}}. \tag{3.23}$$

Denoting either of the quotients as $u(r, t)$, we can write a PDE for V depending on r

$$\frac{\partial V}{\partial t} + \frac{1}{2} g^2(r, t) \frac{\partial^2 V}{\partial r^2} - u(r, t) \frac{\partial V}{\partial r} - rV = 0, \tag{3.24}$$

where functions g and u depend on market structures like yield curves.

Now, we will introduce some well-known one state variable short-rate models.

3.2.6.3. Vasicek and Cox-Ingersoll-Ross Models

The one state variable models describe the dynamics of the short-rate r as follows

$$dr_t = b(\theta - r_t)dt + \sigma r^{\beta} dW(t), \tag{3.25}$$

where b, θ, σ, and β are positive constants. The model proposed in [24] considers $\beta = 0$ and well known as the Vasicek model while the Cox-Ingersoll-Ross model introduced in [25] considers $\beta = 0.5$. One of the drawbacks of the Vasicek model is that the short-rate can be negative. On the other hand, in the case of the CIR model, the square root term does not allow negative interest rates. However, the major drawback of these models is that the model parameters are constant, so we can not fit the model to the market structures like yield curves.

3.2.6.4. Hull-White Model

The Hull-White model [26, 27] is an extension of the Vasicek model which can be fitted to the market structures like yield curves. The SDE is given as

$$dr_t = b(t)(\theta(t) - r_t)dt + \sigma(t)dW(t), \tag{3.26}$$

where model parameters $b(t)$, $\theta(t)$, and $\sigma(t)$ are time dependent. Equation (3.26) can also be represented as

$$dr_t = (a(t) - b(t)r_t)dt + \sigma(t)dW(t), \tag{3.27}$$

where $a(t) = b(t)\theta(t)$ is a deterministic function of time. The term $(a(t) - b(t)r_t)$ is a drift term and $a(t)$ is known as deterministic drift.

Definition 5. ***Drift.*** *The drift is a rate at which the expected value of the process changes.*

We can define a PDE for any underlying instrument based on the Hull-White model depending on r. We recall (3.24)

$$\frac{\partial V}{\partial t} + \frac{1}{2}g^2(r,t)\frac{\partial^2 V}{\partial r^2} - u(r,t)\frac{\partial V}{\partial r} - rV = 0.$$

In the case of a Hull-White model, $g(r,t) = \sigma(t)$, and $-u(r,t) = (a(t) - b(t)r)$. Substituting $g(r,t)$ and $u(r,t)$, we get

$$\frac{\partial V}{\partial t} + (a(t) - b(t)r)\frac{\partial V}{\partial r} + \frac{1}{2}\sigma^2(t)\frac{\partial^2 V}{\partial r^2} - rV = 0, \tag{3.28}$$

where the parameters $a(t)$, $b(t)$, and $\sigma(t)$ depend on the yield curve.

3.2.7. Yield Curve Simulation

The calibration of financial models is based on market structures like yield curves. These time-dependent parameters are derived from yield curves which determine the average direction in which short-rate r moves. The PRIIP's regulation demands to perform yield curve simulations for at least 10,000 times. We collect historical interest rates for these simulations. The data set must contain at least 2 years of daily interest rates for an underlying instrument or 4 years of weekly interest rates or 5 years of monthly interest rates. For example, we have collected the daily interest rate data at 20-30 tenor points in time over the past five years. Each year has approximately 260 working days also known as observation periods. Thus, there are $n \approx 1306$ observation periods and $m \approx 20$ tenor points in time. We then obtain the simulated yield curves by 'bootstrapping' process for the recommended holding period in the future.

Definition 6. ***Holding period.*** *A holding period is a period between the acquisition of an asset and its sale. It is the length of time during which an underlying instrument is 'held' by an investor.*

Remark 1. *The recommended holding period (RHP) gives an idea to an investor that for how long should an investor hold the product to minimize the risk. Generally, the RHP is given in years.*

The model parameters $a(t)$, $b(t)$, and $\sigma(t)$ are then calibrated based on these 10,000 simulated yield curves. In matrix form, we can write the simulated yield curves as

$$Y = \begin{bmatrix} y_{11} & \cdots & y_{1m} \\ \vdots & \vdots & \vdots \\ y_{s1} & \cdots & y_{sm} \end{bmatrix}_{s\times m} \tag{3.29}$$

where m is the number of tenor points and s is the total number of simulations, e.g., 10,000.
The authors of [28] observe that the small perturbations in the market data may lead to significant deviations in the parameter values of the short-rate models. Thus, this makes the parameter calibration problem as an ill-posed problem. We implement a classical Tikhonov regularization approach to stabilize the ill-posed problems. In this work, we use the inbuilt UnRisk PRICING ENGINE functions for the parameter calibrations [29]. UnRisk considers the model parameters as piecewise constant functions, i.e., a_i, b_i, σ_i are constants in $((i+1).\Delta T, i.\Delta T)$ where $0 < T < T^*$. Here, T^* is the last tenor point. In the case of the Hull-White model, we obtain $a(t)$ using simulated yield curves and $b(t), \sigma(t)$ from matrices of cap and swaption prices for various strikes, expiries, and maturities [29]. Based on 10,000 different simulated yield curves, we obtain s =10,000 different parameter vectors $a(t)$, $b(t)$, and $\sigma(t)$. In the matrix form, we write

$$\mathbf{a(t)} = \begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & \vdots & \vdots \\ a_{s1} & \cdots & a_{sm} \end{bmatrix} \mathbf{b(t)} = \begin{bmatrix} b_{11} & \cdots & b_{1m} \\ \vdots & \vdots & \vdots \\ b_{s1} & \cdots & b_{sm} \end{bmatrix} \boldsymbol{\sigma(t)} = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1m} \\ \vdots & \vdots & \vdots \\ \sigma_{s1} & \cdots & \sigma_{sm} \end{bmatrix} \tag{3.30}$$

All parameters are assumed to be piecewise constant changing their values only on tenor points, i.e., on model term dates. Thus, if there are m tenor points, then there will be m values for a single parameter vector. See [30] for the detailed parameter calibration procedures of the prominent financial models.

3.3. Numerical Methods

3.3.1. Finite Difference Method

The PDEs obtained for the short-rate models can be interpreted as convection-reaction-diffusion PDEs [31]. Consider a Hull-White PDE given by (3.28)

$$\frac{\partial V}{\partial t} + \underbrace{(a(t) - b(t)r_t)\frac{\partial V}{\partial r}}_{\text{Convection}} + \underbrace{\frac{1}{2}\sigma^2(t)\frac{\partial^2 V}{\partial r^2}}_{\text{diffusion}} - \underbrace{rV}_{\text{reaction}} = 0. \tag{3.31}$$

In this work, we apply the finite difference method to solve the Hull-White PDE. The convection term in the above equation may lead to numerically unstable results. Thus, we implement the so-called upwind scheme [32] to obtain a stable solution. We incorporate the semi-implicit scheme called the Crank-Nicolson method [33] for the time discretization.
The discretization of the PDE generates a parametric HDM of the following form (3.32). We need to solve this system at each time step n with an appropriate boundary condition and a known initial value of the underlying instrument.

$$A(\rho_s(t))V^{n+1} = B(\rho_s(t))V^n, \qquad V(0) = V_0, \tag{3.32}$$

where the matrices $A(\rho) \in \mathbb{R}^{M\times M}$, and $B(\rho) \in \mathbb{R}^{M\times M}$ are parameter dependent matrices. $V \in \mathbb{R}^M$ is a high dimensional state vector. M is the total number of spatial discretization points. t is the time variable. $t = [0, T]$ where T is the final term date. The time period $[0, T]$ is split into $n = 0, \ldots, N-1$ time intervals where N is the total number of time discretization points. Within the interval $[0, T]$, if we have m tenor points, then we will have m values for each parameter such that $a_s = a_{s1}, \ldots, a_{sm}$, $b_s = b_{s1}, \ldots, b_{sm}$, and $\sigma_s = \sigma_{s1}, \ldots, \sigma_{sm}$. For example, if we consider an instrument with a contract period of 10 years, then in this case $T = 10$ years. If there are m tenor points such that $m := 0, 1y, 2y, \cdots, 10y$, then we will get 11 different values for each parameter vector. That is, considering the time interval of 1 day, for the period of $n = 0$ to 260, $a_n = a_{s1}$, $b_n = b_{s1}$, and $\sigma_n = \sigma_{s1}$ (1 year $\approx$ 260 days).
For the simplicity of notations, we denote $\rho_s = [(a_{s1}, \ldots, a_{sm}), (b_{s1}, \ldots, b_{sm}), (\sigma_{s1}, \ldots, \sigma_{sm})]$ as the sth group of model parameters where $s = 1, \ldots, 10000$. Each parameter group ρ_s has values ranging from ρ_{s1} to ρ_{sm} where m is the total number of tenor points. We need to solve the system (3.32) for at least 10,000 parameter groups ρ generating a parameter space P of $10000 \times m$.

3.3.2. Parametric Model Order reduction

We employ the projection based MOR technique to solve the HDM (3.32). The idea is to project a high dimensional space onto a low dimensional subspace, Q as

$$\bar{V}^n = QV_d^n, \tag{3.33}$$

where $Q \in \mathbb{R}^{M\times d}$ is a reduced order basis with $d \ll M$, V_d is a vector of reduced coordinates, and $\bar{V} \in \mathbb{R}^M$ is the solution obtained using the reduced order model. Substituting (3.33) into the system of equations (3.32) gives the residual of the reduced state as

$$r^n(V_d, \rho_s) = A(\rho_s)QV_d^{n+1} - B(\rho_s)QV_d^n. \tag{3.34}$$

In the case of the Galerkin projection, the residual $r(V_d, \rho_s)$ is orthogonal to the reduced order basis Q

$$Q^T r^n(V_d^n, \rho_s) = 0. \tag{3.35}$$

Multiplying (3.34) by Q^T, we get

$$\begin{aligned} Q^T A(\rho_s)QV_d^{n+1} &= Q^T B(\rho_s)QV_d^n, \\ A_d(\rho_s)V_d^{n+1} &= B_d(\rho_s)V_d^n, \end{aligned} \tag{3.36}$$

where the matrices $A_d(\rho_s) \in \mathbb{R}^{d\times d}$ and $B(\rho_s) \in \mathbb{R}^{d\times d}$ are the parameter dependent reduced matrices. We obtain the parametric reduced order model (3.36) based on a proper orthogonal decomposition method (POD) [12, 13]. POD generates an optimal order orthonormal basis Q which serves as a low dimensional subspace in the least square sense for a given set of computational data. We aim to obtain the subspace Q independent of parameter space P. In this work, we obtain the optimal basis set by the method of snapshots. We compute snapshots by solving the HDM (3.32) for the selected parameter groups (i.e., snapshots taken for some parameter groups $\rho_1 \cdots \rho_l \in [\rho_1\ \rho_s]$). Further, we construct a snapshot matrix composed of these snapshots. Finally, we generate an optimally ordered orthonormal basis by performing a singular value decomposition of the snapshot matrix.

A solution V_s of the HDM for a single parameter group ρ_s can be represented as

$$V_s = \begin{bmatrix} V_s(r_1,t_1) & V_s(r_1,t_2) & \cdots & V_s(r_1,t_N) \\ V_s(r_2,t_1) & V_s(r_2,t_2) & \cdots & V_s(r_2,t_N) \\ \vdots & \vdots & \vdots & \vdots \\ V_s(r_M,t_1) & V_s(r_M,t_2) & \cdots & V_s(r_M,t_N) \end{bmatrix} \tag{3.37}$$

We obtain such solutions of the HDM for selected parameter groups and combined them to form a snapshot matrix, $\hat{V}$ such that

$$\hat{V} = \begin{bmatrix} V_1(r_1,t_1) & \cdots & V_1(r_1,t_N) & \cdots & V_l(r_1,t_1) & \cdots & V_l(r_1,t_N) \\ V_1(r_2,t_1) & \cdots & V_1(r_2,t_N) & \cdots & V_l(r_2,t_1) & \cdots & V_l(r_2,t_N) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ V_1(r_M,t_1) & \cdots & V_1(r_M,t_N) & \cdots & V_l(r_M,t_1) & \cdots & V_l(r_M,t_N) \end{bmatrix} \tag{3.38}$$

We perform a truncated SVD of the matrix V to obtain a reduced basis Q

$$\begin{aligned} \hat{V} &= \sum_{i=1}^{k} \Sigma_i \phi_i \psi_i^T, \\ \hat{V} &= \Phi \Sigma \Psi^T, \end{aligned} \tag{3.39}$$

where ϕ_i and ψ_i are the left and right singular vectors of the matrix $\hat{V}$ respectively. Σ_i are the singular values.

$$\hat{V} = \begin{bmatrix} \phi_1 & \cdots & \phi_k \end{bmatrix}_{M\times k} \begin{bmatrix} \Sigma_1 & 0 & \cdots \\ 0 & \ddots & \vdots \\ \vdots & \cdots & \Sigma_k \end{bmatrix}_{k\times k} \begin{bmatrix} \psi_1 & \cdots & \psi_k \end{bmatrix}_{k\times k} \tag{3.40}$$

The truncated (economy-size) SVD computes only first k columns of matrix Φ. The optimal projection subspace Q consists of d left singular vectors ϕ_i known as POD modes. Here, d makes the dimension of the reduced order model. The quality of the parametric reduced model mainly depends on the selection of parameters for which the snapshots are computed. Thus, it necessitates defining an efficient sampling technique for the high dimensional parameter space. We could consider standard sampling techniques like uniform or random sampling to generate snapshots. However, the computational cost for the uniform sampling may become too expensive due to the combinatorial explosion of samples used to cover the parameter domain [34]. On the other hand, random sampling might neglect the vital regions in the parameter space. The greedy sampling method introduced in [35] is proven to be an efficient method for sampling a high dimensional parameter space in the framework of MOR.

3.3.2.1. Adaptive Greedy Sampling Technique

The greedy sampling technique selects the parameter groups at which the error between the reduced order model and the full model is maximum. Further, we compute the snapshots using these parameter groups so that we can obtain the best suitable reduced order basis Q. However, the computation of a relative error $\|e(.,\rho)\| = \|V(.,\rho) - \bar{V}(.,\rho)\|/\|V(.,\rho)\|$ between the full model and the reduced model is expensive. Thus, usually, the error is replaced by the error bounds or the relative residual for the approximate solution $\bar{V}$. We discuss the error estimators ε in Sect. 3.3.2.2.

For a high dimensional parameter space, it is not feasible to compute the error estimate for each parameter group. Thus, we run the greedy sampling algorithm for a pre-defined set of parameter groups $\Pi \subseteq P$. The selection of this subset could be random. However, the random selection of a parameter set may not contain

the parameter vector corresponding to the most significant error. Therefore, instead of selecting Π randomly, we propose to select them adaptively. We construct a surrogate model $\bar{\varepsilon}$ to approximate the error estimator ε over the entire parameter space. Further, we use the surrogate model to locate the parameter groups Π, where the probability of having the larger values of ε is the highest. The adaptive greedy sampling methods in the framework of MOR are well addressed in [36, 37, 34]. This approach is also implemented to obtain the reduced order models for a parameterized steady Navier Stokes equation [38], for elliptical PDEs [39], and for parabolic PDEs [40].

3.3.2.2. Error Estimation

To avoid the computational expenses, the error $\|e(.,\rho)\|$ is usually replaced by the error estimators ε like error bounds and the norm of the residual.

- **Error Bonds:** Error bounds $\delta e(\rho)$ tell the maximum possible error in the approximation and can be given as

$$\|e(.,\rho)\| \leq \delta e(\rho) \quad \forall \rho \in P \tag{3.41}$$

  However, in some cases, it is not possible to define the error bounds or the error bounds do not exist. In such cases, the relative residual of the approximate solution is a common alternative.

- **Residual:** The relative error $\|e(.,\rho)\|$ is bounded by the relative residual as

$$\frac{\|r(V_d,\rho_s)\|}{\|A(\rho_s)\|} \leq \|e(.,\rho)\| \leq \|A^{-1}(\rho_s)\| \cdot \|r(V_d,\rho_s)\| \tag{3.42}$$

  This error bound holds if and only if $A(\rho)$ is a well-conditioned matrix [34, 41]. The quantity $\frac{\|r(V_d,\rho)\|}{\|A(\rho)\|\cdot\|V\|}$ is known as the relative residual.

3.4. Benchmark Cases

In the previous sections, we explain the mathematical framework for the analysis of financial instruments. To test the developed numerical methods, we present here some benchmark cases.

3.4.1. Benchmark Case 1

The first proposed benchmark case is to validate the numerical methods implemented for the simulation of yield curves and parameter calibration. The implemented numerical methods are following the guidelines provided by the PRIIP regulations.
We perform a principal component analysis on the collected historical data to ensure that the simulation results in a consistent curve. Further, using the principal components corresponding to their maximum energies, we calculate the consistent interest rates and composed them into a matrix called as the matrix of returns. Finally, we obtain the simulated yield curve by applying the bootstrapping procedure on the matrix of returns. To fulfill the regulations demand, we perform the bootstrapping process for at least 10,000 times. The detailed procedure can be found in the PRIIPs regulations. In this work, we implement the parameter calibration as described in [30]. We use the inbuilt UnRisk functions for the parameter calibration. UnRisk PRICING ENGINE integrates the pricing and calibration engines into Mathematica. Highly advanced numerical schemes are used to provide fast and robust solutions which are all programmed in C++.
The numerical methods are tested with real market based historical data. We have collected the daily interest rate data at 26 tenor points in time over the past five years. Each year has 260 working days. Thus, there are 1300 observation periods. We have retrieved this data from the State-of-the-art stock exchange information system "Thomson Reuters EIKON" [42].

3.4.2. Benchmark Case 2

The main task of this research is to evaluate the financial instrument based on short-rate models. Second benchmark case is to verify the implemented MOR algorithm. We use a finite difference method for simulating the convection-diffusion-reaction PDE. The projection-based MOR approach has been implemented, and the reduced-order basis is obtained using the proper orthogonal decomposition approach.
We aim to validate these numerical methods and designed algorithms using two different 'real-world' financial instruments. The first numerical example is of the floater with a cap and a floor. The details are as given in Table 3.1. Another numerical example is of a puttable bond (see Table 3.2).

Table 3.1: Numerical Example of a floater with a cap and a floor.

Reference interest rate	Euribor3M
Fixing of Euribor3M	in advance
Coupon frequency	quarterly
Cap rate	2.25 % p.a.
Floor rate	0.5 % p.a.
Maturity and Nominal value	10 years with face value of €1

Table 3.2: Numerical Example of a puttable bond.

Interest rate	5 % p.a. fixed
Coupon frequency	Annually
Put rate	€1
Face/nominal value	€1
Maturity	10 years
Put dates	Annually starting one year after issuing a bond

We justify the numerical results obtained from the reduced-order model by comparing to the results obtained using the HDM. We calculate the relative error between the ROM and HDM as follows

$$\|e(.,\rho)\|_2 = \frac{\|V(.,\rho) - \bar{V}(.,\rho)\|_2}{\|V(.,\rho)\|_2}, \tag{3.43}$$

where $\|\cdot\|_2$ is the 2-norm.

Bibliography

[1] European Commission, "Commission delegated regulation (EU) 2017/653 OJ L 100," *Off. J. EU*, vol. 1, pp. 1–52, 2017.

[2] A. Gupta and M. Subrahmanyam, "Pricing and hedging interest rate options: Evidence from cap-floor markets," *J. Bank. Finance*, vol. 29, pp. 701–733, 2005.

[3] J. Bicksler and A. Chen, "An economic analysis of interest rate swaps," *J. Finance*, vol. 3, pp. 645–655, 1986.

[4] D. Brigo and F. Mercurio, *Interest Rate Models - Theory and Practice*, 2nd ed. Berlin: Springer-Verlag, 2006.

[5] M. Aichinger and A. Binder, *A Workout in Computational Finance*, 1st ed. West Sussex, UK: John Wiley and Sons Inc., 2013.

[6] E. Ekström, P. Lötstedt, and J. Tysk, "Boundary values and finite difference methods for the single factor term structure equation," *J. Appl. Math. Finance*, vol. 16, pp. 253–259, 2009.

[7] T. Haentjens and K. I. Hout, "Alternating direction implicit finite difference schemes for the heston-hull-white partial differential equation," *J. Comput. Finance*, vol. 16, pp. 83–110, 2012.

[8] A. Falcó, L. Navarro, and C. Cendón, "Finite difference methods for hull-white pricing of interest rate derivatives with dynamical consistent curves," *SSRN Elec. J.*, 2014.

[9] A. Sepp, "Numerical implementation of hull-white interest rate model: Hull-white tree vs finite differences," Working Paper, 2002, available from `www.hot.ee/seppar`.

[10] A. Cohen and R. DeVore, "Approximation of high-dimensional parametric pdes," *Acta Numerica*, vol. 24, pp. 1–159, 2015.

[11] I. Piotr and R. Motwani, "Approximate nearest neighbors: Towards removing the curse of dimensionality," in *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing*. New York, NY, USA: ACM Press, 1998, pp. 604–613.

[12] A. Chatterjee, "An introduction to the proper orthogonal decomposition," *Curr. Sci.*, vol. 78, pp. 808–817, 2000.

[13] G. Berkooz, P. Holmes, and J. Lumley, "The proper orthogonal decomposition in the analysis of turbulent flows," *Annu. Rev. Fluid Mech.*, vol. 25, no. 1, pp. 539–575, 1993.

[14] M. Graham and I. Keverekidis, "Alternative approaches to the karhunen-loéve decomposition for model reduction and data analysis," *Comput. Chem. Eng.*, vol. 20, no. 5, pp. 495–506, 1996.

[15] I. Jolliffe, *Principal Component Analysis*, 1st ed. Berlin: Springer-Verlag, 2014.

[16] M. Graham and I. Keverekidis, "Proper orthogonal decomposition and its applications part I: Theory," *J. Sound Vib.*, vol. 252, no. 3, pp. 527–544, 2002.

[17] L. Sirovich, "Turbulence and the dynamics of coherent structures. Part I: coherent structures," *Quart. Appl. Math.*, vol. 45, no. 3, pp. 561–571, 1987.

[18] R. Dudley, *Unifrom Central Limit Theorems*, 1st ed. New York, NY, US: Cambridge University Press, 2010.

[19] A. Corhay and A. Rad, "Statistical properties of daily returns: Evidence from european stock markets," *J. Bus. Finan. Account.*, vol. 21, no. 2, pp. 271–282, 1994.

[20] H. Albrecher, A. Binder, V. Lautscham, and P. Mayer, *Introduction to Quantitative Methods for Financial Markets*, 1st ed. Berlin: Springer-Verlag, 2010.

[21] M. Grigoriu, *Stochastic Calculus: Applications in Science and Engineering*, 1st ed. Basel: Birkhäuser, 2002.

[22] R. Mahnke, J. Kaupuzs, and I. Lubashevsky, *Physics of Stochastic Processes: How Randomness Acts in Time*, 1st ed. Weinheim: WILEY-VCH Verlag GmbH and Co., 2008.

[23] P. Wilmott and S. Howson, *The Mathematics of Financial Derivatives: A Student Introduction*, 1st ed. London: Cambridge University Press, 2002.

[24] O. Vasicek, "An equilibrium characterization of the term structure," *J. Financial Econ.*, vol. 5, no. 2, pp. 177–188, 1977.

[25] J. Cox, J. Ingersoll Jr., and S. Ross, "A theory of the term structure of interest rates," *Econometrica*, vol. 53, no. 2, pp. 385–407, 1985.

[26] J. Hull and A. White, "Pricing interest-rate-derivative securities," *Rev. Financial Stud.*, vol. 3, no. 4, pp. 573–592, 1990.

[27] ——, "One-factor interest-rate models and the valuation of interest-rate derivative securities," *J. Financ. Quant. Anal.*, vol. 28, no. 2, pp. 235–254, 1993.

[28] H. Engl, "Calibration problems–an inverse problems view," *Wilmott*, pp. 16–20, 2007.

[29] MathConsult, "Calibration of interest rate models," MathConsult GmbH, Linz, Austria, Report, 2009.

[30] S. Shreve, *Stochastic Calculus and Finance*, 1st ed. New York, US: Springer-Verlag, 2004.

[31] D. Anderson, J. Tannehill, and R. Pletcher, *Computational Fluid Mechanics and Heat Transfer*, 3rd ed. London: CRC Press, 2013.

[32] J. Heinrich, P. Huyakorn, O. Zienkiewicz, and A. Mitchell, "An 'upwind' finite element scheme for two-dimensional convective transport equation," *Int. J. Numer. Meth. Engng.*, vol. 11, pp. 131–143, 1977.

[33] G. Sun and C. Trueman, "Efficient implementations of the crank-nicolson scheme for the finite-difference time-domain method," *IEEE Trans. Microw. Theory Tech.*, vol. 11, pp. 131–143, 1977.

[34] T. Bui-Thanh, K. Willcox, and O. Ghattas, "Model reduction for large-scale systems with high-dimensional parametric input space," *SIAM J. Sci. Comput.*, vol. 30, no. 6, pp. 3270–3288, 2008.

[35] C. Prud'homme, D. Rovas, K. Veroy, L. Machiels, Y. Maday, A. Patera, and G. Turinici, "Reliable real-time solution of parametrized partial differential equations: Reduced-basis output bound methods," *J. Fluids Eng.*, vol. 124, no. 1, pp. 70–80, 2001.

[36] A. Paul-Dubois-Taine and D. Amsallem, "An adaptive and efficient greedy procedure for the optimal training of parametric reduced-order models," *Int. J. Numer. Meth. Engng.*, vol. 102, pp. 1262–1292, 2014.

[37] B. Notghi, M. Ahmadpoor, and J. Brigham, "Adaptive reduced-basis generation for reduced-order modeling for the solution of stochastic nondestructive evaluation problems," *Comput. Methods. Appl. Mech. Engrg.*, vol. 310, pp. 172–188, 2016.

[38] K. Veroy and A. Patera, "Certified real-time solution of the parametrized steady incompressible Navier–Stokes equations: rigorous reduced-basis a posteriori error bounds," *Int. J. Numer. Meth. Fluids*, vol. 47, pp. 773–788, 2005.

[39] K. Veroy, C. Prud'homme, and A. Patera, "A posteriori error bounds for reduced-basis approximation of parametrized noncoercive and nonlinear elliptic partial differential equations," in *Proceedings of the 16th AIAA Computational Fluid Dynamics Conference*, Orlando, United States, 2003, p. 3847.

[40] M. Grepl and A. Patera, "A posteriori error bounds for reduced-basis approximations of parametrized parabolic partial differential equations," *M2AN Math. Model. Numer. Anal.*, vol. 39, pp. 157–181, 2005.

[41] W. Oettli and W. Prager, "Compatibility of approximate solution of linear equations with given error bounds for coefficients and right-hand sides," *Numer. Math.*, vol. 6, no. 1, pp. 405–409, 1964.

[42] Thomson Reuters Eikon, "Interest rate data," December 2018, retrieved from `eikon.thomsonreuters.com`.

4. Benchmarks for the modelling of continuous casting molds

Umberto Emil Morelli[1], Peregrina Quintela[2], Patricia Barral[2], Gianluigi Rozza[3], Michele Girfoglio[3], Gianfranco Marconi[4], Conte Riccardo[4]

[1] *Instituto Tecnolóxico de Matemática Industrial*

[2] *Instituto Tecnolóxico de Matemática Industrial, Universidade de Santiago de Compostela*

[3] *Scuola Internazionale Superiore di Studi Avanzati*

[4] *Danieli*

Abstract. The objective of the present research project is the development of a control algorithm for a continuous casting mold equipped with thermocouples. The algorithm will have to solve in real time the inverse and direct heat transfer problem in a water cooled copper mold providing the instantaneous heat flux at the mold inner boundary and the temperature gradients within the mold. This document provides a description of the process and a hierarchy of the mathematical models used for the simulation of continuous casting molds. Moreover, different benchmarks are presented for the validation and analysis of the algorithms which will be developed in the present research.

Keywords: Continuous casting, mold, heat transfer, reduce order modeling, inverse heat transfer problem.

4.1. Introduction and Motivation

The continuous casting (CC) of steel is presently the most used process to produce steel worldwide. In 2017, 96% of the steel was produced using CC [1]. This process has been used for decades and during this time it has undergone improvements based mainly on experience with the commercial operation, aided by physical water modeling to understand the fluid flow behavior and more recently by numerical simulations [2]. In this section an overview of the CC process, illustrated in Figure 4.1, is provided. Finally at the end of the section, the motivation and objectives of the present research are presented.

In continuous casting the metal is first heated until it liquefies. The molten metal is then tapped into the ladle. When it is at the correct temperature, the metal goes into the tundish. The tundish plays the role of a reservoir while one ladle is empty and it is substituted by another. In the tundish the metal flow is regulated and smoothed.

Through the Submerged Entry Nozzle (SEN), the metal is drained into an open-base copper mold. The mold is water cooled to start the solidification of the metal (primary cooling). The SEN drains the metal below a layer of mold powder which is floating at the top of the molten metal in the mold as in Figure 4.2. Part of the powder in contact with the steel melts down creating a liquid layer. This layer fills the gap between the steel and the mold. Due to the heat extraction at the mold face, a portion of the liquid resolidifies in contact with the mold creating a solid layer. Finally, due to the strand shrinkage, an air gap can occur in the lower part of the mold expecially at the strand corners. This air gap reduces the heat extraction from the steel to the mold. Consequently, the molds are generally tapered to reduce this air gap.

The steel solidification starts at the mold and continues all the way down the secondary cooling section. In this region, the strand is supported by a series of cooled rollers and sprayed with water. The complete solidification of the strand is achieved at the end of this section.

The mold design and working conditions directly influence the final product quality and the production capacity of a continuous caster as well as the mold life time. The present investigation arises from the need of the industrial partner of the project, Danieli & C. Officine Meccaniche S.p.A. [4], to have a robust and versatile control system for their continuous casting molds. To analyze and control the behavior of a casting mold, it is essential to know the instantaneous temperature field and heat fluxes in it. The molds are equipped with thermocouples installed within the copper plates of the mold. These thermocouples provide temperature measurements at dif-

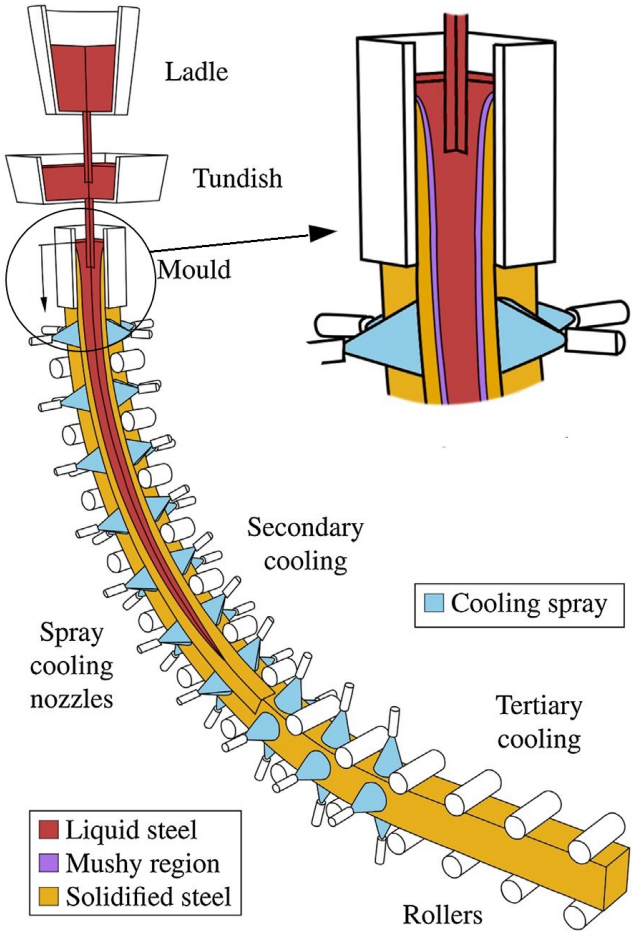


Figure 4.1: Schematic overview of the continuous casting process [3]. The molten steel is tapped from the ladle to the tundish. Then, the SEN injects it into the mold where the steel begins its solidification. Starting from the mold, the steel slab is cooled and it reaches complete solidification in the secondary cooling region.

ferent locations of the mold. To control the mold, the instantaneous heat fluxes at the mold face in contact with the strand have to be computed based on the thermocouples measurements (inverse problem).

Then, knowing the heat fluxes at the boundary, the temperature gradients in the mold can be computed (direct problem) allowing the control of the machine and the adjustment of the casting parameters to optimize the process. The main goal of the present investigation is the development of a control algorithm for the CC mold which can achieve the aforementioned computations in real time within the machine control system. To achieve this goal, Reduced Order Modeling (ROM) techniques will play a fundamental role in the investigation. Consequently, benchmarks are required to validate, study and compare the obtained results, and to compare the performance of different algorithms.

4.2. State of the art

An extensive literature on the modeling of continuous casting direct simulation and the aforementioned inverse problem is available. In this section a brief overview is provided and discussed.

4.2.1. Direct problem

The direct simulation of continuous casting is a very complex problem. Many interacting phenomena occur in the mold region including heat transfer, solidification, multiphase turbulent flow, clogging, complex interfacial

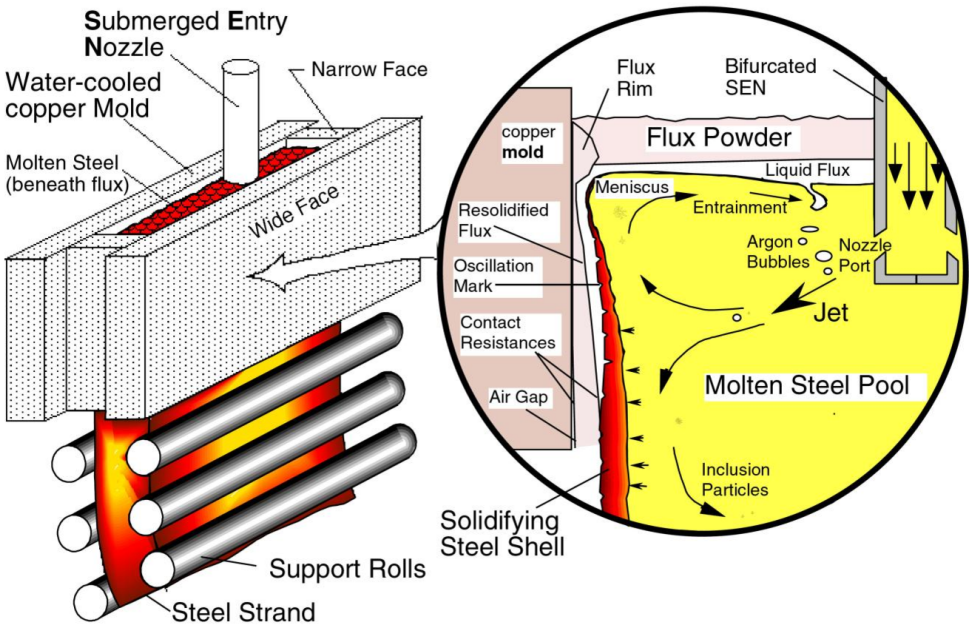


Figure 4.2: Schematic of a vertical section of the mold region [5]. On the left, the main phenomena occurring within the mold are illustrated. In particular, the figure shows the gap which forms betwee the steel and the mold. This gap is filled by the flux powder which liquefies in contact with the hot steel and resolidifies close to the mold face. In the lower part of the mold an air gap can occur due to the slab shrinkage. Molds are generally tapered to reduce this air gap.

behavior, thermo-mechanical distortion, stresses, cracks, segregation and microstructure formation. These phenomena are generally transient, three-dimensional, and operate over wide length and time scales. Consequently, several models have been developed ranging from fast and simple models for implementation in online control systems to sophisticated multiphysics simulations that take into account many coupled phenomena (an updated review on this subject by Thomas [2] is available).

The complexity of the model depends on the physical phenomena considered and the domain used. In this section an overview of the mold models available in the literature is presented. Because the present investigation regards the mold and its cooling system, models related to the steel flow in the slab and its solidification are not discussed here.

A models hierarchy is illustrated in Figure 4.3. In this section, the mathematical models found in the literature are described from the simplest to the more complex.

Considering only the mold copper region, the transient three-dimensional heat equation is [6]

$$\rho_c C_{p_c} \frac{\partial T_c}{\partial t} = \nabla \cdot (k_c \nabla T_c), \tag{4.1}$$

where $T_c = T_c(\mathbf{x}, t)$ is the copper temperature, ρ_c its density, $k_c(T_c)$ the thermal conductivity, and C_{p_c} the specific heat capacity. Assuming a steady-state process and constant properties [7, 8], the heat equation can be simplified as

$$\Delta T_c = 0. \tag{4.2}$$

In any case boundary conditions (BC) are provided. At the portion of the boundary in contact with the slab or with the air if the slab shrinkage creates a gap (so called "hot" face) Γ_{hot}, the empirical heat flux proposed by

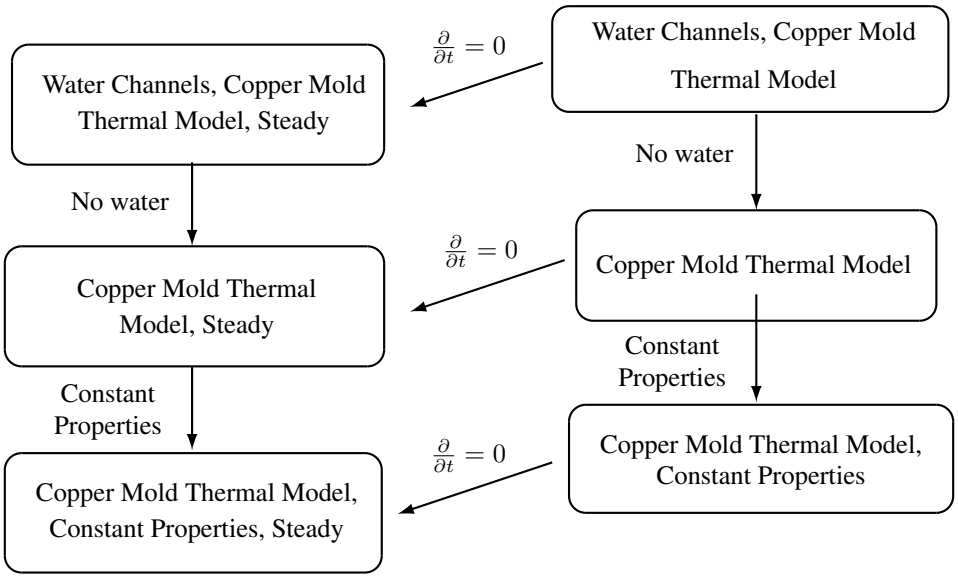


Figure 4.3: Models hierarchy for the direct simulation of continuous casting molds.

Savage and Pritchard [9] is commonly imposed

$$-k_c \frac{\partial T_c}{\partial n}\bigg|_{\Gamma_{hot}} = q(v_c, z) = A_0 + B_0\sqrt{t_{dwell}(v_c, z)}, \tag{4.3}$$

where A_0 and B_0 are coefficients dependent on the casted steel, and $t_{dwell} = z/v_c$ is the slab dwell time which is a function of the cast velocity v_c and of the axial coordinate along the mold z.

Alternatively, the heat flux at this boundary can be computed solving the inverse problem (discussed in the next section). The computed heat flux can then be applied as a boundary condition in the direct problem [10].

At the portion of the boundary in contact with the cooling water Γ_w, a convection boundary condition is applied

$$-k_c \frac{\partial T_c}{\partial n}\bigg|_{\Gamma_w} = h_w(T_c - T_w), \tag{4.4}$$

where T_w is the local water temperature. In the literature when this model has been used, T_w is assumed to increase linearly along the water channel from the inlet temperature [11].

The water channel heat transfer coefficient h_w can be calculated using the Sleicher Rouse equation [12]

$$h_w = (5 + 0.15Re^{C1}Pr^{C2})\frac{K}{D} \tag{4.5}$$

where

$$\begin{aligned} C1 &= 0.88 - \frac{0.24}{4 + Pr}, \\ C2 &= 0.333 + 0.5e^{-0.6Pr}, \\ K &= 0.59 + 0.001(T_w[K] - 273), \\ D &= 4A/P \end{aligned} \tag{4.6}$$

where Re and Pr are the Reynolds and Prandtl numbers, A the cross sectional area of the water channel and D the perimeter.

Commonly, the heat flux at the portions of the boundary in contact with "cold" air (i.e. upper Γ_{up} and lower

Γ_{low} faces of the mold, and the outer part of the mold not in contact with cooling water Γ_{out}) is neglected [13, 7, 14, 11]

$$-k_c \frac{\partial T_c}{\partial n}\bigg|_{\Gamma_{up}} = -k_c \frac{\partial T_c}{\partial n}\bigg|_{\Gamma_{low}} = -k_c \frac{\partial T_c}{\partial n}\bigg|_{\Gamma_{out}} = 0. \tag{4.7}$$

Consequently, in these models the heat flows into the mold from its hot face and flows out from the water channels.

The complexity of the model can be increased including in the domain the water flowing in the cooling system [15]. This improvement of the model is especially required when the cooling system geometry is complex and it is not possible to assume a temperature profile.

Modeling water as a Newtonian fluid, its flow can be modeled using RANS and a turbulence model (i.e. $k-\epsilon$). Assuming steady flow, the water model includes continuity equation

$$\nabla(\rho_w(T_w)\mathbf{u}) = 0, \tag{4.8}$$

momentum conservation equation

$$\nabla \cdot (\rho_w(T_w)\mathbf{u} \otimes \mathbf{u}) = -\nabla p_w + \nabla \cdot [\mu_{eff}(T_w)(\nabla \mathbf{u} + \nabla \mathbf{u}^T)] + \rho_w(T_w)\mathbf{g}, \tag{4.9}$$

and the energy equation

$$\frac{k_w}{\rho_w C_{p_w}} \Delta T_w = \mathbf{u} \nabla T_w. \tag{4.10}$$

In these equations, T_w is the water temperature, $\rho_w(T_w)$ the density, $\mathbf{u}$ the Reynolds averaged water velocity vector, and $\mu_{eff} = \mu + \mu_t$ the effective viscosity composed of the dynamic viscosity $\mu(T_w)$ and the turbulence viscosity μ_t. In a general case, the water properties are calculated as a function of the water temperature.

The heat transfer between water and copper is again expressed by Eq. 4.4. With T_w being the computed water temperature at the boundary while h_w is computed based on local flow-field conditions.

4.2.2. Inverse problem

The mold is equipped with a number of thermocouples that measure the temperature at different locations in the copper mold. Moreover, the cooling water flow is measured as well as the water averaged temperature at the inlet and outlet of the cooling system. As previously mentioned, an inverse problem has to be solved for the computation of instantaneous heat fluxes at the mold hot face.

This is a well known problem in the literature and several inverse problem solution techniques have been used. Such as conjugate gradient method, genetic algorithm, Nelder-Mear algorithm, least square method. Even the use of neural networks has been explored [16].

In the literature, the validation of the inverse problem model has been conducted in two ways: (i) using real mold thermocouples measurements and comparing them to the temperatures computed at the thermocouples positions using the BC of the inverse model [17, 16, 18] or (ii) creating a numerical test case of which the BC are known and validating the inverse problem model by its ability to reconstruct this boundary condition [19, 20].

Being the objective of the present investigation the development of a real time control algorithm, different inverse problem solution techniques will be studied to in order to select the most computationally efficient. More importantly, the implementation of ROM in the inverse problem will be investigated as a possible solution to acchieve real time solution performances.

4.3. Benchmarks

In the literature it is possible to identify notable test cases which geometry and thermal state is publicly available. These test cases have been identified and are proposed here as benchmarks for the validation and analysis

of the algorithms which will be developed in the present research. Moreover, the design of a new valuable test case is proposed in partnership with the industrial partner.

The selected benchmarks can be divided into analytical solutions, experiments and numerical simulations. These test cases have been used in the literature in different ways but mainly for validation of direct simulations, validation of inverse problem solvers, research on mold design and/or optimization of the measurement system [20] .

Table 4.1: Benchmarks selected for the present investigation.

Type	1D	2D	3D
Analytical solutions	Benchmark 1 [21] Benchmark 2 [22]		
Experiments			Benchmark 6 [15] Benchmark 3 [13] Benchmark 5 [23] Benchmark 8
Numerical simulations		Benchmark 3 [20] Benchmark 4 [19] Benchmark 5 [10]	Benchmark 6 [15] Benchmark 7 [15] Benchmark 8

4.3.1. Benchmark 1

Analytical solutions are available for an infinite plate in different conditions. Two cases will be considered [21]:(i) a infinite uniform plate with applied heat flux on one side and heat convection on the other in steady condition, and (ii) a infinite uniform plate with uniform temperature at time t_0 and convection on both sides.

Assuming constant properties, the solution of the one-dimensional problem (i), shown in Figure 4.4 is:

$$T(x) = q_{in}\left(\frac{L-x}{k} + \frac{1}{h}\right) + T_\infty \tag{4.11}$$

where q_{in} is the heat flux on the left side of the plate, L the plate thickness, k the thermal conductivity on the plate, h the convection heat transfer coefficient on the right side, and T_∞ the surrounding temperature.

Assuming constant properties, the solution of the one-dimensional problem (ii) for the dimensionless time $\tau = \frac{\alpha t}{k} > 0.2$ can be approximated as [21]:

$$\theta(x,t) = \frac{T(x,t) - T_\infty}{T_i - T_\infty} = A_1 e^{-\lambda_1^2 \tau} cos(\lambda_1 x/L), \quad \tau > 0.2 \tag{4.12}$$

where A_1 and λ_1 are dependent on the Biot number $Bi = \frac{hL}{k}$ only and T_i is the temperature at t_0.

4.3.2. Benchmark 2

The classical steady state conjugated problem in Figure 4.5 (Luikov problem) has been selected as a benchmark because it involves convection on one face of the plate as in a simplified mold model.

For thin plates, the solution of the problem is [22]

$$\theta_T(\tau) = \frac{T_b - T_(x)}{T_b - T_\infty} = \sum_{i=1}^{\infty} c_i Br_x^i \tag{4.13}$$

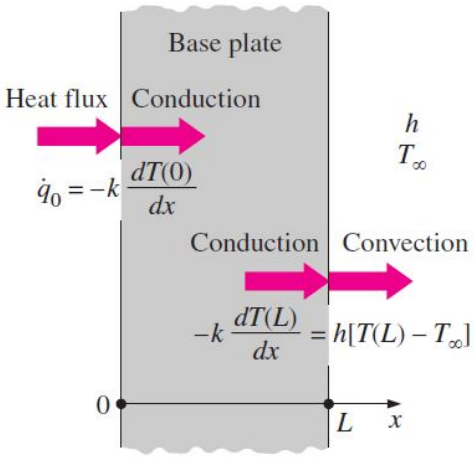


Figure 4.4: Schematic of the infinite uniform plate in steady condition with applied heat flux on one side and heat convection on the other one (Benchmark 1.i) [21].

where $Br_x = \frac{k_f}{k_s}\frac{b}{x}Re_x^m Pr^n$ is the Brun number and the constants c_i are obtained from

$$\begin{aligned} c_1 &= \frac{-1}{a_0 + a_1}, \\ c_i + 1 &= \left(\sum_{j=1}^{i+1} \frac{\Gamma(i+2)}{\Gamma(i+2-j)} a_j\right)^{-1} c_i, \quad i = 1, 2, 3, \cdots \end{aligned} \tag{4.14}$$

where Γ is the Gamma function and

$$a_j = -\frac{\Gamma(\beta)\Gamma(\frac{1}{\gamma})}{C\gamma} \sum_{i=j}^{\infty} \sum_{r=0}^{j} \binom{-\beta}{i} \binom{i}{r} \frac{(-1)^r \prod_{p=-i+1}^{0} \left(\frac{j-r}{\gamma}(m-1) + p\right)}{\Gamma(\frac{1}{\gamma} + \beta + i)\Gamma(j+1)} \tag{4.15}$$

where C, γ, m, and β are contants dependent on the condition of the flow (turbulent or laminar).

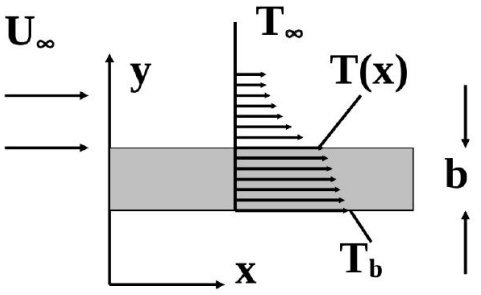


Figure 4.5: Plate with uniform temperature T_b at lower surface and convectively cooled at upper surface (Luikov problem). [22].

4.3.3. Benchmark 3

Udayraj et a. [20] developed a benchmark case based on the experimental results by Samarasekera and Brimacombe [13]. The mold considered has a total height $H + H_1 = 700\ mm$ and a width $L = 96\ mm$. The domain

considered is a rectangular two dimensional vertical slice of the mold. The mold is made up of copper with thermal conductivity $k_c = 390\ W/mK$. The meniscus is located at $y = H = 617.6\ mm$ and the water enters in the channel from the bottom ($y = 0$) at a speed $U_w = 7\ m/s$ and with temperature $T_{w_{in}} = 303.15\ k$. Water properties are summarized in Table 4.2. The casting speed is $v_c = 0.0375\ m/s$.

Table 4.2: Water properties of Benchmark 3 [20].

Density, ρ_w	$999.97\ kg/m^3$
Specific heat, c_{p_w}	$4180\ J/kgK$
Dynamic viscosity, μ_w	$79.77 \cdot 10^{-5}\ Ns/m^2$
Thermal conductivity, k_w	$0.6\ W/mK$
Heat transfer coefficient, h_w	$22500\ W/m^2K$

The Savage and Pritchard [9] heat flux is applied at the hot face

$$q_s(y) = 2680 - 335\sqrt{\frac{H-y}{v_c}} \quad \left[\frac{kW}{m^2}\right]. \tag{4.16}$$

For this test case, measured and simulated temperature distributions on hot and cold face are publicly available.

4.3.4. Benchmark 4

Zhang et al. [19], developed the 2D test problem shown in Figure 4.6. The test problem is a rectangular area ABEF with $H = 21\ mm$ and $W = 12\ mm$, starting at $288\ K$. The AF and BE boundaries are insulated. The EF boundary is cooled off by water with convection heat transfer coefficient $h_w = 1 \cdot 10^4\ W/m^2K$ and temperature $T_w = 288\ K$. The heat flux $q(\partial\Omega_2, t)$ is applied to boundary AB, where $\partial\Omega_2 = \{(x, y) : x = 0\ mm \leq y \leq 21\ mm\}$ and $0\ s \leq t \leq 25\ s$. The heat flux $q(\partial\Omega_2, t)$ profile is decreasing linearly in the y-axis and varies triangularly in time,

$$q(\partial\Omega_2, t) = \begin{cases} 1 \cdot 10^6, & 0\ s \leq t \leq 5\ s \\ 1 \cdot 10^6 \left(1 - \frac{y[mm]}{21\ mm}\right) \frac{t[s]}{5\ s} + 1 \cdot 10^6, & 5\ s \leq t \leq 10\ s \\ 1 \cdot 10^6 \left(1 - \frac{y[mm]}{21\ mm}\right) \frac{20\ s - t[s]}{5\ s} + 1 \cdot 10^6, & 10\ s \leq t \leq 15\ s \\ 1 \cdot 10^6. & 15\ s \leq t \leq 20\ s \end{cases} \tag{4.17}$$

In the test problem calculation, eight virtual response thermocouples of 1st column are spaced 3 mm apart in y-axis and located 3 mm away from the AB boundary to provide the temperature for minimizing the objective function of inverse calculation. The other eight virtual thermocouples of the 2nd column spaced 3 mm apart in y-axis and located 8 mm away from the AB boundary to provide the boundary condition as show in Figure 4.6(b). The direct problem on this test domain is then solved with a 0.2 mm x 0.2 mm grid size and 0.01 s time-step. The temperature is stored every 0.1 s. Subsequently, the measured simulated temperatures are given as input of the inverse model for the estimation of the heat flux on $\partial\Omega_2$.

Using this benchmark case, Zhang et al. tested the robustness of the inverse problem model. A Gaussian noise signal $\omega\sigma$ was added to the simulated temperature to mimic the measurement error, where the standard deviation of noise σ is set as 0, 0.1 and 0.2, and $-2.576 \leq \omega \leq 2.576$ is a random variable.

4.3.5. Benchmark 5

A three-dimensional test case have been developed by Du et al. [10, 23] to validate their 2D model. They experimentally measured the temperature at different locations of the curved caster of Figure 4.7. The caster has a curvature radius of $10.75\ m$ and metallurgical length of $28.8\ m$ with a casting speed $v_c = 0.65\ m/min$.

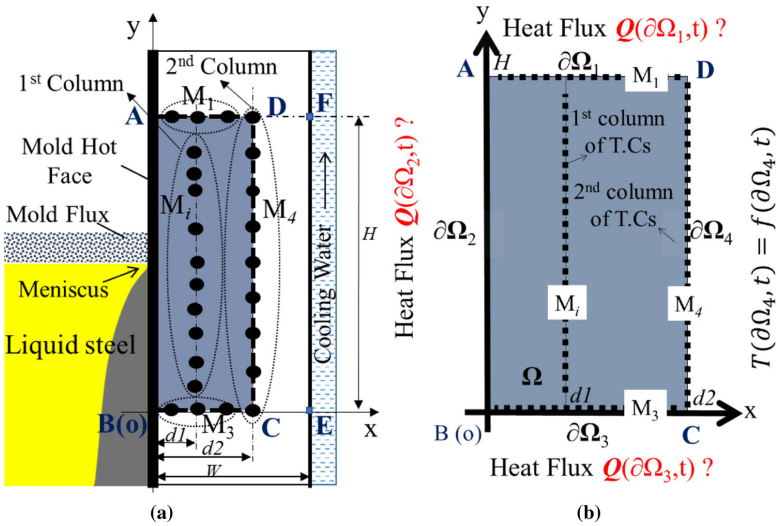


Figure 4.6: Scheme of the benchmark developed by Zhang et al. [19]: (a) the locations of thermocouples in the copper mold, where the dots represent thermocouples, and (b) the inverse problem computational domain and the boundary conditions.

The copper mold is coated with 1 mm of nickel on the hot face. The thermophysical properties of the materials are summarized in Table 4.3. Other geometry and process parameters are summarized in Table 4.4. A detail of the water slots at the outside radius is illustrated in Figure 4.8.

Table 4.3: Thermophysical properties of materials used in Benchmark 5 [20].

Material	Specific heat [J/kgK]	Thermal conductivity [W/mK]	Density [kg/m^3]
Copper	390	340	8900
Nickel	460	80	8910
Water	4200	0.5	998

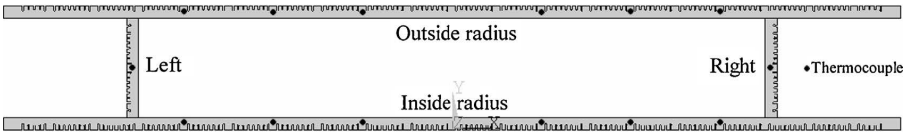


Figure 4.7: Horizontal section of the mold used in Benchmark 5 [23].

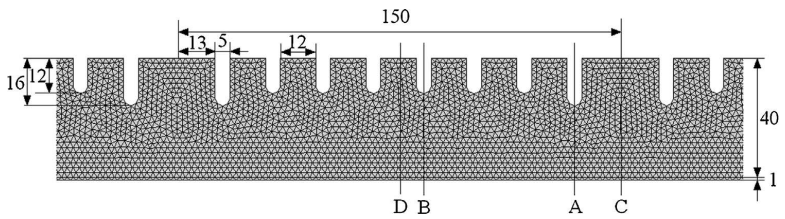


Figure 4.8: Detail of the water slots at the outside radius of the mold of Benchmark 5 [23].

In Figure 4.9 the measured temperatures are compared with the temperature profile along the width direction at the buried sections of the thermocouples. As shown in the figure, three rows of thermocouples were buried in the mold 210 mm, 325 mm and 445 mm from the mold top.

Table 4.4: Geometry and process parameters of Benchmark 5 [20].

Item	Value
Mold height	900 mm
Copper plate thickness	40 mm
Slab width	2100 mm
Slab thickness	320 mm
Water slot width	5 mm
Water slots depth	
Inside radius	16 mm, 20 mm
Outside radius	12 mm, 16 mm
Narrow face	12 mm, 16 mm
Casting speed	0.65 m/min
Meniscus level	100 mm
Water temperature	
Inlet	304.8 K
Outlet for inside radius	308.9 K
Outlet for outside radius	310.5 K
Outlet for narrow faces	310.4 K
Water flow	
Inside radius	0.097 m^3/s
Outside radius	0.074 m^3/s
Narrow faces	0.0108 m^3/s
Thermocouples distance from hot face	
Inside radius	14 mm
Outside radius	28 mm

4.3.6. Benchmark 6

Meng and Thomas [14] developed a benchmark case for the validation of their COND1D model which was used later on by other investigations in the literature. The benchmark is based on measurements presented in previous research.

The water channels and copper plate domain dimensions and spacing are shown in Figure 4.10(a) and Figure 4.11. Water properties are all calculated as function of temperature. The thermophysical properties of the copper used in this test case are summarized in Table 4.5. The temperatures measured at the wall illustrated in Figure 4.11 and the thermocouples position are shown in Figure 4.10(b). The inlet water temperature is 303.15 K and the velocity 7.8 m/s. The averaged water temperature rise is 7.1 K.

4.3.7. Benchmark 7

This benchmark is based on numerical simulations conducted by Xie and Chen [15]. A section of the studied mold is shown in Figure 4.12. The geometric and process parameters are summarized in Table 4.6. The temperature profile applied at the hot face of the mold as a first kind of boundary condition is shown in Figure 4.13. The thermophysical properties of copper are the same as in Table 4.5. For this benchmark case, the temperature distribution computed by Xie and Chen is available for the copper mold and water channels at different sections

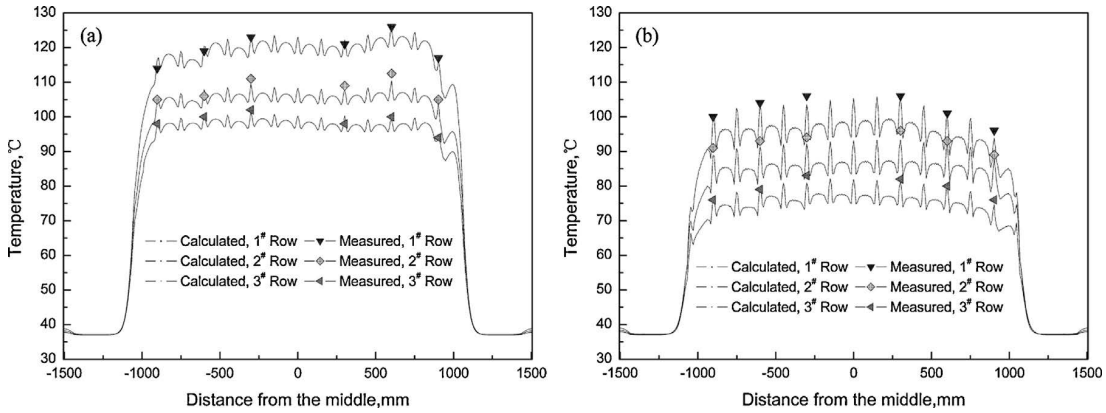


Figure 4.9: Measured temperature of Benchmark 5. In the figure these measurements are compared with the temperature profiles computed by Du et al. [23]. The measurement are the averaged temperatures over an interval of 20 min. (a) Inside radius; (b) outside radius.

Table 4.5: Thermophysical properties of copper used in Benchmark 6 [15].

Material	Specific heat [J/kgK]	Thermal conductivity [W/mK]	Density [kg/m^3]
Copper	410	335 at 298 K	8940
		315 at 393 K	
		310 at 623 K	

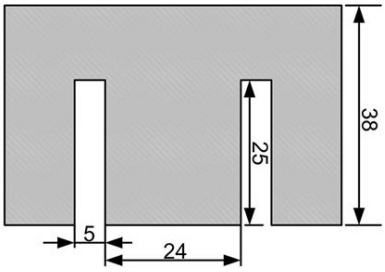
of the mold.

Table 4.6: Geometry and process parameters of Benchmark 7 [15].

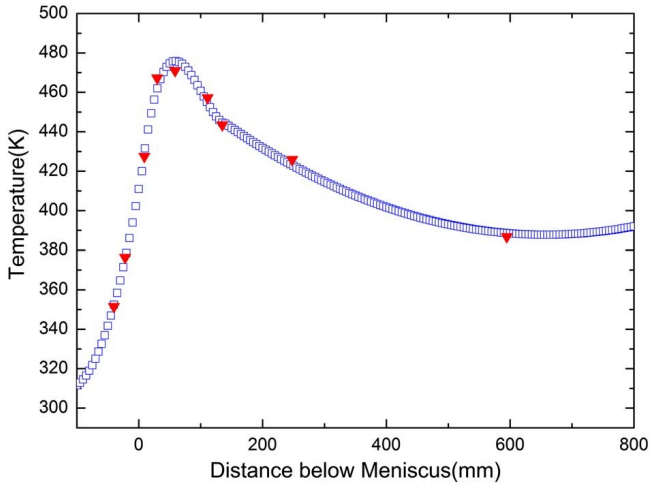
Item	Value
Slab cross-section	250x2000 mm^2
Mold height	900 mm
Mold working length	800 mm
Casting speed	1.15 m/min
Water flow	
Wide plate	5009 L/min
Narrow plate	515 L/min
Water inlet velocity	
Wide plate	8.382 m/s
Narrow plate	8.086 m/s
Water inlet pressure	1 MPa

4.3.8. Benchmark 8

A benchmark case will be designed base on data provided by Danieli. The idea is to design a three dimensional test case in which the water channels have three-dimensional geometry as in Danieli's mold. In this case is not possible to assume the water temperature along the channels. Consequently, they would have to be included in the domain of the simulation and properly simulated.



(a) Water channel geometry and thermocouples plane position.



(b) Temperature measurements (triangles) and temperature profile used for validation (squares).

Figure 4.10: Geometry of Benchmark 6 [15]
(a) and relative temperature profile (b). In (b) the temperature measured by the thermocouples are indicated as red triangles while the temperature profile was computed by Xie and Chen [15] fitting the measurements, to validate their model.

Bibliography

[1] World Steel Association, "World steel in figures 2018," *World Steel Association: Brussels, Belgium*, 2018.

[2] B. G. Thomas, "Review on modeling and simulation of continuous casting," *Steel Research Int.*, vol. 89, no. 1, p. 1700312, 2018.

[3] L. Klimeš and J. Štětina, "A rapid gpu-based heat transfer and solidification model for dynamic computer simulations of continuous steel casting," *Journal of Materials Processing Technology*, vol. 226, pp. 1–14, 2015.

[4] "Danieli & C. Officine Meccaniche S.p.A." https://www.danieli.com/en/_1.htm, accessed: 2018-09-11.

[5] B. G. Thomas and F. M. Najjar, "Finite element modelling of turbulent fluid flow and heat transfer in continuous casting," *Applied Mathematical Modelling*, vol. 15, no. 5, pp. 226 – 243, 1991.

[6] B. Thomas, G. Li, A. Moitra, and D. Habing, "Analysis of thermal and mechanical behavior of copper molds during continuous casting of steel slabs," *Iron and Steelmaker(USA)*, vol. 25, no. 10, pp. 125–143, 1998.

[7] A. Gupta, R. Singh, A. Paul, and S. Kumar, "Effect of mould geometry, coating, and plate thickness on

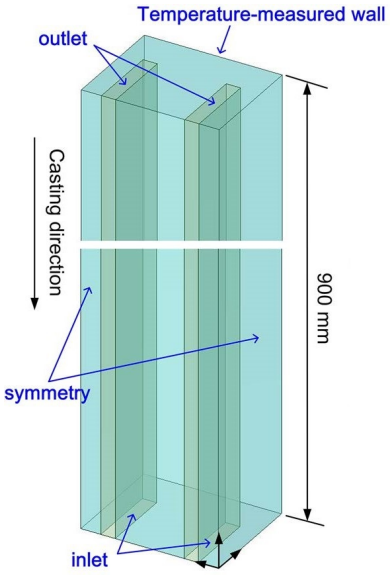


Figure 4.11: Domain of Benchmark 6 [15]. The lateral faces are plane of symmetry, the other walls were adiabatic.

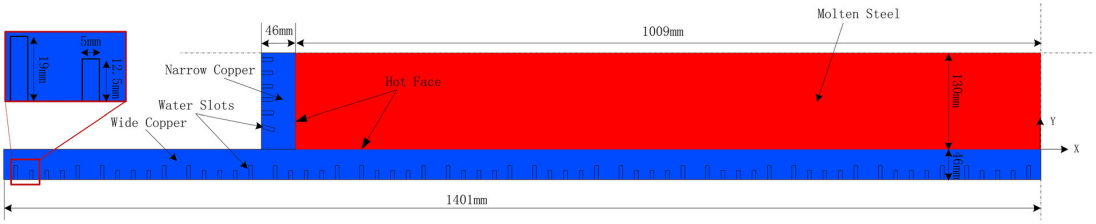


Figure 4.12: Cross section of the mold of Benchmark 7 [15]. Only a quarter of the mold is shown because of the symmetry.

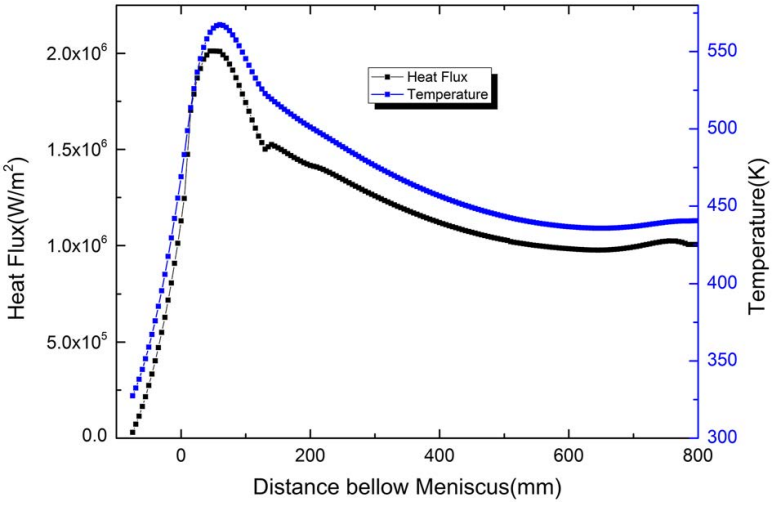


Figure 4.13: Temperature profile applied as boundary condition at the hot face of the mold in Benchmark 7 [15] and related heat flux.

the thermal profile of continuous casting moulds," *Journal of the Southern African Institute of Mining and Metallurgy*, vol. 118, no. 5, pp. 505–513, 2018.

[8] R. B. Mahapatra, J. K. Brimacombe, and I. V. Samarasekera, "Mold behavior and its influence on quality in the continuous casting of steel slabs: Part ii. mold heat transfer, mold flux behavior, formation of oscillation marks, longitudinal off-corner depressions, and subsurface cracks," *Metallurgical and Materials Transactions B*, vol. 22, no. 6, pp. 875–888, Dec 1991.

[9] J. Savage and W. Pritchard, "The problem of rupture of the billet in the continuous casting of steel," *Journal of the Iron and Steel Institute*, vol. 178, no. 3, pp. 269–277, 1954.

[10] D. Fengming, W. Xudong, L. Yu, W. Shanjiao, Z. Ze, and Y. Man, "Investigation on thermo-mechanical behavior of mold corner for continuous casting slab," *ISIJ International*, vol. 55, no. 10, pp. 2150–2157, 2015.

[11] J. Yang, Z. Cai, and M. Zhu, "Transient thermo-fluid and solidification behaviors in continuous casting mold: Evolution phenomena," *ISIJ International*, vol. 58, no. 2, pp. 299–308, 2018.

[12] C. Sleicher and M. Rouse, "A convenient correlation for heat transfer to constant and variable property fluids in turbulent pipe flow," *International Journal of Heat and Mass Transfer*, vol. 18, no. 5, pp. 677 – 683, 1975.

[13] I. Samarasekera and J. K. Brimacombe, "The thermal field in continuous-casting moulds," *Canadian Metallurgical Quarterly*, vol. 18, no. 3, pp. 251–266, 1979.

[14] Y. Meng and B. G. Thomas, "Heat-transfer and solidification model of continuous slab casting: Con1d," *Metallurgical and Materials Transactions B*, vol. 34, no. 5, pp. 685–705, Oct 2003.

[15] X. Xie, D. Chen, H. Long, M. Long, and K. Lv, "Mathematical modeling of heat transfer in mold copper coupled with cooling water during the slab continuous casting process," *Metallurgical and Materials Transactions B*, vol. 45, no. 6, pp. 2442–2452, Dec 2014.

[16] X. Wang and M. Yao, "Neural networks for solving the inverse heat transfer problem of continuous casting mould," in *2011 Seventh International Conference on Natural Computation*, vol. 2, July 2011, pp. 791–794.

[17] H. Zhang, W. Wang, F. Ma, and L. Zhou, "Mold simulator study of the initial solidification of molten steel in continuous casting mold. part i: Experiment process and measurement," *Metallurgical and Materials Transactions B*, vol. 46, no. 5, pp. 2361–2373, Oct 2015.

[18] F. ming DU, X. dong WANG, Y. LIU, T. yi LI, and M. YAO, "Analysis of non-uniform mechanical behavior for a continuous casting mold based on heat flux from inverse problem," *Journal of Iron and Steel Research, International*, vol. 23, no. 2, pp. 83 – 91, 2016.

[19] H. Zhang, W. Wang, and L. Zhou, "Calculation of heat flux across the hot surface of continuous casting mold through two-dimensional inverse heat conduction problem," *Metallurgical and Materials Transactions B*, vol. 46, no. 5, pp. 2137–2152, Oct 2015.

[20] Udayraj, S. Chakraborty, S. Ganguly, E. Chacko, S. Ajmani, and P. Talukdar, "Estimation of surface heat flux in continuous casting mould with limited measurement of temperature," *International Journal of Thermal Sciences*, vol. 118, pp. 435 – 447, 2017.

[21] Y. A. Cengel, *Heat transfer: a practical approach*. Boston, Mass: WBC McGraw-Hill, 2007.

[22] A. Lehtinen and R. Karvinen, "Analytical solution for a class of flat plate conjugate convective heat transfer problems," *Frontiers in Heat and Mass Transfer*, vol. 2, pp. 1–6, 2011.

[23] F. Du, X. Wang, M. Yao, and X. Zhang, "Analysis of the non-uniform thermal behavior in slab continuous casting mold based on the inverse finite-element model," *Journal of Materials Processing Technology*, vol. 214, no. 11, pp. 2676 – 2683, 2014.

5. Benchmark for numerical simulation of thermo-mechanical phenomena arising in blast furnaces

Nirav Shah[1], Gianluigi Rozza[1], Peregrina Quintela[2], Patricia Barral[2], Michele Girfoglio[1]

[1] *Scuola Internazionale Superiore di Studi Avanzati*

[2] *Instituto Tecnolóxico de Matemática Industrial, Universidade de Santiago de Compostela*

Abstract. The high temperature process occurring inside blast furnace hearth and correspondingly high thermal stresses induced in blast furnace walls pose significant challenge for mechanical design. We present here steady state thermo-mechanical coupled model to numerically compute thermal stresses. The model involves classical equations of energy conservation and momentum conservation from continuum mechanics. In the present report, axisymmetric formulation consisting of conduction heat transfer and momentum conservation is considered. Further, the domain considered is isotropic and homogeneous. Subsequently, separate benchmark cases for each individual model complexity ensure step by step error resolution. At the end of the report, envisioned roadmap for reduced basis method and computation is presented. These methods are aimed at quick and reliable transfer of numerical results to real industrial problem. Additionally, model hierarchy for full order model, taking into account practical complexities, gives future extensions from the current simplified model.

Keywords: Blast furnace hearth, thermo-mechanical coupled model, finite element method, reduced basis method, multiphysics problem, benchmark verification.

5.1. Conceptual model

Blast furnace is a metallurgical furnace used to produce iron from ”charge”. The ”charge” contains iron ore, limestone and coke. The process requires high temperature to increase rate of oxidation of carbon, an exothermic reaction. In the hearth the hot air, supplied through Tuyere, acts as source of oxygen and is used for oxidation of carbon. The temperature in the hearth is as high as 1500°C. The general layout of blast furnace is shown in Figure 5.1.

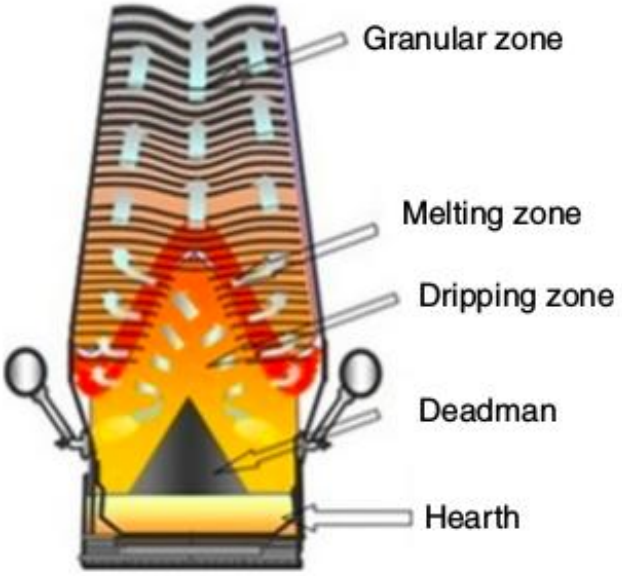


Figure 5.1: Blast furnace divided in zones [1]

As mentioned by Swartling M. et. al. [2] the hearth is made up of several refractory zones (Figure 5.2), with each zone having specific function and accordingly, design requirements. The outside of the hearth is covered with a steel shell (Figure 5.2). The wall and bottom are constructed by five types of refractory materials, each with its specific properties depending on type of environment the region is exposed to. The wall is made up of carbon material having relatively high heat conductivity to keep the inner wall at a low temperature. The upper layer of the bottom is a ceramic plate with the composition 23.5 % SiO_2 , 73.5 % Al_2O_3 and some additives. It has high resistance to mechanical wear, which helps to keep the bottom layer intact and avoid cracks. That is important to avoid penetration and solidification of liquid iron in the lower bottom layers, since this would have impact on the heat flow. Between the steel shell and the bottom refractory is a layer of ramming paste. The taphole region is made up of carbon refractory. It has very high heat conductivity to transport heat from the taphole during tapping and reduce the thermal stresses in the region.

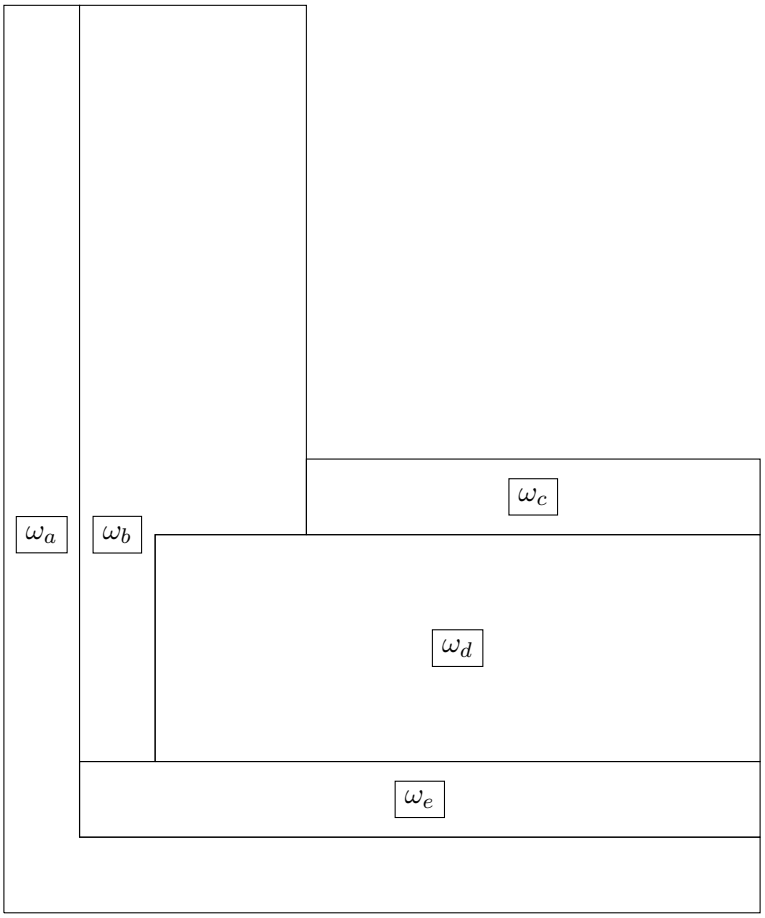


Figure 5.2: Scheme of Hearth axisymmetric geometry ω : ω_a) Steel shell, ω_b) Carbon wall, ω_c) Ceramic plate, ω_d) Lower bottom layer, ω_e) Ramming paster [2]

The proper materials selection ensures less heat loss and also less thermal stresses. Minimising the thermal stresses and the mechanical effects of the hot metal flow constitute a design challenge. The wall of the hearth is subjected also to abrasive wear due to oxidation, abrasion/deterioration due to chemical attack and erosion due to hot liquid flow. The heat transfer through the furnace wall is due to conduction process, and through the region of fluid flow is due to convection process.

5.2. Practical significance of the project

Wear of carbon refractory limits the hearth lifetime, which in turn restricts lifetime of blast furnace. An optimum design enlarges the hearth lifetime and consequently, the blast furnace campaign. Among different

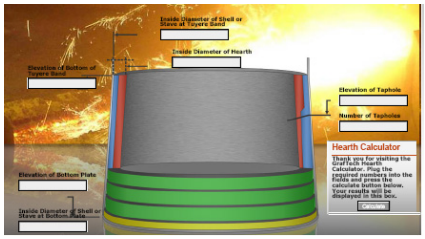
families of design, the most commonly used is considered in this work.

The design parameters which affect the hearth design are material properties, operating conditions and geometric dimensions. If carbon blocks are used, frictional contact could occur between block surfaces. Other materials used are ceramic cup bricks or castable material for ramming mix. The parameters related to geometry of hearth such as thickness of hearth material are of interest for geometric parametrization (Figure 5.4).

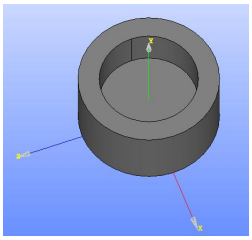
The benchmark cases introduced in this document correspond to some thermo-mechanical models arising in the blast furnace hearth, and are based on the extensive experience of Global Research and Development Center of ArcellorMittal in Asturias, Spain. The project is carried out under able supervision from ITMATI and SISSA.

5.3. Mathematical model

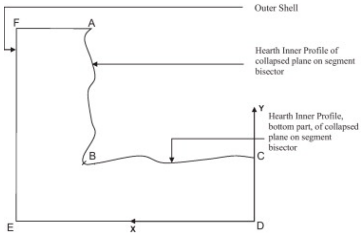
The governing equations for thermo-mechanical modelling are linear momentum conservation and energy conservation from continuum mechanics.



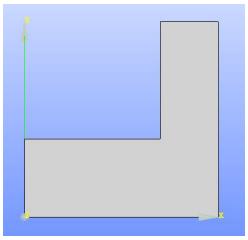
(a) 3 Dimensional hearth geometry [3]



(b) Hearth 3 dimensional computational domain



(c) Axisymmetric hearth geometry [4]



(d) Hearth axisymmetric computational domain

Figure 5.3: Hearth geometry and computational domain

The mathematical model is based on the following assumptions which lead to reasonable simplifications :

1. The model is assumed to be axisymmetric. Taphole operation is not part of this study, i.e. only normal operating conditions are considered.
2. We only focus on the steady state operations.
3. At first iteration hearth is made up of single material and is isotropic. After successful benchmark tests anisotropy and inhomogeneity will be considered.
4. Initially linear elasticity, heat transfer only by conduction within hearth walls and by convection with surrounding and no contact between bricks are considered. After successful benchmark tests of simplified models, the following non linearities will be considered : Contact between refractory blocks, radiation heat transfer and material property dependence on temperature.

We describe the circular cross section of blast furnace by r, θ and vertical axis of blast furnace by y (cylindrical coordinate system). We introduce now simplified 3-dimensional domain Ω, corresponding to the solid part of the hearth, in cylindrical coordinates as $\omega \times [0, 2\pi)$. Ω refers to 3-dimensional computational domain and its vertical section ω, refers to axisymmetric computational domain (Figure 5.3). We denote the boundary of Ω as $\partial\Omega$ and the boundary of ω as $\partial\omega$. Let r_0 be the outer radius of Ω and y_{min}, y_{max} are its minimum and maximum $y-$coordinates respectively. We represent $\overrightarrow{e_y}$, $\overrightarrow{e_r}$ and $\overrightarrow{e_\theta}$, the unit vectors corresponding to $y-, r-$ and $\theta-$ directions respectively. $\overrightarrow{n}$ is the unit normal vector pointing outwards. We use the following notations

for boundaries of Ω (Figure 5.4).

$$\begin{aligned} \Gamma_{out} &= \partial\Omega \cap (r \equiv r_0) \,, \\ \Gamma_{+} &= \partial\Omega \cap (y \equiv y_{max}) \,, \\ \Gamma_{-} &= \partial\Omega \cap (y \equiv y_{min}) \,, \\ \Gamma_{sf} &= \partial\Omega \backslash (\Gamma_{out} \cup \Gamma_{+} \cup \Gamma_{-}) \,. \end{aligned}$$

The corresponding two dimensional boundaries of ω are denoted using γ instead of Γ (Figure 5.4). Notice that a new boundary appears on the axi-symmetric domain:

$$\gamma_s = \partial\omega \cap (r \equiv 0) \,.$$

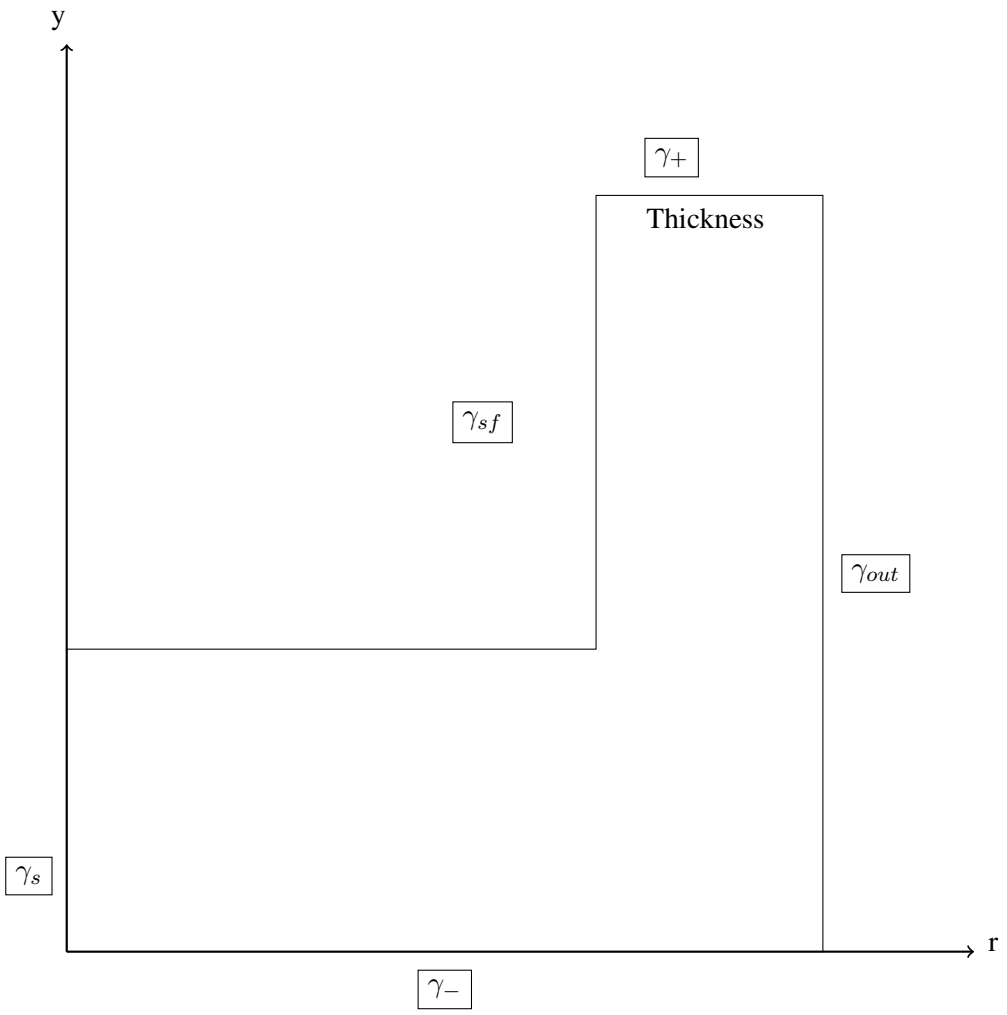


Figure 5.4: Boundaries of axi-symmetric computational domain

Based on the assumptions listed above, the momentum conservation (for small displacements) and energy conservation can be written as,

$$-Div(\boldsymbol{\sigma}) = \overrightarrow{0} \text{ in } \Omega \,, \tag{5.1}$$

$$-Div(\boldsymbol{K}\nabla T) = 0 \text{ in } \Omega \,. \tag{5.2}$$

Here $\boldsymbol{K}$ refers to thermal conductivity tensor, $\boldsymbol{\sigma}$ refers to stress tensor, and $T = T(r, \theta, y)$ refers to temperature. The stress tensor $\boldsymbol{\sigma}$ is related to strain tensor through Hooke's law:

$$\boldsymbol{\sigma}(\overrightarrow{u}) = \lambda \operatorname{Tr}(\boldsymbol{\varepsilon}(\overrightarrow{u}))\boldsymbol{I} + 2\mu\boldsymbol{\varepsilon}(\overrightarrow{u}) - (2\mu + 3\lambda)\alpha(T - T_0)\boldsymbol{I} \,, \tag{5.3}$$

where $\boldsymbol{I}$ refers to the identity tensor, strain tensor $\boldsymbol{\varepsilon}(\overrightarrow{u})$ is defined as,

$$\boldsymbol{\varepsilon}(\overrightarrow{u}) = \frac{1}{2}(\nabla\overrightarrow{u} + \nabla\overrightarrow{u}^T) \,, \tag{5.4}$$

$\overrightarrow{u}$ being the displacement vector,

$$\overrightarrow{u} = u_r(r, \theta, y)\overrightarrow{e_r} + u_\theta(r, \theta, y)\overrightarrow{e_\theta} + u_y(r, \theta, y)\overrightarrow{e_y} \,, \tag{5.5}$$

T_0 is the reference temperature, α is the thermal expansion coefficient, and λ and μ are the Lame parameters of the material. The latter can be expressed in terms of Young modulus, E, and the Poisson ratio, ν as:

$$\mu = \frac{E}{2(1+\nu)} \,,\ \lambda = \frac{E\nu}{(1-2\nu)(1+\nu)} \,. \tag{5.6}$$

On the upper boundary (Γ_+), the weight of other blast furnace components is acting and no conduction heat transfer are considered :

$$\begin{aligned} (-\boldsymbol{K}\nabla T)\cdot\overrightarrow{n} &= 0 \,, \\ \overrightarrow{\sigma_t} = \overrightarrow{0}, \sigma_n &= |\overrightarrow{W_s}| \,, \end{aligned} \tag{5.7}$$

where $\overrightarrow{W_s}$ is the weight from other furnace components supported by hearth walls, $\overrightarrow{\sigma_t}$ and σ_n are tangential and normal forces :

$$\begin{aligned} \sigma_n &= (\boldsymbol{\sigma}\overrightarrow{n})\cdot\overrightarrow{n} \,, \\ \overrightarrow{\sigma_t} &= \boldsymbol{\sigma}\overrightarrow{n} - \sigma_n\overrightarrow{n} \,. \end{aligned}$$

On the bottom boundary (Γ_-), the temperature is assumed to be known, the normal displacement is null and shear force is assumed to be zero, therefore on this boundary,

$$\begin{aligned} T &= T_D \,, \\ \overrightarrow{u}\cdot\overrightarrow{n} = 0, \overrightarrow{\sigma_t} &= \overrightarrow{0} \,. \end{aligned} \tag{5.8}$$

On inner boundary (Γ_{sf}), a convection heat transfer with the liquid phase occurs and hydrostatic pressure is acting :

$$\begin{aligned} (-\boldsymbol{K}\nabla T)\cdot\overrightarrow{n} &= \alpha_f(T_f - T) \,, \\ \boldsymbol{\sigma}\overrightarrow{n} &= -p_h\overrightarrow{n} \,, \end{aligned} \tag{5.9}$$

where T_f is the fluid temperature assumed to be constant at the steady state, α_f the convective heat transfer coefficient, and p_h is the hydrostatic pressure computed as:

$$p_h = \rho_f g h, \tag{5.10}$$

being ρ_f the mass density of the fluid, g the gravitational accelaration, and h height of fluid column.

On outer boundary (Γ_{out}), convective heat flux and no deformation in radial direction are assumed :

$$\begin{aligned}(-\boldsymbol{K}\nabla T)\cdot\overrightarrow{n} &= \alpha_{out}(T_{out}-T)\ ,\\ \overrightarrow{u}\cdot\overrightarrow{n} = 0, \overrightarrow{\sigma_t} &= \overrightarrow{0}\ ,\end{aligned} \tag{5.11}$$

α_{out} being the heat transfer coefficient, and T_{out} the ambient temperature.

It can be observed that the introduced model (boundary conditions, body forces and heat source term) is independent of θ and hence an axisymmetric model hypothesis is applicable. The axisymmetric model leads to significant computational savings as the 3-Dimensional model in (r, y, θ) can now be described by only 2-coordinates (r, y). In this axisymmetric system, we represent the displacement $\overrightarrow{u}$ and temperature T, both independent of θ as,

$$\begin{aligned}\overrightarrow{u} &= u_r(r,y)\overrightarrow{e_r} + u_y(r,y)\overrightarrow{e_y}\ ,\\ T &= T(r,y)\ .\end{aligned} \tag{5.12}$$

The associated axisymmetric model is reduced to consider conservation equations (5.1), (5.2) and boundary conditions (5.7), (5.8), (5.9) and (5.11) replacing the Ω domain by its vertical section ω, and the Γ boundaries by γ such that $\Gamma = (\gamma\backslash\gamma_s)\times[0, 2\pi)$, adding the usual condition of symmetry over γ_s :

$$(-\boldsymbol{K}\nabla T)\cdot\overrightarrow{n} = 0\ ,\ \overrightarrow{u}\cdot\overrightarrow{n} = 0\ ,\ \overrightarrow{\sigma_t} = \overrightarrow{0} \text{ over } \gamma_s\ . \tag{5.13}$$

5.4. Numerical model

The finite element based simulation of coupled thermo mechanical system arising in blast furnace is well known in various scientific literatures such as [5]. It solves energy and momentum conservation equation. The model is considered steady and the bricks as isotropic and homogeneous.

The thermo mechanical behavior for blast furnace hearth lining was also investigated by Brulin et. al. [6] based on the modified Cam-Clay material model using Finite Element Method. This approach takes into account non linear elastic and plastic behavior. It introduces a yield function and plastic state based on an assumed yield function.

Kaymak [7] introduced a fast and simplified contact model for blast furnace hearth thermo-mechanical behavior simulation. This model uses the fact that contact occurs under compression whilst during tension there is no contact. This eliminates tensile stresses between contacts.

Different softwares packages are available for the simulation such as ANSYS (https://www.ansys.com/), Code_Aster [8] or FEniCS [9]. We plan to use Code_Aster [8] which is an open access software developed for analysis of structures and thermomechanics for studies and research. During benchmark tests we compare the solutions provided by Code_Aster against solution provided by FEniCS [9].

The basic workflow for solving coupled thermo mechanical problem involves solving the energy equation first to get the values of temperature at the degrees of freedom. Their temperature values are then inserted into the momentum conservation equation to compute the effects related to thermal stresses (Figure 5.5).

5.5. Parametrization

The parameters of interest are thermal and mechanical properties of the material and geometric variations i.e. geometric parameters.

The relevant material properties for mechanical design are Lame parameters or equivalently, Young's modulus and Poisson's ratio or shear modulus. From thermal design point of view thermal conductivity and convective heat transfer coefficients are relevant.

The geometric parameters for the current problem are thickness of hearth lining. Depending on the design adopted the geometry and accordingly geometric parameters could considerably change.

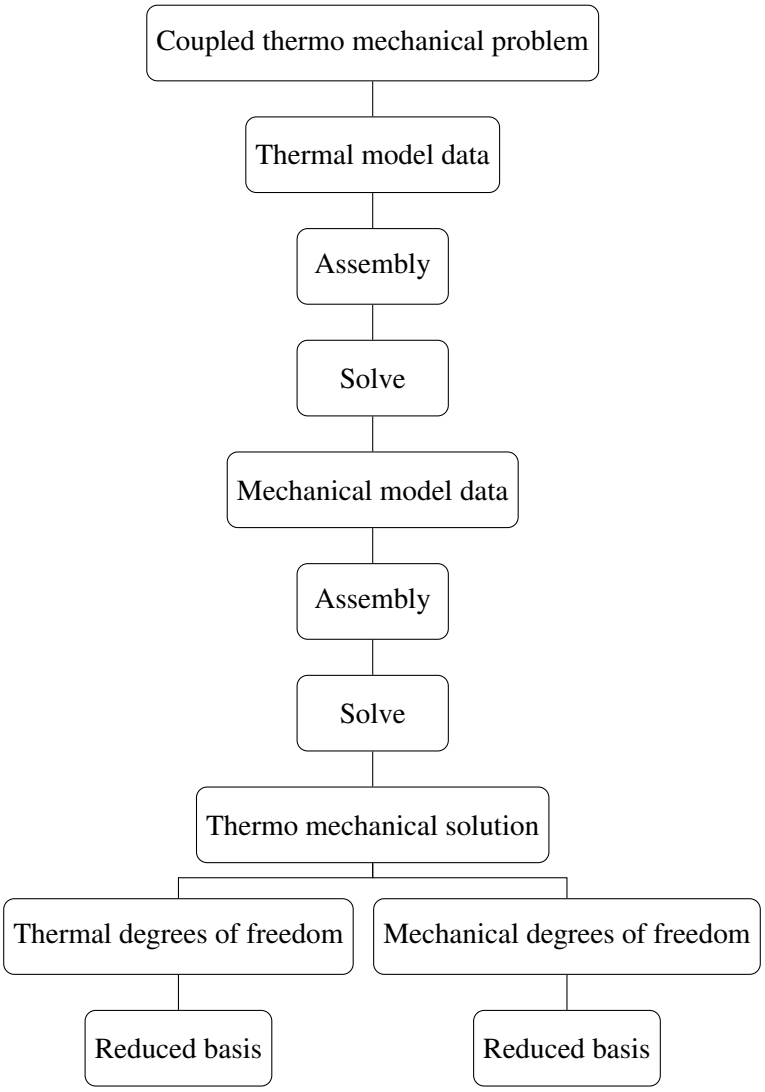


Figure 5.5: Workflow for numerical simulation of thermo-mechanical coupled system

5.6. Model order reduction

Model order reduction (MOR), also known as reduced basis methods, for multiphysics problems have recently gained attention within model order reduction community. The projection based methods are well established for many problems, however, preservation of block structure and/or preservation of passivity of original system has given rise to different approaches. The challenge is to transfer these favorable properties from full order model to reduced order model. We discuss below several methods for model order reduction. A careful assessment will be made after implementation and verification of the simulation using a finite element method for the present problem. If required, comparison will be made between different approaches. It should be noted that reduced order solution is calibrated against full order solution, which in turn is compared against benchmark solution. Therefore, separate benchmark problem for reduced basis solution is not necessary. As discussed in section 5.7, established reduced basis library RBniCS (http://mathlab.sissa.it/rbnics, [10]) is foreseen to be used.

We discuss now some multiphysics reduced basis methods. For further details, we refer to [11].

5.6.1. Standard Krylov subspace method

A system of the form $Ax = b$ with $A \in \mathbb{R}^{m\times m}$, $x \in \mathbb{R}^m$, $b \in \mathbb{R}^m$, and $r_k = Ax - b$, with $r_k \in \mathbb{R}^m$ and k being the number of iteration of solver, the n-dimensional standard Krylov subspace $\tilde{V}$ is created by,

$$\tilde{V} = \mathcal{K}_n(A, r_0) = colspan\{r_0, Ar_0, A^2r_0, ..., A^{n-1}r_0\} \, . \tag{5.14}$$

The standard Krylov subspace method, however, is not structure preserving and hence one loses the original matrix structure in reduced basis space [11]. In some structure preserving methods the Krylov subspace method is modified in order to preserve the block matrix structure of the original system. The moment matching property is realized by projecting the original system matrices onto the appropriate input and output based Krylov subspaces [11].

5.6.2. Order reduction by projection

In this method the original matrix A is reduced by projecting onto its subspace with matrix V,

$$\tilde{A} = V^T AV \, . \tag{5.15}$$

As pointed out by Freund [12], the projection approach preserves the passivity of the original system.

5.6.3. Structure-Preserving Reduced-order Interconnect Macromodelling (SPRIM)

The main computational step in SPRIM is identical to PRIMA : Passive reduced-order interconnect macromodelling algorithm [13]. This step aims at creating n-dimensional Krylov subspace $\mathcal{K}_n(A, r)$. The PRIMA method creates the reduced basis space using matrix $\tilde{V}$ (equation (5.14)) containing orthonormal basis for the Krylov space. The strong advantage of this method is preservation of stability and passivity of the original system [14]. In SPRIM, the vectors in matrix $\tilde{V}$ are placed diagonally in matrix $\tilde{V}_{sprim}$ and matrix $\tilde{V}_{sprim}$ is used in place of $\tilde{V}$.

5.7. Problem characteristics and Benchmark cases

Based on the description in conceptual and mathematical model, the following characteristics problems need to be tested with benchmark cases. Full order simulation is performed by Code_Aster [8] (www.code-aster.org). Additionally, for benchmark comparison, full order simulation will also be performed by FEniCS [9] (https://fenicsproject.org/) and/or ANSYS (www.ansys.com). For reduced basis simulation routines from RBniCS [10] (www.mathlab.sissa.it/rbnics) will be used. The assembled matrices from Code_Aster will be used in compatibility with individual RBniCS modules.

We propose separate benchmark cases for each individual characteristic for step by step error resolution. After successful analytical numerical tests, a benchmark problem will be verified with experimental data from industrial partner. The data here refers to temperature measurement by thermocouples, strain gauge measurements and / or stress gauge measurements based on practice followed by industrial partner such as [15].

5.7.1. Energy equation

As mentioned in solution workflow (Figure 5.5), first energy equation is solved. A known analytical temperature function and source term function will be applied. Temperature at each degree of freedom computed with Code_Aster will be compared with Temperature computed with FEniCS and/or ANSYS for the same problem.

5.7.2. Momentum equation

We apply analytical stress function, such as known spherical stress tensor or known shear stress on the domain and compare results provided by FEniCS and ANSYS with Code_Aster simulations.

5.7.3. Coupling

We apply known spherical stress test in the domain. In another test, we subject the domain to only thermal stress. When the domain is subjected to combined effect of known spherical stress test and thermal stress, the principal stresses should be the sum of spherical stress and thermal stress.

5.7.4. Condition number

Due to large difference between order of magnitude of thermal conductivity and that of material parameters such as Young's modulus, the coupled stiffness matrix might have bad condition number. However, as the equations are solved sequentially, the coupled stiffness matrix is not required to be assembled and hence, condition number is not expected to be a major problem.

5.8. Future work

The simplified model introduced here will be modified to consider practical complexities. Specifically, radiation heat transfer and non linear material behavior will be simulated and tested. After simulating this non linear model, the contact between refractory blocks will be simulated and verified with benchmark tests. The model will include necessary complexities and verified with experimental data. Accordingly, benchmark tests will be modified for the verification of simulations with benchmark problems.

Bibliography

[1] D. H. B. Andrade, R. P. Tavares, A. C. B. Quintas, V. E. de Souza Moreira, A. O. Viana, and V. M. Gasparini, "Evaluation of the permeability of the dripping zone and of flooding phenomena in a blast furnace," *Journal of Materials Research and Technology*, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2238785417303769

[2] M. Swartling, B. Sundelin, A. Tilliander, and P. G. Jönsson, "Heat transfer modelling of a blast furnace hearth," *Steel research international*, vol. 81, no. 3, pp. 186–196, 2010. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/srin.200900145

[3] "Graftech international," http://archive.constantcontact.com/fs072/1101483160833/archive/1102208996928.html, webpage.

[4] A. Bhattacharya, A. Debjani, and S. Debjani, "Estimation of operating blast furnace reactor invisible interior surface using differential evolution," *Applied Soft Computing*, vol. 13, no. 5, pp. 2767 – 2789, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1568494612004619

[5] D. Gruber, K. Andreev, and H. Harmuth, "Fem simulation of the thermomechanical behaviour of the refractory lining of a blast furnace," *Journal of Materials Processing Technology*, vol. 155-156, pp. 1539 – 1543, 2004, proceedings of the International Conference on Advances in Materials and Processing Technologies: Part 2. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0924013604006715

[6] J. Brulin, F. Roulet, A. Rekik, E. Blond, A. Gasser, R. Mc Nally, and M. Micollier, "Latest Evolution in Blast Furnace Hearth thermo-Mechanical Stress Modelling," in *4th International Conference on Modelling and Simulation of Metallurgical Processes in Steelmaking*, Dusseldorf, France, Jun. 2011, p. CD rom. [Online]. Available: https://hal.archives-ouvertes.fr/hal-00611744

[7] Y. Kaymak, "A simplified approach to the contact in thermo-mechanical analysis of refractory linings," *Sohnstr*, vol. 65, p. 40237, 2007.

[8] Electricité de France, "Finite element Code_Aster, Analysis of Structures and Thermomechanics for Studies and Research," 1989–2017.

[9] M. S. Alnæs, J. Blechta, J. Hake, A. Johansson, B. Kehlet, A. Logg, C. Richardson, J. Ring, M. E. Rognes, and G. N. Wells, "The fenics project version 1.5," *Archive of Numerical Software*, vol. 3, no. 100, 2015.

[10] J. S. Hesthaven, G. Rozza, and B. Stamm, *Certified Reduced Basis Methods for Parametrized Partial Differential Equations*, 1st ed., ser. Springer Briefs in Mathematics. Switzerland: Springer, 2015.

[11] W. H. A. Schilders and A. Lutowska, *A Novel Approach to Model Order Reduction for Coupled Multiphysics Problems*. Cham: Springer International Publishing, 2014, pp. 1–49. [Online]. Available: https://doi.org/10.1007/978-3-319-02090-7_1

[12] R. W. Freund, *The SPRIM Algorithm for Structure-Preserving Order Reduction of General RCL Circuits*. Dordrecht: Springer Netherlands, 2011, pp. 25–52. [Online]. Available: https://doi.org/10.1007/978-94-007-0089-5_2

[13] A. Odabasioglu, M. Celik, and L. T. Pileggi, "Prima: passive reduced-order interconnect macromodeling algorithm," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 17, no. 8, pp. 645–654, Aug 1998.

[14] W. Schilders, *Introduction to Model Order Reduction*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 3–32. [Online]. Available: https://doi.org/10.1007/978-3-540-78841-6_1

[15] R. M. Duarte, I. Ruiz-Bustinza, D. Carrascal, L. F. Verdeja, J. Mochón, and A. Cores, "Monitoring and control of hearth refractory wear to improve blast furnace operation," *Ironmaking & Steelmaking*, vol. 40, no. 5, pp. 350–359, 2013. [Online]. Available: https://doi.org/10.1179/1743281212Y.0000000045

Part III.

Optimization methods

6. Benchmark for high-performance algorithms in adaptive optics control

Bernadett Stadler[1], Ronny Ramlau[1]

[1] *Industrial Mathematics Institute, Johannes Kepler Universität Linz*

Abstract. The new generation of ground-based extremely large telescopes require highly efficient algorithms to achieve an excellent image quality in a large field of view. These systems rely on adaptive optics (AO), where one aims to compensate in real-time the rapidly changing optical distortions in the atmosphere. Due to the steady growing of telescope sizes, the computational load is increasing drastically and current algorithms become infeasible. After a brief introduction to astronomical adaptive optics we compare the frequently used reconstruction method MVM to two novel ones. The first one, called Cumulative Reconstructor with Domain Decomposition (CuReD), is an extremely fast reconstruction algorithm for wavefront sensor data. The second approach is the Finite Element Wavelet Hybrid Algorithm (FEWHA), which tackles the problem of atmospheric tomography efficiently and accurately. Finally, we describe the simulation environment OCTOPUS, which is provided by the ESO and used in our benchmark cases for quality evaluation.

Keywords: Adaptive optics, wavefront reconstruction, atmospheric tomography.

6.1. Introduction

The new generation of planned earthbound Extremely Large Telescopes (ELT) aims at excellent image quality in a large field of few. Such systems rely on Adaptive Optics (AO) systems with the task to correct optical distortions caused by atmospheric turbulences.

To achieve such a correction, the deformations of optical wavefronts, emitted by natural or artificial guided stars, are measured via wavefront sensors and, subsequently, corrected using deformable mirrors (DMs). Many of such systems require the reconstruction of the turbulence profile in the atmosphere, which is called atmospheric tomography. Mathematically, such problems are ill-posed, i.e., the recovery of the solution from noisy measurements is unstable. These underlying ill-posed problems have to be solved in real-time, as the atmospheric turbulences changes within milliseconds. The complex setup of such a system, in particular the huge amount of data that has to be processed in real-time, lead to non-trivial conditions on the used algorithms.

6.1.1. Industrial Partner

Microgate [1] is an Italian company located in Bolzano. The engineering division mainly operates in the field of astronomy. Their focus lies on the development of control systems for adaptive mirrors, real time computers for adaptive optics, adaptive optics sensors and radioastronomy. These systems have become the centre of the most innovative research worldwide in the field of astronomy and are used in some of the most important new generation telescopes, e.g., VLT (Very Large Telescope), which was built by the ESO. Furthermore, the company is involved in the ELT project, covering the tasks of control system design and simulation, electronic design and electronic manufacturing.

6.2. Physical Description of the Problem - Astronomical Adaptive Optics

Turbulent air motion in the atmosphere admits fluctuations of the refractive index. Light, that is initially travelling through the atmosphere as planar waves is distorted. The aim of an AO system is to mechanically correct this distortions through deformable mirrors.

6.2.1. Basics of Imaging

In this section the complete mapping from the object emitting light to the image of this object is defined, see, e.g., [2]. We assume an object to be a cloud of point sources of light. A perfect image is the point where light rays converge in absence of perturbations due to the low of geometric optics. The real image, including optical errors, is defined by the energy distribution over the image plane and is expressed by the so called point spread function (PSF). This setting leads to the following definition of the intensity of an object

$$I_R(x,y) = \int_{\mathbb{R}^2} \mathcal{P}(x-\xi, y-\eta) \cdot I_G(\xi,\eta) d\xi d\eta, \tag{6.1}$$

which can be written as convolution

$$I_R = \mathcal{P} * I_G. \tag{6.2}$$

Using the well-known convolution theorem yields

$$\mathcal{F}\{I_R\} = \mathcal{H} \cdot \mathcal{F}\{I_G\}, \tag{6.3}$$

whit $\mathcal{H} := \mathcal{F}\{\mathcal{P}\}$ called optical transfer function (OTF).

6.2.2. AO Components

The aim of an AO system is to correct optical distortions caused by atmospheric turbulences. To achieve such a correction, the system uses natural or artificial guided stars to emit the deformation of the wavefronts. The wavefronts are measured via wavefront sensors and, subsequently, corrected using DMs. In the following subsections the main components of an AO system, i.e., the guide star, the wavefront sensor and the deformable mirror are briefly described. For more details, we refer to [2].

6.2.2.1. Guide Stars

In order to reach the goals of the AO system, i.e., to correct the distortions of the wavefront, the atmosphere needs to be sufficiently known. For that purpose, light from a bright guide star (GS) is used as a reference for the AO system. Such a GS can be either a star in the sky, near the scientific object of interest, or an artificial GS generated by a powerful laser beam called natural guided star (NGS) or laser guided star (LGS), respectively. The laser beams of the LGS stimulates the sodium layer of the atmosphere, leading to disruptions in the emitted light. Therefore, in each AO setting at least one NGS is required to overcome this drawback.

6.2.2.2. Wavefront sensor

A wavefront sensor (WFS) indirectly measures the distortions of the wavefront using the light from GS. Various WFS are used in AO, however, within the framework of this project we focus on two main pupil plane WFS called Shack-Hartmann WFS and pyramid WFS.

Shack-Hartmann WFS

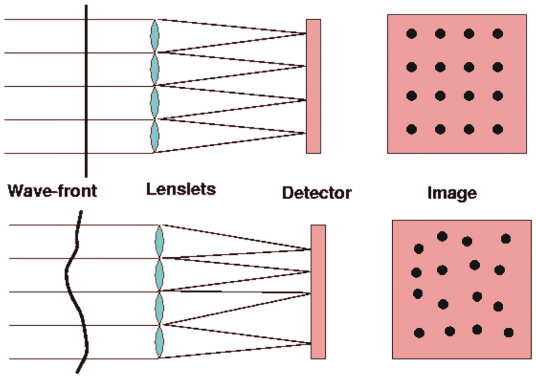


Figure 6.1: Shack-Hartman WFS [3]

A Shack-Hartmann WFS [4] consists of a quadratic array of small lenslets and a CCD photon detector lying behind that array. The quadratic subdomains of the CCD, covered by one single lenslet, are called subapertures. On each subapertures the associated lenslet focusses the light, where the x- and y-coordinates are measured. The slope of the incoming WF is related to the sensor measurements by the centre of gravity

$$S_x(x,y) = \frac{1}{|\Omega_{jk}|}\int_{\Omega_{jk}} \varphi_x(x,y)dxdy, \tag{6.4}$$

$$S_y(x,y) = \frac{1}{|\Omega_{jk}|}\int_{\Omega_{jk}} \varphi_y(x,y)dxdy, \tag{6.5}$$

where $\varphi_x := \frac{\partial \varphi}{\partial x}$ and $\varphi_y := \frac{\partial \varphi}{\partial x}$.

Pyramid WFS

The main component of the pyramid WFS [5] is a four-sided glass pyramidal prism in the focal plane of the telescope. This prism splits the incoming light into four beams. The relay lens, located behind the prism, re-images the beams leading to four different images I_1, I_2, I_3 and I_4 on the CCD camera (Figure 6.2). The two sensor measurements S_x and S_y are given by

$$S_x(x,y) = \frac{(I_1(x,y) + I_2(x,y)) - (I_3(x,y) + I_4(x,y))}{I_0}, \tag{6.6}$$

$$S_y(x,y) = \frac{(I_1(x,y) + I_4(x,y)) - (I_2(x,y) + I_3(x,y))}{I_0}, \tag{6.7}$$

where I_0 denotes the average intensity. Modulating the incoming beam dynamically allows a linearisation of the sensor and to increase its dynamic range. The modulation can be performed either by oscillating the pyramid itself or by using a static diffusive optical element. Two modulation scenarios are possible, namely circular and linear modulation.

The four-sided pyramidal prism of the sensor can be approximated via 2 orthogonally placed two-sided roof prisms. Each of the roofs creates two different images on the detector. The sensor measurements S_x and S_y are obtained by subtracting the two intensity patterns. Due to the physical decoupling of the prisms, the measurements contain information only in $x-$ and $y-$direction, respectively.

6.2.2.3. Deformable Mirror

A DM consists of a thin, flexible, highly reflecting surface, which is controlled by a set of actuators that drive the mirror. There are several possibilities to drive the actuators, such as, electromechanical, electromagnetic

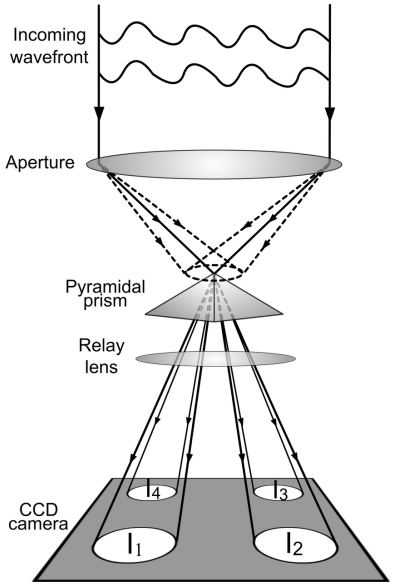


Figure 6.2: Pyramid WFS

and piezoelectric. Within the ELT, a flat adaptive mirror with approximately 5200 actuators is planned, allowing a readjusting of the surface with a very high frequency [6].

6.2.3. AO Systems

Within the framework of the benchmark cases, four AO systems were handled. These systems are briefly described in the following subsections.

6.2.3.1. Single Conjugate AO

If the object of interest, e.g., a star or a galaxy, is located near a bright NGS, the classical AO system Single Conjugate AO (SCAO) is used. In a SCAO system the wavefront is reconstructed using one WFS, that measures the data, and one DM, where the shape is chosen according to the reconstruction. The biggest problem of SCAO systems is that the further away the object of interest is from the NGS, the worse is the correction of the wavefront.

6.2.3.2. Laser Tomography AO

If no NGS is available in the vicinity of the object of interest, the usage of an SCAO system in not possible. The idea is to generate NGS to obtain a good correction. Because one NGS near the object would corrupt the image, due to the finite distance, several NGS are used in the surrounding of the field of view.

Within the framework of a laser tomography AO (LTAO) G_{LGS} and G_{NGS} are used in combination with a single mirror to reconstruct the wavefront. The correction is performed through two steps. The first step is called atmospheric tomography, where the turbulent layers are reconstructed from the sensor measurements. In the second step, the shape of the DM is chosen according to the projection of the wavefront through the reconstructed layers in the direction of interest.

6.2.3.3. Multi Object AO

In contrast to LTAO multi object AO (MOAO) corrects for multiple directions of interest, simultaneously, by using several mirrors. Each mirror corrects for a specific direction. As in the LTAO case a combination of NGS and LGS is used for reconstructing the layers.

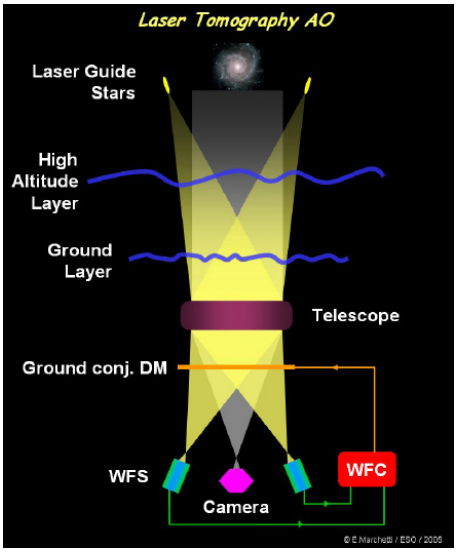


Figure 6.3: Principle of LTAO [7]

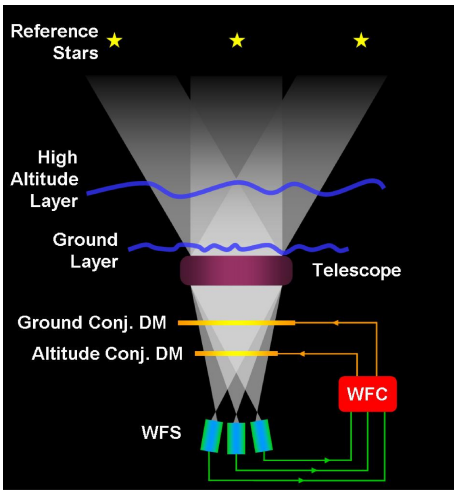


Figure 6.4: Principle of MCAO [7]

6.2.3.4. Multi Conjugate AO

As in MOAO, a Multi Conjugate AO (MCAO) system corrects for multiple directions, however, with the aim to achieve a uniformly optimal correction over the whole field of view and not into specific directions. For that purpose, several DMs are used conjugated to different heights in the atmosphere.

6.2.4. Statistics of the Atmosphere

The distortions of the wavefront are mainly caused by atmospheric turbulence which arises due to an irregular mixing of hot and cold air. The surface of the earth is heated by the sun during the day while it cools down during the night. Together with wind shears, this leads to turbulent air motions. The refractive index of the air is strongly related to the temperatures and, thus, varies due to the irregularities throughout the atmosphere. For this reason, the wavefront, which initially propagates as plane waves, is distorted. In many applications it is sufficient to use a layered model of the atmosphere, i.e., to assume that the turbulence is concentrated on $L \in \mathbb{N}$ layers at certain height.

Usually, the effects of the turbulences are not predictable and, therefore, modelled by random processes. In AO, one assumes to sufficiently know the physics of the atmosphere. There are two very important models, according to Kolmogorov and von Karman, which are briefly described in the following subsections.

6.2.4.1. Kolmogorov Turbulence Model

According to Kolomogorov [8] the behaviour of the atmosphere is modelled via an isotropic, stationary random process. The structure of this random process is based on the structure function, describing the expected difference of values at two points and the covariance function, measuring the spatial covariance. Because we are dealing with a stationary process, both functions only depend on the separation Δx and not on a specific point x.

The structure function of a stationary process f is given by

$$D_f(\Delta x) := \mathbb{E}((f(x+\Delta x) - f(x))^2), \tag{6.8}$$

and the covariance is defined by

$$C_f(\Delta x) := \mathbb{E}((f(x) - \mathbb{E}(f))(f(x + \Delta x) - \mathbb{E}(f))). \tag{6.9}$$

The power spectral density (PSD) characterizes the behaviour of the covariance function in the Fourier Domain and is defined by

$$\Phi_f(\Delta x) := \mathbb{E}((f(x) - \mathbb{E}(f))(f(x + \Delta x) - \mathbb{E}(f))). \tag{6.10}$$

Kolmogorov's theory is constricted by two quantities, namely l_0 and L_0 called inner and outer scale, respectively. The inner scale represents the size of the smallest eddy in the turbulence, whereas the outer scale corresponds to the largest size. Within this range, Kolmogorv stated that the structure function of the refractive index of the atmosphere at a certain height h is given by

$$D_f(\Delta x) := C_n^2(h)|\Delta x|^{2/3} \text{ for } l_0 < |\Delta x| < L_0. \tag{6.11}$$

The PSD of the refractive index n is then defined as

$$\Phi_n(\kappa) := 0.033 C_n^2(h)|\kappa|^{-11/3} \text{ for } 2\pi L_0^{-1} < |\kappa| < 2\pi l_0^{-1}. \tag{6.12}$$

The behaviour of the PSD due to Kolomogorov's theory is only given within the so called inertial range, defined by the inner and outer scale. For a very small $|\kappa|$, i.e., turbulent eddies larger than the outer scale, Kolmogorov's model shows problems due to the singularity.

6.2.4.2. Von Karman Turbulence Model

The von Karman model [9] modifies the Kolmogorov model in order to overcome the problem with the singularity at $\kappa = 0$. This leads to the following definition of the power spectral density

$$\Phi_f(\kappa) = \frac{0.033 C_n^2(h)}{(|\kappa| + \kappa_0^2)^{11/16}} \exp(-\frac{|\kappa|^2}{\kappa_m^2}), \tag{6.13}$$

with $\kappa_0 = 2\pi L_0^{-1}$ and $\kappa_m = 5.92 l_0^{-1}$.

6.3. Mathematical Description of the Problem

Within the framework of this project, we are dealing with two different mathematical problems. The first one consists of the reconstruction of the incoming, distorted wavefront, either from a Shack-Hartman WFS or a Pyramid WFS. The second problem deals with atmospheric tomography, where we assume a layered model of the atmosphere, with the aim to reconstruct the turbulent layers.

6.3.1. Wavefront Reconstruction

In general, the data s, obtained from a wavefront sensor, is related in a non-linear way to the wavefront ϕ

$$s = Q\phi + \eta, \tag{6.14}$$

where Q denotes the non-linear WFS operator and η models the noise with variance σ_n^2. The reconstruction of the wavefront from sensor data is an inverse problem and, thus, requires regularization techniques to obtain a stable solution. Beside a wavefront sensor, an AO systems requires an active control device, that corrects the wavefront, and a control algorithm, that connects the measurements s with the control commands a_i. The

whole process leads to the wavefront correction

$$\phi_{correct} = Ha = \sum_i a_i h_i, \tag{6.15}$$

where h_i is called influence function. This problem can be formulated as minimization of the corresponding noise-weighted least squares functional

$$J_c(a) = ||QHa - s||^2_{C_\eta^{-1}}. \tag{6.16}$$

In the following subsections two common algorithms for solving the problem of wavefront reconstruction are briefly described. The first is the standard approach in AO, however, has some problems from the computational point of view while dealing with very large telescope as the ELT. The second algorithm was developed by the Austrian Adaptive Optics team (AAO) composed of JKU, RICAM and MathConsult.

6.3.1.1. Matrix-Vector Multiplication

Matrix-vector multiplication (MVM), see, e.g. [10], is the most utilized algorithm in the area of AO. Within the setting of MVM, the WFS is assumed to be linearly related to the DM via an interaction matrix P. This interaction matrix, also called control matrix, maps the DM commands to the sensor measurements. The wavefront reconstruction is then obtained via the inverse of the control matrix. In particular, regularisation is performed using the pseudo inverse $P^\dagger$ as an approximation, since, P is ill-conditioned. The big advantage of this method is that it is highly parallelizable and piplineable. Furthermore, it can handle the modulated as well as the non-modulated pyramid WFS. However, it is extremely time consuming and, thus, not feasible for large scale AO settings, as it is the case within the ELT.

6.3.1.2. Cumulative Reconstructor with Domain Decomposition

To overcome the run-time problem of the MVM an alternative algorithm was developed by the AAO team, called Cumulative Reconstructor (CuRe) or in the enhanced version Cumulative Reconstructor with Domain Decomposition (CuReD) [11]. Instead of $\mathcal{O}(n^2)$, which denotes the computational costs of the MVM algorithm, the CuRe can perform wavefront reconstruction in $\mathcal{O}(n)$ and is still parallelizable and piplineable. Originally, the algorithm was developed for Shack-Hartman WFS, however, using a pre-processing step makes an extension to pyramid WFS possible.

The idea of CuRe is to utilize a modified Hudgin geometry to discretize the domain. In contrast to the Fried geometry, where sensor measurements are obtained using corner points of the subapertures, here, the Shack-Hartman sensor measurements are obtained by subtracting the center of sides

$$s_x[i,j] := \phi[i, j - \frac{1}{2}] - \phi[i-1, j - \frac{1}{2}], \tag{6.17}$$

$$s_y[i,j] := \phi[i - \frac{1}{2}, j] - \phi[i - \frac{1}{2}, j-1]. \tag{6.18}$$

This decouples the measurements in $x-$ and $y-$ direction and, thus, allows the construction of independent chains along one-dimension. The reconstruction of the wavefront is performed from averaged gradients via an integration method, using an iterative scheme

$$l_x[i+1, j - \frac{1}{2}] := l_x[i, j - \frac{1}{2}] + s_x[i+1, j], \tag{6.19}$$

$$l_y[i-\frac{1}{2},j+1] := l_y[i-\frac{1}{2},j] + s_y[i,j+1]. \tag{6.20}$$

The remaining step is to correct the alignment, i.e., to connect the chains, which is done by a shift operation such that the mean value is equal to zero. The construction of the trend chains is based on rectangular domains, however, telescope apertures have usually a circular domain with central obstruction, thus, an adaption is required. Instead of this general trend, the common boundary of the respective subapertures is utilized to compute the alignment of two neighbouring chains.

A drawback of the CuRe is the bad noise propagation for large WFS, as it is the case within the ELT. For that purpose, the CuRe has been extended by a domain decomposition step leading to the CuReD algorithm. The idea is to divide the aperture into smaller domains (domain decomposition) and apply the CuRe on each subdomain. The small parts are then stitched together via their boundaries, to obtain the whole reconstructed wavefront.

As mentioned in the beginning of this section, originally, the CuReD was developed only for Shack-Hartman WFS. To deal with pyramid type WFS as a first step the data preprocessing transforms the pyramid sensor measurements into Shack-Hartman data. In a second step, the CuRed is applied to the preprocessed data. This concept is then called P-CuReD algorithm [12].

For the preprocessing step, a roof approximation of the pyramid WFS is utilized, which decouples the two measurement directions, allowing to perform operation on s_x only row-wise and on s_y only column-wise. The analytical Fourier domain relation between the two sensor data is then used to map the pyramid sensor data into Shack-Hartman like data.

6.3.2. Atmospheric Tomography

Atmospheric Tomography is the fundamental problem in many AO systems, e.g., LTAO, MOAO and MCAO. Assuming a layered model of the atmosphere, the goal of the atmospheric tomography problem is to reconstruct the turbulent layer from the WFS measurements. This turbulent layers are related to the WFS measurements by

$$s_g = \mathcal{T}_g \mathcal{P}_g \phi, \tag{6.21}$$

where $\phi = (\phi_1, \cdots, \phi_L)$ denotes the L turbulent layers a heights $0 \leq h_1 < \cdots < h_L$. The operator $\mathcal{P}_g$ is a geometric propagation operator towards one of the GS $g = 1, \cdots, G$ and $\mathcal{T}_g$ is the WFS operator. The vector s_g contains the WFS measurements for the GS g.

The atmospheric tomography problem is defined by

$$s = A\phi, \tag{6.22}$$

where $A = (A_1, ..., A_G)$ with $A_g := \mathcal{T}_g \mathcal{P}_g$ for $g = 1, ..., G$.

This problem is extremely ill-posed, due to the small angle of separation. Often, it is formulated in the Bayesian framework, since, this allows incorporating statistics of turbulence and noise into the model [13].

Fur that purpose, random variables for the measurements and the turbulent layers in combination with additive noise are used to define (6.22) in a stochastic setting

$$S = A\Phi + \eta. \tag{6.23}$$

The random variables Φ and η are modelled as Gaussians with zero mean and covariance matrices C_Φ and C_η, respectively. The layers are statistically independent, hence, the covariance matrix C_Φ has a block diagonal

structure. The optimal solution to the above described setting is the maximum a-posteriori estimate (MAP)

$$\phi_{MAP} = argmin_{\phi}||C_{\phi}^{-1/2}\phi||^2 + ||C_{\eta}^{-1/2}(s - A\phi)||^2. \tag{6.24}$$

Solving this equation is equivalent to solve the linear system of equations

$$(A^* C_{\eta}^{-1} A + C_{\Phi}^{-1})\phi = A^* C\eta^{-1} s. \tag{6.25}$$

The inverse of C_{η} has a sparse representation, however, the sparsity of C_{Φ}^{-1} and A depends highly on the discretization of the problem.

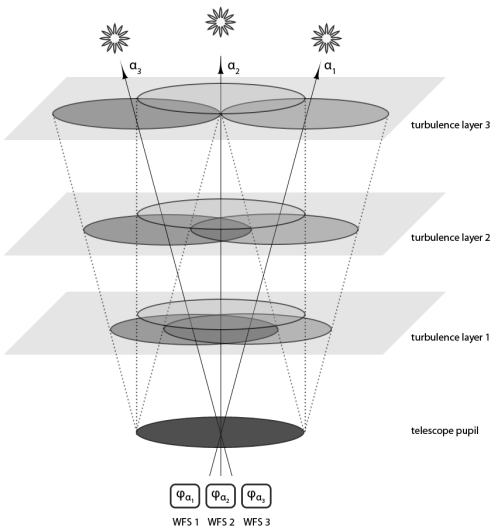


Figure 6.5: Atmospheric Tomography

In the following subsections two algorithms for solving the atmospheric tomography problem are described. The first one is the MVM, which can be generalized in order to perform wavefront reconstruction and atmospheric tomography. The second approach was, as the CuReD, developed by the AAO team.

6.3.2.1. MVM

The MVM method, as described in Section 6.3.1.1, can be generalized in order to perform wavefront reconstruction and atmospheric tomography, see, e.g., [13], by applying one matrix $P^{\dagger}$ to the sensor measurements. Certainly, the run-time problem gets of course more present, since we are dealing here with an even bigger amount of data and, thus, rather large matrices P and $P^{\dagger}$.

Within the Bayesian framework the control matrix is given by

$$P = (A^* C_{\eta}^{-1} A + C_{\Phi}^{-1}) A^* C_{\eta}^{-1}, \tag{6.26}$$

and can be computed explicitly. For every specific geometry, every certain noise and turbulence parameters of the layers the inverse has to be recomputed. P is a full matrix, hence the method scales with $\mathcal{O}(n^2)$ operations.

The MAP estimate is then obtained by a matrix-vector multiplication

$$\phi_{MAP} = Ps. \tag{6.27}$$

As already mentioned in Section 6.3.1.1, due to parallelization and pipelining the method is still efficient, although, for large telescopes as the ELT not feasible.

6.3.2.2. Finite Element Wavelet Hybrid Algorithm

The Finite Element Wavelet Hybrid Algorithm (FEWHA) [14] is an iterative method to compute the MAP estimate. The idea behind is to use compactly supported orthonormal wavelets for representing the turbulent layers.

In the frequency domain, these wavelet representation allows a completely diagonal approximation of the penalty term in (6.24). To achieve a sparse representation for the fitting term, discretization is applied using a piecewise bilinear basis of finite elements.

Turbulence Statistics in the Wavelet Domain
In order to obtain a diagonal approximation of the penalty term, first, we discretize the turbulence layers in the wavelet domain

$$\phi(x,y) = \sum_{j\in\mathbb{Z}}\sum_{k\in\mathbb{Z}}\sum_{t=1,2,3} < \phi, \psi^t_{jk} > \psi^t_{jk}(x,y), \tag{6.28}$$

for $(x,y) \in \mathbb{R}^2$.
Furthermore, we assume that the spectral density of the turbulent field satisfies the von Karman model. Thus, the covariance operator of the layer ϕ at height h can be written as

$$C_\phi = c(h)\mathcal{F}^{-1} m \mathcal{F}, \tag{6.29}$$

where m is the spectral density and $c(h)$ a height dependent constant defined by

$$c(h) = \frac{0.023 r_0^{-5/3}\lambda^2 C_n^2(h)}{4\pi^2}. \tag{6.30}$$

For any $f \in C_0^\infty(\mathbb{R}^2)$ the penalty term of (6.24) can be approximated by

$$||C_\phi^{-1/2} f||^2_{L^2} \simeq \frac{1}{c(h)}(\kappa_0^{11/3}||f||^2_{L_2} + ||(-\Delta)^{11/12} f||^2_{L_2}). \tag{6.31}$$

Since, we are only interested in reconstructing the turbulent domain on a bounded region (region of interest), we consider the periodically extended wavelets on the domain

$$\Omega_\phi = [0, \delta(2^J - 1)^2] - \xi, \tag{6.32}$$

where J denotes the number of wavelet scales in the discretization, $\delta > 0$ is a scaling factor and $\xi \in \mathbb{R}^2$ represents the shift away from the origin.

In the discrete case, the function f in (6.31) is represented by a finite number of wavelet coefficients. Using the Bernstein-Jackson inequalities, this leads to an equivalent representation for the regularizing term by a diagonal matrix D with

$$D_{\lambda\lambda} = \frac{1}{c(h)}(\kappa_0^{11/3} + 2^{(11/3j)}), \tag{6.33}$$

where $\lambda = 0, \ldots, 2^{2J-1}$ is the global wavelet index and $j = 0, \ldots, J-1$ corresponds to the scale index. The described concept is extended to L turbulent layers $\mathbf{\Phi} = (\phi_1, \ldots, \phi_L)$ at heights $0 \le h_1 < \cdots < h_L$ by introducing the square domain Ω_l and the diagonal matrix D_l. For the full problem we define the block-diagonal matrix

$$\mathbf{D} = diag(D_1, \ldots, D_L). \tag{6.34}$$

Finally, the penalty term is approximated by

$$||\mathbf{C_E}^{-1/2}\phi||_{L^2}^2 = \sum_{l=1}^{L} ||C_l^{-1/2}\phi_l||_{L^2}^2 \simeq \sum_{l=1}^{L} (D_l c_l, c_l)_2 = (\mathbf{Dc}, \mathbf{c}). \tag{6.35}$$

Atmospheric Tomography in the Wavelet Domain
To obtain a sparse representation of the fitting term in (6.24) the atmospheric tomography operator A is discretized in the wavelet domain. The computation of the operator can be reduced to an one-dimensional evaluation of one-dimensional wavelet functions.

The operator A has a block-component representation

$$\mathbf{A} = \begin{pmatrix} \Gamma_{11}^{LGS} & \dots & \Gamma_{1L}^{LGS} \\ \vdots & & \vdots \\ \Gamma_{G1}^{LGS} & \dots & \Gamma_{GL}^{LGS} \end{pmatrix}, \tag{6.36}$$

where Γ_{gl}^{LGS} and $\Gamma_{g'l}^{NGS}$ denote the Shack-Hartmann operators in LGS and NGS direction, respectively.

The domains, observed by a WFS in direction $\theta_g = (\theta_g^x, \theta_g^y, 1)$ of a LGS and NGS, respectively, are given by

$$\Omega_{gl}^{LGS} := (1 - \frac{h_l}{H})\Omega + (\theta_g^x, \theta_g^y)h_l \tag{6.37}$$

$$\Omega_{g'l}^{NGS} := \Omega + (\theta_{g'}^x, \theta_{g'}^y)h_l. \tag{6.38}$$

Each turbulent layer ϕ_l, defined by (6.32), is observed by all WFSs, thus, the following condition must hold,

$$(\bigcup_{g=1}^{G_{LGS}} \Omega_{gl}^{LGS}) \cup (\bigcup_{g'=G_{LGS+1}}^{G} \Omega_{g'l}^{NGS}) \subseteq \Omega_l. \tag{6.39}$$

The two Shack-Hartman operators Γ_{gl}^{LGS} and $\Gamma_{g'l}^{NGS}$ are defined on the domains Ω_{gl}^{LGS} and $\Omega_{g'l}^{NGS}$, respectively. The LGS operator consists of a concatenation of two components, associated to the $x-$ and $y-$ direction, respectively, where each component has a matrix representation. Using the fact that the scaling and wavelet functions have a multiplicative representation, the assembly of the matrices Γ_{gl}^{LGS} and $\Gamma_{g'l}^{NGS}$ can be reduced to a one-dimensional evaluation of one dimensional scaling and wavelet functions over the edges of the projected subapertures.

Finally, this concepts allows to define the discretization of equation (6.25) in the wavelet domain

$$(\widetilde{A}^T C_\eta^{-1} + \widetilde{A} + \alpha D)c = \widetilde{A}^T C_\eta^{-1} s, \tag{6.40}$$

where $\widetilde{A}$ is the atmospheric tomography operator in the wavelet domain.

6.4. Quality Evaluation

In order to evaluate the benchmark cases, i.e. the quality performance of an AO system, a variety of measures are available. In the following subsections we briefly describe two of them. However, in further deliveries we will focus on the Strehl ratio for evaluating the quality in the numerical examples.

6.4.1. Strehl Ratio

The Strehl ratio is the rate between the real energy distribution $I(x, y)$ and the hypothetical distribution obtained in diffraction-limited imaging and is defined by

$$S := \frac{max_{(x,y)} I(x, y)}{max_{((x,y))} I_D(x, y)}. \tag{6.41}$$

This formula leads to a quantity between 0 and 1, that increases with quality and reaches its maximum value 1 in diffraction-limited case. It is common to state the Strehl ratio in $100 \cdot S\%$.

Marechal Criterion
If the system fulfils the Marechal criterion

$$S \geq 0.8, \tag{6.42}$$

it is called well-corrected [2]. In this case the Strehl ration can be approximated by

$$S \approx \exp(-\int \Delta\overline{\varphi}^2 d\mathbf{x}). \tag{6.43}$$

This expression allows a much faster evaluation, since the Fourier transform is avoided. The problem, however, is that in practice it is often not fulfilled. Nevertheless, the approximation can be used to compare different reconstruction methods, which is the case when comparing various benchmarks. Although, it might not be a good approximation to S, it is still known that the reconstruction quality is related monotonically to it.

6.4.2. Full Width at Half Maximum

Another common quality measure is the Full Width at Half Maximum (FWHM), which is deduced from the PSF. It is defined as the width of the PSF at the point where the intensity is one half of the maximum intensity. For the optical density I it is defined by

$$FWHM(I) := diam\{\mathbf{x} \in \mathbb{R}^2 | I(\mathbf{x}) \geq \frac{1}{2} max|I|\}. \tag{6.44}$$

6.5. Description of Data

To evaluate the quality of the algorithms, they are tested using numerical simulations. For that purpose, the official end-to-end simulation tool established by the ESO, called OCTOPUS, is used. OCTOPUS [15] simulates a complete telescope and all its components, as well as the perturbation of light travelling through the turbulent atmosphere. The simulation of the atmosphere is based on the assumption of a layered structure of perturbations.

The big advantage of OCTOPUS is that external reconstruction algorithms can be implemented in any language and started as a separate process parallel to OCTOPUS. The transfer of data between the two processes is managed via files, that have a specified syntax. In a first step, OCTOPUS writes the sensor measurements to a file and, furthermore, a signal to be finished. Then the external reconstruction algorithm, waiting for this signal, reads the sensor measurements and performs wavefront reconstruction or atmospheric tomography. The result of the algorithm are actuator commands in order to adapt the DM, which again are written to a file. The algorithm finishes, by writing a signal for which OCTOPUS is waiting. OCTOPUS then updates the DM shape and provides the new sensor measurements. This loop is continued until a predefined number of time steps.

Bibliography

[1] Microgate, "Description of industrial partner," Online Description. [Online]. Available: http://www.microgate.it

[2] F. Roddier, "Adaptive optics in astronomy," Cambridge, U.K. ; New York :Cambridge University Press, 1999.

[3] A. Tokovinin, "Adaptive optics tutorial at ctio," Cambridge University Press, 1999.

[4] J. Primot, "Theoretical description of shack–hartmann wave-front sensor," Optics Communications 222(1):81–92, 2003.

[5] C. Verinaud, "On the nature of the measurements provided by a pyramid wave-front sensor," Optics Communications 233:27–38, 2004.

[6] E. Vernet, M. Cayrel, N. Hubin, R. Biasi, G. Angerer, M. Andrighettoni, D. Pescoller, D. Gallieni, M. Tintori, M. Mantegazza, and et. al., "The adaptive mirror for the e-elt," In Proceedings of the Third AO4ELT Conference, 2013.

[7] E. S. Observatory, "Eso," Online Description. [Online]. Available: http://www.eso.org

[8] A. N. Kolmogorov, "The local structure of turbulence in incompressible viscous fluid for very large reynolds numbers," In Dokl. Akad. Nauk SSSR volume 30, pages 299–303, 1941.

[9] T. von Karman, "Mechanische Ähnlichkeit und turbulenz," rd International Congress of Applied Mechanics, 1930.

[10] N. F. Law and R. G. Lane, "Wavefront estimation at low light levels," Optics Communications 126:19–24, 1996.

[11] M. Rosensteiner, "Wavefront reconstruction for extremely large telescopes via cure with domain decomposition," J. Opt. Soc. Am. A 29(11):2328–2336, 2012.

[12] I. Shatokhina, A. Obereder, M. Rosensteiner, and R. Ramlau, "Preprocessed cumulative reconstructor with domain decomposition: a fast wavefront reconstruction method for pyramid wavefront sensor," Applied Optics 52(12):2640–2652, 2013.

[13] T. Fusco, J.-M. Conan, G. Rousset, L. Mugnier, and V. Michau, "Optimal wavefront reconstruction strategies for multi conjugate adaptive optics," J. Opt. Soc. Am. A 18(10):2527–2538, 2001.

[14] T. Helin and M. Yudytskiy, "Wavelet methods in multi-conjugate adaptive optics," Inverse Problems 29(8):085003, 2013.

[15] ESO, "Online description of octopus," Online description. [Online]. Available: http://www.eso.org/sci/facilities/develop/ao/tecno/octopus.html

7. Sinkhorn algorithm for point source far field FreeForm Optics problem

Jean-David Benamou[1], Wilbert Ijzerman[2], Giorgi Rukhaia[1]

[1] *Institut National de Recherche en Informatique et en Automatique*

[2] *Signify*

Abstract. FreeForm Optics is the branch of Optics concerned with the design of non-conventional asymmetric refractive and reflective optical elements or systems of such elements. This research is important to improve the energy efficiency of lighting devices and reduce light pollution (for example of street lighting). A classic application of FreeForm Optics (amongst many) is the irradiance tailoring problem: design an optical system transferring a given light source emittance (e.g a car headlight bulb) to a prescribed irregular target irradiance (e.g. the angular far field distribution of projected light). At the industrial level, FreeForm Optics design has remained so far largely heuristic.
On the academic side, two classes (collimated or point source illuminance) of idealized tailoring irradiance problems can be exactly modeled and solved using Optimal Transport theory. Optimal Transport defines a unique map or a coupling between prescribed distributions representing given illuminance and irradiance. This map can then be used to construct the optical element shape. Recent advances in Optimal Transport numerical solvers allow tackling systems described by millions of degrees of freedom. This offers a sound mathematical and numerical background to FreeForm Optics.

Keywords: Reflector problem, FreeForm Optics, optimal transport, entropic regularization, Sinkhorn algorithm.

7.1. Introduction

A light source, also called "illuminance", is sufficiently small compared to the reflecting surface so that it can be regarded as a point in space. It can therefore be modelled as a probability distribution on the sphere, it will be denoted μ in this paper. The light hits a perfect mirror and we are also given a desired target light distribution, the "illumination" in the far field. From the far field the reflecting surface can be regarded as a point and the illumination again modelled as a probability distribution, denoted ν, on the sphere. Total light conservation is assumed. The reflector problem is to determine the shape of the mirror which produces the specular reflection from the source to the target distribution. This can be interpreted as the inverse problem of generating some illumination given an illuminance and a reflector (see figure 7.1).

7.1.1. Optimal Transport model

This problem has an elegant mathematical modelization and solution based on the optimal transportation (OT) theory due to [1] and [2]. We briefly recall the main result as presented in [2]. In its Kantorovich primal and dual form (see [3]) :

Theorem 2 (Kantorovich duality)**.** *Given two compact manifold X and Y endowed with a continuous, bounded from below cost function $c : X \times X \to \mathbf{R}$ and two borel probability measures $(\mu, \nu) \in \mathcal{P}(X) \times \mathcal{P}(Y)$. Then, Kantorovich problem in primal and dual forms (7.1) has solutions.*

$$OT(\mu,\nu) := \min_{\gamma \in \Pi(\mu,\nu)} \langle c, \gamma \rangle_{X\times Y} = \max_{f,g \in C} \langle f, \mu \rangle_X + \langle g, \nu \rangle_Y \tag{7.1}$$

htdp

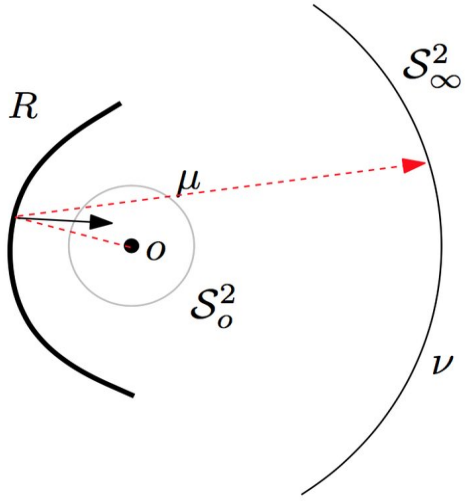


Figure 7.1: Reflector problem from Point source O to Far Field.

with respectively primal :

$$\Pi(\mu,\nu) := \{\gamma \in \mathcal{P}(X \times Y), \langle 1_X, \gamma\rangle_Y = \nu \ \langle 1_Y, \gamma\rangle_X = \mu\},$$

and dual :

$$C = \{(f,g) \in \mathcal{C}(X) \times \mathcal{C}(Y),\ f \oplus g \leq c\},$$

constraints sets

The notation $\langle f, \alpha\rangle_\Omega$ stands for the duality product $\int_\Omega f\, d\alpha$ between bounded continuous functions $f \in \mathcal{C}(\Omega)$ and probability measures $\alpha \in \mathcal{P}(\Omega)$, $\{f \oplus g\}(x,y) = f(x) + g(y)$ is the direct sum and $\mu \otimes \nu \in \mathcal{P}(X \times Y)$ the tensor product. Finally 1_Ω is the characteristic function, i.e. a constant 1 on Ω.

Under suitable hypothesis on c (as they are technical and satisfied for the costs in this paper, we skip this part), the OT problem is well posed and the the optimal transference plan γ is concentrated on a graph of the OT map $y = T(x)$ implicitely defined by the saturation of the dual constraint :

$$f(x) + g(T(x)) = c(x, T(x)),\ \mu\, a.e. \tag{7.2}$$

The pair (f,g) are called the Kantorovich potentials and is unique up to an additive constant .

By construction T is a measure preserving map characterizing the transport. The measure preserving property is usually denoted $\nu = T\#\mu$ (T pushes forward μ to ν). The pushforward of μ is the measure defined as

$$\nu(A) = T_\#\mu(A) = \mu(T^{-1}(A)) \text{ for all } \nu \text{ measurable subset } A \tag{7.3}$$

Remark 2 (L^p Wasserstein metric)**.** *For complete separable metric space X and L^p costs $c := 1/pd^p(x,y)$, this OT problems defines a separable metric on the set of probability measures with finite second moments: the "Wasserstein" distance, which is given by $W_p^p(\mu,\nu) := OT(\mu,\nu)$.*
This metric metrizes weak convergence of measures is a fundamential tool in image processing (see [4]).

In [2], Wang showed that the point source reflector model can be translated to an OT problem. More precisely,

he proved the following theorem :

Theorem 3. *Let $S_0 \in \mathbb{S}^{d-1}$ and $S_\infty \in \mathbb{S}^{d-1}$ be connected domains in northern and southern hemispheres respectively, μ and ν which represent the given illuminance and illumination probability distributions. Then theorem 2 applies to the cost function*

$$c(x, y) = -\log(1 - x \cdot y). \tag{7.4}$$

A transport map T satisfying (7.2) exists and the solution of the corresponding OT problem can be used to build the desired reflector.

The construction of the reflector can be summarized as follows : Taking the exponential of the dual constraints and the saturation property (7.2) we get

$$\frac{e^{-g(T(x))}}{1 - x \cdot T(x)} = e^{f(x)} \leq \frac{e^{-g(y)}}{1 - x \cdot y}, \ \mu \otimes \nu \ a.e. \tag{7.5}$$

We now define in $\mathbb{R}^d$ a family of parabolic reflectors with axis $y \in S_\infty : x \in S_0 \to P_y(x) := \dfrac{e^{-g(y)}}{1 - x \cdot y}$. And directly infer that the reflector shape parameterized over the directions in S_0 and given as :

$$R = \{x e^{f(x)} | x \in S_0\}. \tag{7.6}$$

Under this choice the map $x \to T(x)$ can be interpreted as the specular reflection of an optical ray at $R(x)$ onto a parabola of axis $T(x)$ while the illumination and illuminance constraints are enforced by (7.3).

7.1.2. Entropic Regularization of Optimal Transport

Entropic regularization has been introduced for OT computations in [5] (see [4] for a comprehensive review). The entropic regularization of the Kantorovich problem (2) is based on the following KullBack-Leibler divergence or "relative entropy" (KL) penalization :

$$\begin{aligned} OT_\epsilon(\mu, \nu) := \quad & \min_{\gamma_\epsilon \in \Pi(\mu,\nu)} \langle c, \gamma_\epsilon \rangle_{X \times Y} + \epsilon \ \mathrm{KL}(\gamma_\epsilon \,|\, \mu \otimes \nu) = \\ & \max_{f_\epsilon, g_\epsilon} \langle f_\epsilon, \mu \rangle_X + \langle g_\epsilon, \nu \rangle_Y - \epsilon \left\langle \exp(\tfrac{1}{\epsilon}(f_\epsilon \oplus g_\epsilon - c)) - 1, \mu \otimes \nu \right\rangle_{X \times Y} \end{aligned} \tag{7.7}$$

where $\epsilon > 0$ is a small "temperature" parameter (see [6] for a Statistical Physics interpretation of this problem due to Schroedinger) and

$$\mathrm{KL}(\gamma \,|\, \mu \otimes \nu) := \int_{X \times Y} \log\left(\frac{d\gamma}{d\mu \otimes d\nu}\right) d\gamma \text{ if } \gamma \text{ is absolutely continuous w.r.t to } \mu \otimes \nu \text{ and } +\infty \text{ else.}$$

The primal-dual optimality condition is given by

$$\gamma_\epsilon = \exp(\frac{1}{\epsilon}(f_\epsilon \oplus g_\epsilon - c))\, \mu \otimes \nu. \tag{7.8}$$

The optimal entropic plan is therefore the scaling by the Kantorovich potentials of a fixed Kernel $\exp(-\frac{1}{\epsilon}c)$.

Remark 3. *Of course, altering the desired target functional (2) results in altered solution and thereof γ_ϵ is not the exact transport plan that we are looking for. It is diffuse, i.e. not concentrated on a map, and ϵ can be interpreted as a bandwidth under which the transport is blurred.*

Although, this entropic plan γ_ϵ converges to γ, the minimizer of (2), when ϵ goes to 0.(see [4])

7.1.3. Sinkhorn Algorithm for Regularized Optimal Transport

Numerical solutions are produced using the discretization of this problem, i.e. replacing (X, Y, c, μ, ν) by $(X_N, Y_N, c_N, \mu_N \nu_N)$ in the following way:

$$\mu_N = \sum_{i=1}^{N} p_i \delta_{x_i}, \quad \nu_N = \sum_{j=1}^{N} q_j \delta_{y_j}, \quad where \quad \sum_{i=1}^{N} p_i = \sum_{j=1}^{N} q_j = 1. \tag{7.9}$$

Of course the number of discrete points for μ and ν may differ, we keep N for both to simplify the presentation. This discretisation provides a natural discretization of the OT problem (2). Settting $X_N = \{x_i\}_{i=1..N}$, $Y_N = \{y_j\}_{j=1..N}$, $c_N = \{c(x_i, y_j)\}_{i,j=1..N}$, $q = \{q_j\}_{j=1..N}$ and $p = \{p_i\}_{i=1..N}$. we can again use the $\langle ., . \rangle$. notation :

$$OT_N(p, q) := \min_{\gamma_N \in \Pi(p,q)} \langle c_N, \gamma_N \rangle_{X_N \otimes Y_N} \tag{7.10}$$

where

$$\Pi(p, q) := \left\{ \gamma_N \in \mathbb{R}_+^{N \times N} | \langle 1_{X_N}, \gamma_N \rangle_{Y_N} = p, \ \langle 1_{Y_N}, \gamma_N \rangle_{X_N} = q \right\} \tag{7.11}$$

Similarly, discretization of regularized problem (7.7) gives

$$OT_{\epsilon,N} := \max_{f_\epsilon, g_\epsilon} \langle f_\epsilon, \mu_N \rangle_{X_N} + \langle g_\epsilon, \nu_N \rangle_{Y_N} - \epsilon \left\langle \exp(\frac{1}{\epsilon}(f_\epsilon \oplus g_\epsilon - c_N)) - 1, \mu_N \otimes \nu_N \right\rangle_{X_N \times Y_N}. \tag{7.12}$$

where we use the same notation (f_ϵ, g_ϵ) for discrete vectors in $\mathbb{R}^N$.

We solve (7.12) with Sinkhorn algorithm. It corresponds to a block coordinate (f_ϵ and g_ϵ) ascent : Initialize with $g_\epsilon^0 = 0_Y$ and then iterate (in k) :

$$\begin{aligned} f_\epsilon^{k+1} &= -\epsilon \log(\langle \exp(\tfrac{1}{\epsilon}(g_\epsilon^k - c_N)), \nu_N \rangle_{Y_N}) \\ g_\epsilon^{k+1} &= -\epsilon \log(\langle \exp(\tfrac{1}{\epsilon}(f_\epsilon^{k+1} - c_N)), \mu_N \rangle_{X_N}) \end{aligned} \tag{7.13}$$

As discussed in [4](Remark 4.13), for sufficiently regular data (for example when exact map T is guaranteed to be smooth) following estimate holds for sufficiently large number of iterations k in (7.13):

$$\sup_{X_N} |f_\epsilon(x) - f_\epsilon^k(x)| = O(1 - \epsilon)^k \tag{7.14}$$

Where f_ϵ is an exact regularized potential of (7.12).

7.2. Hierarchical approach to Sinkhorn Algorithm

7.2.1. ϵ scaling

As mentioned in remark (3), decreasing ϵ would result in a more accurate solution for (2). On the other hand, estimate (7.14) suggests that smaller ϵ we take, higher number of iterations will be required for Sinkhorn algorithm to converge. Also, taking ϵ too small, would result into numerical overflows due to the exponential terms of order $e^{\frac{1}{\epsilon}}$ in (7.13)

As discussed in [7], problem of numerical stability can be tackled by working with the increments of the potentials rather then full potentials during the iterative steps.

That is, if we look at the updates f_ϵ^{k+1} and g_ϵ^{k+1} in (7.13) as $f_\epsilon^{k+1} = f_\epsilon^k + \hat{f}_\epsilon^{k+1}$ and $g_\epsilon^{k+1} = g_\epsilon^k + \hat{g}_\epsilon^{k+1}$, then by moving previous approximations to the right hand side, we will get the following new iterative scheme for the increments:

$$\begin{aligned}
\hat{f}_\epsilon^{k+1} &= -\epsilon \log(\langle \exp(\tfrac{1}{\epsilon}(g_\epsilon^k + f_\epsilon^k - c_N)), \nu_N \rangle_{Y_N}) \\
f_\epsilon^{k+1} &= f_\epsilon^k + \hat{f}_\epsilon^{k+1} \\
\hat{g}_\epsilon^{k+1} &= -\epsilon \log(\langle \exp(\tfrac{1}{\epsilon}(f_\epsilon^{k+1} + g_\epsilon^k - c_N)), \mu_N \rangle_{X_N}) \\
g_\epsilon^{k+1} &= g_\epsilon^k + \hat{g}_\epsilon^{k+1}
\end{aligned} \tag{7.15}$$

Those iterations will be more stable due to the saturation property of the optimizing potentials (7.2). This property tells us that quantity $f(x_i) + g_(y_j) - c(x_i, y_j)$ is zero for exact potentials and optimal pairs (x_i, y_j) while being strictly negative for non-optimal pairs. Thereof, when the iterates f_ϵ^k and g_ϵ^k are close to the true potentials, new updating steps would not cause a numerical overflow.

Although, this approach alone would not help at the first steps of the algorithm, since we have no guarantees that initial approximations would be close to the exact potentials, and for small ϵ we would get an overflow at the first step of the iterations. In order to avoid this, possible approach would be to start with higher values of ϵ and gradually decrease it to the desired final value ϵ_{final} (see [7] [8]).

More formally, one can define a sequence of regularization parameters $\epsilon_k \to \epsilon_{final}$ and use ϵ_k at k-th iteration in (7.15). Common choice is to choose starting ϵ_0(usually taken to be 1), chose scaling parameter $\lambda \in (0, 1)$ and define $\epsilon_k := \max\{\epsilon_{final}, \lambda^k \epsilon_0\}$.

Remark 4. *It has been empirially established (see [7] and references therein), that above discussed approach of gradually increasing ϵ_k at each iteration, not only provides more numerically stable scheme, but also increases the convergence speed. In other words, less number of iterations is required for achieving a given error threshold with decreasing ϵ_k at each iteration, then while using fixed ϵ_{final} for all iterations.*

7.2.2. Discretization scaling

In [7] (see also [9]), it is discussed that the entropic regularization with ϵ acts as a smoothing filter on the data, which smoothers out any details that are on the finer scale then ϵ. This means that doing Sinkhorn iterations with discretization such that $\min_{i,j} d(x_i, x_j) << \epsilon$ does not provide valuable improvement over working with discretizations that are on the scale of ϵ.

Thereof, it would be more efficient to also use a sequence of discretizations $(X_{N_k}, Y_{N_k}, c_{N_k}, \mu_{N_k}, \nu_{N_k})$ where $N_k = O(\frac{1}{\epsilon_k})^d$ (where d is the dimension of the problem). In order to implement this approach, one would need to find a way to interpolate approximations f_ϵ^k, g_ϵ^k on the discretization $X_{N_{k+1}}, Y_{N_{k+1}}$ respectively, while they are computed on the grids X_{N_k}, Y_{N_k}.

Luckily, Sinkhorn algorithm provides a canonical way of computing such interpolations, even for the full spaces X and Y. If we expand the definition of scalar product in (7.13) and replace $c_N = c_N(x_i, y_j)$ by $c(x, y_j)$ and $c(x_i, y)$ respectively, we obtain following continuous extensions for given approximations $f_{\epsilon_k}^k$ and $g_{\epsilon_k}^k$:

$$\tilde{f}_{\epsilon_k}^k(x) := -\epsilon_k \log\Big(\sum_{j=1..N_k} \exp\big(\frac{1}{\epsilon_k}(g_{\epsilon_k}^k(y_j) - c(x, y_j))\big)\nu_{N_k}(y_j)\Big), \ \forall x \in X. \tag{7.16}$$

$$\tilde{g}_{\epsilon_k}^k(y) := -\epsilon_k \log\Big(\sum_{i=1..N_k} \exp\big(\frac{1}{\epsilon_k}(f_{\epsilon_k}^k(x_i) - c(x_i, y))\big)\mu_{N_k}(x_i)\Big), \ \forall y \in Y. \tag{7.17}$$

Thereof, at k-th iteration, we can take $k-1$-th approximations to be restrictions of $\tilde{f}_\epsilon^{k-1}(x)$ and $\tilde{f}_\epsilon^{k-1}(x)$ on the spaces X_{N_k} and Y_{N_k} respectively.

Putting it all together, we obtain the following iterative procedure in k:

$$
\begin{aligned}
f_{\epsilon_k}^{k-1} &= \tilde{f}_{\epsilon_{k-1}}^{k-1}|_{X_{N_k}} \qquad g_{\epsilon_k}^{k-1} = \tilde{g}_{\epsilon_{k-1}}^{k-1}|_{Y_{N_k}} \\
\hat{f}_{\epsilon_k}^{k} &= -\epsilon_k \log(\langle \exp(\tfrac{1}{\epsilon_k}(g_{\epsilon_{k-1}}^{k+1} + f_{\epsilon_{k-1}}^{k-1} - c_{N_k})), \nu_{N_k} \rangle_{Y_{N_k}}) \\
f_{\epsilon_k}^{k} &= f_{\epsilon_k}^{k-1} + \hat{f}_{\epsilon_k}^{k} \\
\hat{g}_{\epsilon_k}^{k} &= -\epsilon_k \log(\langle \exp(\tfrac{1}{\epsilon_k}(f_{\epsilon_k}^{k} + g_{\epsilon_k}^{k-1} - c_{N_k})), \mu_{N_k} \rangle_{X_{N_k}}) \\
g_{\epsilon_k}^{k} &= g_{\epsilon_k}^{k-1} + \hat{g}_{\epsilon_k}^{k}
\end{aligned}
\tag{7.18}
$$

In this setting, taking ϵ_{final} to go to 0 means also refining discretization. Thereof, estimate (7.14) would not be applicable. To the best of our knowledge the joint convergence in N and ϵ has only be studied in [9]:

Theorem 4 (Berman joint convergence - corollary 1.3 [9])**.** *We assume μ and ν are in $\mathcal{C}^{2,\alpha}$ and positive, and that N and ϵ are dependent parameters : $N = (1/\epsilon)^d$ where d is the dimension of the problem. A technical condition on the sequence of discretization $(X_N, Y_N, c_N, \mu_N \nu_N)$ called "density propery" (see remark 6 below) is also necessary. Then there exists a positive constant A_0 such that for any $A > A_0$ the folowing holds : setting $m_\epsilon = [-A\log(\epsilon)/\epsilon]$ the continuous interpolation provided by $\tilde{f}_\epsilon^{m_\epsilon}$, built using the cannonical extension (7.16) from the discrete Sinkhorn iterate at $k = m_\epsilon$, satisfies the estimate*

$$
\sup_X |\tilde{f}_\epsilon^{m_\epsilon} - f| \leq -C\epsilon\log(\epsilon) \tag{7.19}
$$

for some constant C (depending on A) and f an optimal potential for (2).

Remark 5. *Assumptions of Theorem 4 holds on the sphere for the reflector cost (see section 6.3.3 [9]). Although, while estimating the necessary number of iterations m_ϵ, this theorem does not take into account the improved effect on the convergence, coming from the ϵ-scaling.*

Remark 6 (Density property Lemma 3.1 [9])**.** *For any given open set U intersecting the support X of μ (same for Y and ν)*

$$
\liminf_{\epsilon \to 0} \epsilon \log(\mu_N(U)) = 0
$$

For the flat space $X \subset \mathbb{R}^d$, this condition is enough. For curved surfaces, a technical generalization is required. But in both cases, this density property ensures the discretization of X and μ (7.9) is such that, for U the sequence of approximations $\mu_N(U)$ never converges faster to 0 than ϵ (remember that $N = (1/\epsilon)^d$).

For the sphere this can be achieved by either the Quasi Monte-Carlo discretizations that are sampled uniformly with respect to the surface element of the sphere (see [10]), or by adjusting the weights of the discretization points according to the deviation from the surface element (e.g. for the orthogonal grids projected from a plane to the sphere).

7.3. Benchmark cases

We propose 2 Benchmark cases for computing optimal shape, using Sinkhorn algorithm incorporating above-mentioned modifications:

- Computing a reflector for the problem where the Source is a uniform distribution with support on a set, which is the inverse stereographic projection of unit square centered at the origin and the desired light distribution is a uniform distribution with support on a set, which is an inverse stereographic projection of circle centered at the origin and Diameter 1.
- Computing a reflector for the problem where the Source is a uniform distribution with support on a set, which is an inverse stereographic projection of unit square centered at the origin and the desired light distribution is a sum of two gauss distributions which are the centered respectively at inverse stereographic projection of points (0.25,-0.25),(0.25,0.25).

Bibliography

[1] T. Glimm and V. Oliker, "Optical design of single reflector systems and the monge–kantorovich mass transfer problem," *Journal of Mathematical Sciences*, vol. 117, pp. 4096–4108, 09 2003.

[2] X.-J. Wang, "On the design of a reflector antenna ii," *Calculus of Variations and Partial Differential Equations*, vol. 20, no. 3, pp. 329–341, Jul 2004. [Online]. Available: https://doi.org/10.1007/s00526-003-0239-4

[3] C. Villani, *Optimal Transport: Old and New*, ser. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2008. [Online]. Available: https://books.google.fr/books?id=NZXiNAEACAAJ

[4] G. Peyré and M. Cuturi, "Computational Optimal Transport," *ArXiv e-prints*, Mar. 2018.

[5] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," in *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2013, pp. 2292–2300. [Online]. Available: http://papers.nips.cc/paper/4927-sinkhorn-distances-lightspeed-computation-of-optimal-transport.pdf

[6] C. Léonard, "A survey of the schrödinger problem and some of its connections with optimal transport," 2013.

[7] B. Schmitzer, "Stabilized Sparse Scaling Algorithms for Entropy Regularized Transport Problems," *arXiv e-prints*, p. arXiv:1610.06519, Oct 2016.

[8] A. M. Oberman and Y. Ruan, "An efficient linear programming method for optimal transportation," 2015.

[9] R. J. Berman, "The Sinkhorn algorithm, parabolic optimal transport and geometric Monge-Ampere equations," *ArXiv e-prints*, Dec. 2017.

[10] R. S. Womersley, "Efficient Spherical Designs with Good Geometric Properties," *ArXiv e-prints*, Sep. 2017.

8. Data driven model adaptations of coil sensitivities in magnetic particle imaging

Lena Hauberg-Lotte[1], Peter Maass[1]

[1] *University of Bremen*

Abstract. In the first chapter we introduce the basic ideas of magnetic particle imaging (MPI), inverse problems and Deep Learning. In the second chapter we focus on the details of the modeling of the forward operator of MPI, specially we introduce the equilibrium model and outline more complex models. After that we discuss different scenarios of model and data accuracies to lay the groundwork for the following discussion of which Deep Learning approaches to inverse problems could be used for which scenarios. We make the five broad categorizations of Deep Learning approaches to inverse problems: 1. Learned Penalty Terms, 2. Plug-and-Play Prior Methods, 3. Gradient-Descent-by-Gradient-Descent type Methods, 4. Regularization by Architecture (Deep Prior), 5. Image Post-Processing via Deep Learning.

Keywords: Deep learning, neural networks, inverse problems.

8.1. Introduction and literature

8.1.1. Magnetic Particle Imaging

Magnetic particle imaging (MPI) is a relatively new non-invasive tomographic imaging technique that directly detects superparamagnetic iron oxide nanoparticles (SPIO). It combines high tracer sensitivity with submillimetre resolution and imaging is performed in milliseconds to seconds.

MPI is suitable for several medical applications: The nonlinear magnetization behaviour of nanoparticles in an applied magnetic field is employed to reconstruct a spatial distribution of the concentration of nanoparticles in the cardiovascular system. A high temporal resolution and a potentially high spatial resolution make MPI suitable for several *in-vivo* applications without the need for harmful radiation. The potential for imaging blood flow was demonstrated first in *in-vivo* experiments using a healthy mouse [1]. The usability of a circulating tracer for long-term monitoring was recently investigated [2]. The high temporal resolution of MPI is advantageous for potential flow estimation [3] and for tracking medical instruments [4]. Recently, MPI was also shown to be suitable for tracking and guiding instruments for angioplasty [5]. Further promising applications of MPI include cancer detection [6] and cancer treatment by hyperthermia [7].

The main goal of MPI is to reconstruct the spatially dependent concentration of particles and for that computationally efficient reconstruction methods are required to allow real time observations. Therefore mathematical models are advantageous for the development of such new methods.

8.1.2. Deep Learning and Inverse Problems

Inverse Problems have been a key tool in many areas of science, technology in general, and more particularly in the field of medical imaging for many years now. The key idea is to calculate causes from effects, opposed to so-called direct problems trying to predict effects from causes.

Deep Learning (DL) on the other hand is a relatively new field which studies big machine learning models. It is not clear what all the strength of DL are, but a major one is the ability to predict labels from data via

supervised learning, e.g., a label of a computer tomographic (CT) image could be cancer or no cancer. A major problem with this is the fact that one usually needs massive amounts of labeled data to train a learning model.

The success of neural networks (NN) in many computer vision tasks in the past has motivated attempts at using Deep Learning to achieve better performance in solving Inverse Problems [8]. It was proposed to apply data-driven approaches to inverse problems, using neural networks as a regularization functional. The network learns to discriminate between the distribution of ground truth images and the distribution of unregularized reconstructions and the approach was used for computer tomography reconstruction [9]. Furthermore, deep learning was used in medical applications such as tumour classification with MALDI Imaging [10], magnetic resonance imaging (MRI) [11] [12], low-dose X-ray CT [13] as well as positron emission tomography (PET) [14].

The need for massive amounts of data is also a major challenge in bringing Deep Learning and Inverse Problems together, since it creates a chicken-and-egg-problem. The problem lies therein that one can think about the ”cause” in Inverse Problems as (or at least connected to) the label in Deep Learning. This means that one usually doesn't have labels for Inverse Problems which are required to train a deep model to solve the Inverse Problem. Most approaches that try to bring Deep Learning to Inverse Problems ignore this fact and only focus on problems where there is enough ground truth data for the cause already – due to some special circumstances. In fact they can simplify the practical application of already solved Inverse Problems massively, but they can usually not be applied to novel unsolved Inverse Problems. In summary: there exists some kind of ”information gap” that creates a boot strap problem. We are planning on exploring ways to solve this problem, in particular forMPI.

8.2. Mathematical description

8.2.1. Magnetic Particle Imaging

To determine the distribution of nanoparticles, which is the quantity $c(x)$, the nonlinear magnetization behavior of ferromagnetic nanoparticles is exploited as follows, see also [15]: A static magnetic field (selection field), which is given by a gradient field, generates a field free point (FFP) (or alternatively a field free line (FFL) [16]). The larger the distance between nanoparticles and FFP, the more is the magnetization caused by the nanoparticles in saturation. The superposition with a spatially homogeneous but time-dependent field (drive field) moves the field free region along a predefined trajectory defining the field-of-view (FOV). An interplay between gradient strength and drive field amplitude determines the FOV size but when guaranteeing a certain resolution the FOV is strictly limited due to safety reasons. The rapid change of the applied field $\mathbf{H}(x,t)$ causes a measurable change of the magnetization $\mathbf{M}(x,t)$ of the nanoparticles.

In the first approximation the change of the magnetization can be characterized by using the Langevin function. Neglecting the interactions between multiple particles and doing the transition from microscopic to macroscopic scale (see [17]) allows the approximation of the magnetization $\mathbf{M}$ by multiplying the particles' mean magnetic moment vector $\bar{\mathbf{m}}(x,t)$ and the particle concentration $c(x)$.

The temporal change of the particles' magnetization induces a voltage $u^P(t)$ in the receive coil units. Using a quasi-static approximation in the induction principle and the law of reciprocity (see [18]) allows for the description via a linear integral operator with respect to the particle concentration:

$$u^P(t) = -\mu_0 \int_\Omega \mathbf{p}^R(x) \cdot \frac{\partial}{\partial t}\mathbf{M}(x,t)\, dx = \int_\Omega c(x) \underbrace{(-\mu_0 \mathbf{p}^R(x) \cdot \frac{\partial}{\partial t}\bar{\mathbf{m}}(x,t))}_{=s(x,t)}\, dx. \tag{8.1}$$

Here, $\mathbf{p}^R$ is the receive coil sensitivity, which is the magnetic field which is generated by the receive coil unit

when applying a unit current. Analogously, the applied field $\mathbf{H}(x,t)$ also induces a voltage $u^E(t)$ which is known as direct feedthrough. Since this value is several orders of magnitude larger than that of the particle signal, it must be removed prior to digitization. This is done by applying an analog filter which is described by a temporal convolution with a kernel function $a(t)$ ($\tilde{a}$ denotes its periodic continuation). The measured signal $v(t)$ is then given by $v = (u^P + u^E) * a$. One common choice for the analog filter is a band-stop filter such that some frequency bands of the particle signal are also removed. The resulting integral kernel $\tilde{s}(x,t) = (s(x,\cdot) * a)(t)$ is primarily determined by the analog filter (receive-unit-dependent), the receive coil sensitivity (receive-unit-dependent), the behavior of the nanoparticles, the particle parameters, and the applied magnetic fields. Due to missing accurate models for the particle magnetization, the whole function $\tilde{s}$ is commonly determined in a time-consuming calibration process limiting the FOV size as well as the resolution. As the calibration data strictly relies on the particle properties and the measurement sequence, changing tracer material or measurement sequences requires a complete recalibration. Model-based approaches including less simplified behavior of the particles' magnetic moments in the applied fields are highly desirable to reduce the calibration costs and to enable more sophisticated measurement sequences.

8.2.1.1. Scenarios in MPI

The main problem in MPI is given by the forward operator

$$\begin{aligned} F :& X(\Omega) \to Y(0,T)^L \\ & c \mapsto (A_k c + a_k * u_k^E)_{k=1,\dots,L}, \end{aligned}$$

for $L \in \mathbb{N}$ receive coil units with suitable function spaces $X(\Omega)$ and $Y(0,T)$ (assumed to be Hilbert spaces in the following). $A_k : X(\Omega) \to Y(0,T)$, $c \mapsto \int_\Omega \tilde{s}_k(x,t)c(x)\,dx$, $k = 1,\dots,L$, is the forward operator mapping to the analog filtered particle signal for individual receive coil units. The operator F describes the actual measurement process. In MPI we are mainly aiming for solving problems given by the linear operator

$$\begin{aligned} A :& X(\Omega) \to Y(0,T)^L \\ & c \mapsto (A_k c)_{k=1,\dots,L}. \end{aligned}$$

We thus need to get rid of the direct feedthrough u_k^E. In an ideal situation it holds $A = F$ as one assumes $a_k * u_k^E = 0$ but in general this not the case. We can now distinguish two cases in MPI regarding a formal description of the forward operator, the *data-based* case where a full calibration of the linear forward operator is performed and the *model-based* case where a suitable model for the mean magnetic moment $\bar{\mathbf{m}}$ is formulated. Due to the fact that finding a suitable model for the particles' magnetization is still an unsolved problem and the full system matrix calibration is still state of the art in MPI, we distinguish these two cases in the following. As it is possible to specify physical models which might not include relevant aspects, hybrid approaches combining best of both are also desirable. InMPI possible problem setups are as follows (to improve reading convenience, we formulate the scenarios for one coil unit only, i.e., $L = 1$):

- *Data-based* case: Let $\Gamma \subset \mathbb{R}^3$ be a reference volume placed at the origin. The data-based approach uses single measurements of a small sample at predefined positions $\{x^{(i)}\}_{i=1,\dots,N} \in \Omega^N$. The concentration phantoms are given by $c^{(i)} = c_0 \chi_{x^{(i)}+\Gamma}$ for some reference concentration $c_0 > 0$. Typical choices for Γ are small cubes ($\sim$ 1 mm $\times$ 1 mm $\times$ 1 mm). The measurements $v^{(i)} = \frac{1}{c_0} F c^{(i)}$, $i = 1,\dots,N$, are then used to characterize the data-based forward operator.

  Background subtraction case (D1): Let $v^{(0)} = F\mathbf{0}$. Assuming the calibration positions are chosen such that $x^{(i)} + \Gamma$ are pairwise disjoint and $\Omega = \cup_{i=1}^N x^{(i)} + \Gamma$, the specific discretised problem of the model-based approach corresponds to the data-based approach (solving $Ac = v - v^{(0)}$) with $S\tilde{c} = v - v^{(0)}$ with $S = [v^{(1)} - v^{(0)} | \dots | v^{(N)} - v^{(0)}]$ and where $c = \sum_{i=1}^N \tilde{c}_i \chi_{x^{(i)}+\Gamma}$. The full discrete setup is obtained by a finite dimensional approximation in $Y_M \subset Y(0,T)$, i.e., assume $v = \sum_{i=1}^M \langle \phi_i, v \rangle \phi_i$ where $\{\phi_j\}_{j\in\mathbb{N}}$

is an ONB of $Y(0,T)$. It then reads: Find $\tilde{c} \in \mathbb{R}^N$ for given measurement $\tilde{v} \in \mathbb{K}^M$, $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$, ($v \in Y(0,T)$ obtained from an evaluation of F) such that

$$\tilde{S}_{\mathrm{D1}}\tilde{c} = \tilde{v} - \tilde{v}^{(0)}, \quad \tilde{S}_{\mathrm{D1}} = [\tilde{v}^{(1)} - \tilde{v}^{(0)}| \dots |\tilde{v}^{(N)} - \tilde{v}^{(0)}] \tag{8.2}$$

where $\tilde{v} = (\langle \phi_i, v \rangle)_{i=1,\dots,M}$ ($\tilde{v}^{(k)}$ obtained analogously).

Partial temporal information case (D2): Alternative strategies to remove the influence of the background signal $v^{(0)}$ (due to the structure of $v^{(0)}$ in a certain basis) is to restrict the problem to partial data. The discretisation of $Y(0,T)$ is again obtained via the ONB $\{\phi_i\}_{i=1,\dots,M}$. Let $\{\psi_j\}_{j\in\mathbb{N}}$ be another ONB of $Y(0,T)$ (can also be equal to $\{\phi_i\}_{i=1,\dots,M}$). We further assume $v^{(0)}$ is sparse in $\{\psi_j\}_{j\in\mathbb{N}}$, i.e. $|\langle v^{(0)}, \psi_j \rangle| \neq 0$, $j \in J$, $|J| < \infty$. We thus formulate the reduced data problem within the data-based case: It then reads finding $\tilde{c} \in \mathbb{R}^N$ for given measurement $\tilde{v} \in \mathbb{K}^M$, $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$, ($v \in Y(0,T)$ obtained from an evaluation of F) such that

$$\langle S\tilde{c}, \psi_j \rangle = \langle v - v^{(0)}, \psi_j \rangle, \ j \in \mathbb{N} \setminus J, \ S = [v^{(1)} - v^{(0)}| \dots |v^{(N)} - v^{(0)}] \tag{8.3}$$

Exploiting the sparsity assumption on $v^{(0)}$ and using the finite-dimensional approximation in Y_M yields the final problem

$$\begin{aligned} &\langle \tilde{S}_{\mathrm{D2}}\tilde{c}, (\langle \phi_i, \psi_j \rangle)_{i=1,\dots,M} \rangle = \langle \tilde{v}, (\langle \phi_i, \psi_j \rangle)_{i=1,\dots,M} \rangle, \ j \in \mathbb{N} \setminus J, \\ &\tilde{S}_{\mathrm{D2}} = [\tilde{v}^{(1)}| \dots |\tilde{v}^{(N)}] \end{aligned} \tag{8.4}$$

The crucial part is to obtain the correct index set J. In the literature this is commonly done by an SNR-threshold technique with respect to $\{\psi_j\}_{j\in\mathbb{N}}$ being the Fourier basis for T-periodic signals. If the index set is determined incorrectly, one inverts an affine linear system assuming it is linear causing additional artifacts in the reconstruction (i.e., in a noise-free case the reconstruction c^* of a true $c^\dagger$ is obtained via $c^* = A^{-1}Fc^\dagger = c^\dagger + A^{-1}v^{(0)}$).

- *Model-based* case: The challenging part in the model-based case is formulating the correct model for the mean magnetic moment $\bar{\mathbf{m}}$.

  Equilibrium model (monodisperse / polydisperse) (M1): One of the most extensively studied models in MPI is based on the Langevin function. This model is motivated by the assumptions that the applied magnetic field is static and the particles are in equilibrium. Under these assumptions, we assume that the mean magnetic moment vector of the nanoparticles immediately follows the magnetic field, i.e.:

$$\bar{\mathbf{m}}(x,t) = m_0 \mathcal{L}_\beta(|\mathbf{H}(x,t)|) \frac{\mathbf{H}(x,t)}{|\mathbf{H}(x,t)|} \tag{8.5}$$

  where $\mathcal{L}_\beta : \mathbb{R} \to \mathbb{R}$ is given in terms of the Langevin function by the following:

$$\mathcal{L}_\beta(z) = \left(\coth(\beta z) - \frac{1}{\beta z} \right) \tag{8.6}$$

  for $m_0, \beta > 0$. The final problem with respect to the Langevin function is to obtain the concentration c from the following system of equations:

$$\begin{cases} v(t) = -\displaystyle\int_0^T \int_\Omega c(x) \tilde{a}_k(t-t') s_k(x,t') \, dx \, dt' \\ s = \mu_0 m_0 (\mathbf{p}^R)^T \dfrac{\partial}{\partial t} \left(\mathcal{L}_\beta(|\mathbf{H}|) \dfrac{\mathbf{H}}{|\mathbf{H}|} \right) \end{cases} \tag{8.7}$$

$I_3 \in \mathbb{R}^{3\times 3}$ being the identity matrix . The *equilibrium model* in (8.7) can be extended to polydisperse tracers by adapting the function defining the length of the mean magnetic moment vector in (8.6). The tracer material is then modeled by a distribution of particles with different diameters $D > 0$. Assuming that the particle's diameter distribution is given by the density function $\rho : \mathbb{R}^+ \to \mathbb{R}^+ \cup \{0\}$ with $\|\rho\|_{L^1(\mathbb{R}^+)} = 1$, we obtain the extended problems by the following:

$$\begin{cases} v(t) = -\displaystyle\int_0^T \int_\Omega c(x)\tilde{a}(t-t')s(x,t')\,dx\,dt' \\ s = \mu_0(\mathbf{p}^R)^T \dfrac{\partial}{\partial t}\left(L_\rho(|\mathbf{H}|)\dfrac{\mathbf{H}}{|\mathbf{H}|}\right) \end{cases} \tag{8.8}$$

where $L_\rho : \mathbb{R} \to \mathbb{R}$ is given in terms of the Langevin function by

$$L_\rho(z) = \int_{\mathbb{R}^+} \rho(D)m_0(D)\mathcal{L}_{\beta(D)}(z)\,dD \tag{8.9}$$

with $m_0, \beta : \mathbb{R}^+ \to \mathbb{R}^+$ describing the influence of the particle diameter on the volume of the core, respectively the magnetic moment.

Imperfect models suitable for lower quality image reconstruction (M2): This category includes a rather large number of possible model approaches. Generally spoken, this comprises models which approximate the behavior of the system matrix (commonly fitting some model parameters to real data) but cannot reach the reconstruction quality of the data-based case. For example, one can treat the analog filter a as one unknown parameter and fit the model in (M1) to real data.

Suitable model for magnetization dynamics (M3): The operator is properly described by a mathematical model including a sufficient model for the magnetization behavior of the tracer. This requires considering (or approximating) Brownian and Neél rotation mechanisms in the magnetic moment rotation of the nanoparticles. The important difference to (M2) is that here we assume that this model is qualitatively an alternative to the data-based case. This is still an unsolved problem but it is added to the list of possible cases to emphasize the opportunities in the context of Deep Learning approaches.

Using the previously formulated standard setups in MPI we can formulate different problem scenarios (S) which are discussed in the context of Deep Learning in the remainder of this article:

- **Scenario (S1):**
  Given information (D1): measured and background-corrected system matrix ($\tilde{S}_{\mathrm{D1}}$), background measurements $\tilde{v}^{(0)}$, a phantom measurement $\tilde{v}$.
  Possible targets: (i) reduce number of calibration scans (larger delta sample and/or missing regions), (ii) obtain fast and improved concentration reconstructions, (iii) obtain memory-efficient representation, (iv) obtain cleaned particle signal
- **Scenario (S2):**
  Given information (D2): measured system matrix ($\tilde{S}_{\mathrm{D2}}$), background measurements $\tilde{v}^{(0)}$ (optional), a phantom measurement $\tilde{v}$.
  Possible targets: (i) reduce number of calibration scans (larger delta sample and/or missing regions), (ii) obtain fast and improved concentration reconstructions, (iii) obtain memory-efficient representation, (iv) obtain correct index set J, (iv) obtain completed particle signal
- **Scenario (S3):**
  Given information (D1 or D2, M1): measured system matrix ($\tilde{S}_{\mathrm{D1}}$ or $\tilde{S}_{\mathrm{D2}}$), background measurements $\tilde{v}^{(0)}$ (in (D2) optional), a phantom measurement $\tilde{v}$, similar but oversimplified model for $\bar{\mathbf{m}}$ (see M1).
  Possible targets: (i) reduce number of calibration scans (larger delta sample and/or missing regions), (ii) obtain fast and improved concentration reconstructions, (iii) obtain memory-efficient representation, (iv)

obtain correct index set J (in (D2)) , (iv) obtain completed particle signal, (v) measured signal correction (background), completion, and mapping to range of oversimplified model (reconstruction is than obtained via oversimplified model).

- **Scenario (S4):**
  Given information (M1): background measurements $\tilde{v}^{(0)}$, a phantom measurement $\tilde{v}$, similar but oversimplified model for $\bar{\mathbf{m}}$ (see M1).
  Possible targets: (i) obtain fast and improved concentration reconstructions, (ii) obtain improved operator, (iii) obtain cleaned particle signal from measured phantom data.
- **Scenario (S5):**
  Given information (M2): background measurements $\tilde{v}^{(0)}$, a phantom measurement $\tilde{v}$, imperfect model for $\bar{\mathbf{m}}$ which allows reasonable reconstruction (see M2).
  Possible targets: (i) obtain fast and improved concentration reconstructions, (ii) obtain improved operator, (iii) obtain cleaned particle signal from measured phantom data.
- **Scenario (S6):**
  Given information (D1 or D2, M2): measured system matrix ($\tilde{S}_{\mathrm{D1}}$ or $\tilde{S}_{\mathrm{D2}}$), background measurements $\tilde{v}^{(0)}$ (in (D2) optional), a phantom measurement $\tilde{v}$, imperfect model for $\bar{\mathbf{m}}$ which allows reasonable reconstruction (see M2).
  Possible targets: (i) reduce number of calibration scans (larger delta sample and/or missing regions), (ii) obtain fast and improved concentration reconstructions (e.g., post-processed reconstruction of lower quality), (iii) obtain memory-efficient representation, (iv) obtain correct index set J (in (D2)) , (iv) obtain completed particle signal, (v) measured signal correction (background), completion, and optional mapping to range of imperfect model (reconstruction is than obtained via oversimplified model).
- **Scenario (S7) (hypothetical):**
  Given information (M3): background measurements $\tilde{v}^{(0)}$, a phantom measurement $\tilde{v}$, suitable model for $\bar{\mathbf{m}}$ (see M3).
  Possible targets: (i) obtain fast and improved concentration reconstructions, (ii) obtain cleaned particle signal from measured phantom data.

8.2.2. Deep Learning and Inverse Problems

In the following subsections we will present the basic ideas behind scenarios of applying Deep Learning to Inverse Problems and briefly introduce explicit methods to deploy them.

We use the following naming conventions:

- The forward operator:
  $$A : X \to Y,$$
  where X (a Banach space) the reconstruction space and Y (a Banach space) the signal space.
- $p_X : X \to \mathbb{R}_{g_\epsilon 0}$ the probability density function of the elements/images/concentrations that we want to reconstruct.
- $p_Y : Y \to \mathbb{R}_{g_\epsilon 0}$ the probability density function of the elements/signals/voltages that we are measuring.
- $p_{XY} : X \times Y \to \mathbb{R}_{g_\epsilon 0}$ the joint probability density function.

8.2.2.1. A: Learned Penalty Terms

In variational approaches to Inverse Problems one usually incorporates a so-called penalty function

$$\phi : X \to \mathbb{R}_{g_\epsilon 0}$$

to favor good reconstructions, via minimizing a Tikhonov-type functional

$$\min_x ||Ax - y||_2^2 + \phi(x)$$

(or similar). This penalty function is usually handcrafted and at best a very rough approximation of what one would want to have as a penalty term. Given enough data to be representative of p_X, one can use neural networks to learn the "ideal" penalty function. This approach has been explored in [19] [20].

Like for the Plug-and-Play Prior Methods, (B), these methods only require knowledge about p_X. This can be provided via handcrafted parts, via known ground truth data or via surrogate, which represents p_X good enough.

- **What do we learn:**
  The parametrization θ of a penalty function $\phi_\theta : X \to \mathbb{R}_{g_\epsilon 0}$ (for example represented by a neural network).
- **Intrinsically required knowledge:**
  Ground truth data about p_X e. g. in the form of many samples.

8.2.2.2. B: Plug-and-Play Prior Methods

This approach is in some sense (w. r. t. proximal operators) dual to approach A.

Plug-and-Play Prior where introduced in [21], outside the Deep Learning context. The authors pointed out that the alternating direction method of multipliers (ADMM) algorithm used for reconstruction in Inverse Problems could easily be adapted to incorporate any type of denoising method (or more general, any type of algorithm) as a prior during reconstruction. The prior is incorporated by replacing the proximal-operator, $\text{prox}_\phi : X \to X$, of some penalty function, ϕ (within the reconstruction algorithm e.g. ADMM) with some other operator. It is worth noting, that this operator does not have to be a proximal operator of some penalty function anymore. With the rise of Deep Learning methods people started to learn these proximal-operator which are replaced by neural networks.

It was also possible to extend the basic idea to other algorithms like proximal descent or primal dual hybrid gradient, not just ADMM, see for example [22, 23].

- **What do we learn:**
  The parametrization θ of an operator $p_\theta : X \to X$ that – incorporated in some reconstruction algorithm, in which it replaces the proximal operator of some penalty function – improves the reconstruction.
- **Intrinsically required knowledge:**
  Ground truth data about p_X, e. g., in the form of many samples.

8.2.2.3. C: Gradient-Descent-by-Gradient-Descent type Methods

A major field of research in Inverse Problems and Deep Learning is to learn an iterative method from data. This field makes implicit use of the Plug-and-Play prior idea but goes way further. Two very prominent papers in this field are the "Learning Fast Approximations of Sparse Coding" (LISTA) paper [24] and the "Learning to learn by gradient descent by gradient descent" paper [25]. Despite the fact that the paper [25] is not explicitly solving an Inverse Problem their method can be seen as the most data driven form of this type of method. Other authors applied the ideas from [25] directly to solve Inverse Problems [26].

The basic idea of this method has be extended in multiple ways to incorporate prior knowledge e.g. by unrolling existing iteration methods up until a fixed number of iterations and replacing different parts of them by neural networks [27, 28, 29, 30, 31, 32].

One tries to learn parameters θ of a function $f_\theta : Y \to X$, such that it produces "nice" reconstructions via

$$\min_\theta \mathbb{E}_{p_{XY}}[d(f_\theta(y), x)],$$

where d is some measure of distance.

- **What do we learn:**
  The parametrization θ of an operator $f_\theta : Y \to X$ that directly produces "nice" reconstructions.
- **Intrinsically required knowledge:**
  Ground truth data about p_{XY} e. g. in the form of many samples.

8.2.2.4. D: Regularization by Architecture

From an abstract point most (if not all) of machine learning methods – and therefore all methods above – can be seen as parameter identification problems. Despite this being the case for all of the above mentioned types this one is the closest to the original notion of parameter identification problems, since no training data is required; one solely fits the parameters θ of a function

$$F_\theta : G \to \mathbb{R}_{g_\epsilon 0}$$

for (usually) a single point of data, $g \in G$ via

$$\min_\theta F_\theta(g).$$

These methods are very new, there is only some internal work by us and one very recent paper by a group from the University of Texas [33]. Both works are inspired by a recent paper [34] that noticed, that the inherent structure of so-called convolutional neural networks (CNN) is a good prior for images (even without training). This inspired the authors in [34] to solve Inverse Problems with the prior that the reconstructed image has to lie in the range of a CNN-architecture (non-specific to a given parameterization of the network).

- Deep Image Prior:
  $$F_\theta(u) = ||u - Ac(\theta)||_2^2,$$
  where u is the measured signal, A the MPI operator and c (the output of an untrained neural network parameterized by θ whose input is a constant) the concentration.
- Our "Deep Operator Priors" are all of the form:
  $$F_\theta(A) = \sum_{i=1}^{N} w_i ||A_i - f_\theta(c_i)||_2^2 + \phi(f_\theta, c_i),$$
  where f_θ is a neural network representing and enforcing the form of the forward operator in its architecture, A_i usually the columns of a measured operator corresponding to sample concentrations c_i and ϕ a penalty function enforcing structures on weights in the internal representations of f_θ. The w_i are weights to incorporate SNR and similar knowledge. Possible architectures are mixtures of CNN for the spatial structure and RNN (recurrent neural network) for the temporal (frequency) structure.

It is also a still open question whether the $||\cdot||_2$-loss is really the best way to enforce the regression, we are also thinking about using the Wasserstein loss [29].

We see huge potential in these methods, since they allow one to incorporate abstract structural knowledge about the object one ones to regularize. The main difference to "classic" parameter identification is, that one uses deep models as structures (instead of e. g. differential equations). This allows one to incorporate more abstract and less precise knowledge about some underlying structure of a given point of data These methods

relate to traditional parameter identification problems, like Deep Learning relates to machine learning.

We not only want to use these type of methods via the one described in [33], but also via own approaches. We see massive potential in applying these types of ideas to the forward operator via casting abstract knowledge about it into into the architectural design of a neural network that in-turn is fitted to a measured (noisy and error prone, maybe incomplete) version of the forward operator.

These methods are especially interesting for Magnetic Particle Imaging, since they do not rely on ground truth data. This is crucial, since forMPI – as well as for nearly any other novel Inverse Problem – one can not expect to have sufficient ground truth data, see chicken-and-egg-problem.

- **What do we learn:**
  A regularized version of some data point (e. g. an image or even an operator itself).
- **Intrinsically required knowledge:**
  Any kind of abstract knowledge about the data point could potentially be incorporated.

8.2.2.5. E: Image Post-Processing via Deep Learning

Of course, all kinds of image post-processing techniques can be applied to improve MPI reconstructions: inpainting, denoising, deblurring, etc. The leading methods in this field are mostly Deep Learning based nowadays. The amount of literature is expanding rapidly, see for examples [35, 36, 37, 38, 39, 40, 41].

8.2.3. Applying Deep Learning to Magnetic Particle Imaging

In MPI we have to deal with the full extend of the chicken-and-egg problem described earlier. Therefore many of the methods described above that rely on ground truth data (and which in general do not solve new Inverse Problems) are not applicable. This makes MPI an example par excellence for bringing Inverse Problems and Deep Learning together in solving previously unsolved Inverse Problems.

Possible approaches to tackle this union:

- Use surrogate data sets.
- Regularization via Architecture.

Using surrogate data sets means to use data that is presumably similar to MPI data, like MRI data, to boot strap a Deep Learning approach. Approaches that lean to that are especially approaches of the kind, where on learns a penalty term, since theses methods a intrinsically from the Inverse Problem itself.

Using architecture as a regularization is a very new field. Obviously Deep Image Prior should be implement for MPI, but there are a myriad of other opportunities to evaluate. For example one could use the structure provided by an recurrent neural network, like a long-short-term-memory network (LSTM) combined with a convolutional neural network to do inpainting/deblurring on the measured operator to reconstruct measurements with bad signal-to-noise ratios. This could be done via one big optimization in which one optimizes the structure of an extremely deep LSTM that reaches over all frequency-(or time) measurements of an operator at the same time. The goal of the optimization is to fit the output of the network to the measured operator (learning its structure) weighted by the signal-to-noise ratio of the measurements.

8.3. Data sets

On GitHub (https://magneticparticleimaging.github.io/OpenMPIData.jl/latest/index.html) an open magnetic particle imaging data set is available in the MPI data format (MDF) that can be red from any programming language like Matlab, Python, Julia and C. MDF provides a common data format for the storage of MPI raw data, calibration data and reconstruction data. The data set is available under the creative commons license and therefore for everyone free to use, to share and to adapt the data by citing the project and the MDF publication (https://arxiv.org/abs/1602.06072v6).

DL scenario \ MPI scenario	S1	S2	S3	S4	S5	S6	S7
A	0.10	0.10	0.15	0.05	0.75	0.85	0.90
B	0.10	0.10	0.15	0.05	0.75	0.85	0.90
C	0.10	0.10	0.15	0.00	0.80	0.90	1.00
D	0.70	0.70	0.75	0.00	0.00	0.80	0.85
E	**	**	**	**	**	**	**

Table 8.1: The bigger the number the more potential we see for improving image quality via Deep Learning Methods. For the methods (A), (B) and (C) we always assume the existence if good surrogate data to model p_X, e. g. MRI data or MPI X-space data. ** denotes the cases that require further investigation.

DL scenario \ MPI scenario	S1	S2	S3	S4	S5	S6	S7
A			[19]		[19]	[19]	
B			[23]		[23]	[23]	
C			[30]		[30]	[30]	
D	[42],*	[42],*	[42],*			[42],*	
E	**	**	**	**	**	**	**

Table 8.2: Relevant short- and mid-term projects are marked with citations for papers which should be adapted to MPI. All scenarios associated with S7 are long-term projects. * denotes ideas approached in the near future. (E) needs some work to identify most promising approaches to be adapted to MPI. ** denotes the cases that requre further investigation.

The MPI data set has been measured with a Bruker PreclinicalMPI Scanner (Bruker Biospin, Ettlingen, Germany) and the dextran based magnetic nano-particles perimag (micromod, Partikeltechnologie GmbH, Rostock, Germany) have been used as tracers for the experiments. For the MPI measurements phantoms printed with the stereolithography 3D printing technique were used. All phantoms share the same robot mount defining the coordinate system. The robot mount is fixed in positive X direction. The flat side marks the positive Z direction. The positive Y direction is then defined by right hand rule.

To resemble a well-known **shape phantom** a cone defined by a 1 mm radius tip and an apex angle of 10 deg and a height of 22 mm. The total volume is 683.9 μl. As tracer perimag in a concentration of 50 mmol per l were used. The phantom can be rendered in 3D resembling the cone or can be cut in layer view to see either a circle (YZ plane) or a triangle with flattened tip (XZ or XY plane).

The **resolution phantom** consists of five tubes filled with perimag featuring a concentration of 50 mmol. The five tubes have a common origin on one side of the phantom. From there the extend in different angles from this origin within the XY and YZ plane are chosen smaller (10 and 15 deg) than in XY direction (20 and 30 deg). By choosing different planes, one can determine the capable resolution due to different distances of the tubes.

The **concentration phantom** consists of eight cubes 2 mm length resulting in 8 μl volume each. The distance of the cubes are 12 mm between centres within the XY plane and 6 mm between centres within Z direction. The concentrations in the eight sample chambers are different, ranging from 5.85 to 44.4 mmol per l.

Several MPI sequences have been used for data acquisition.

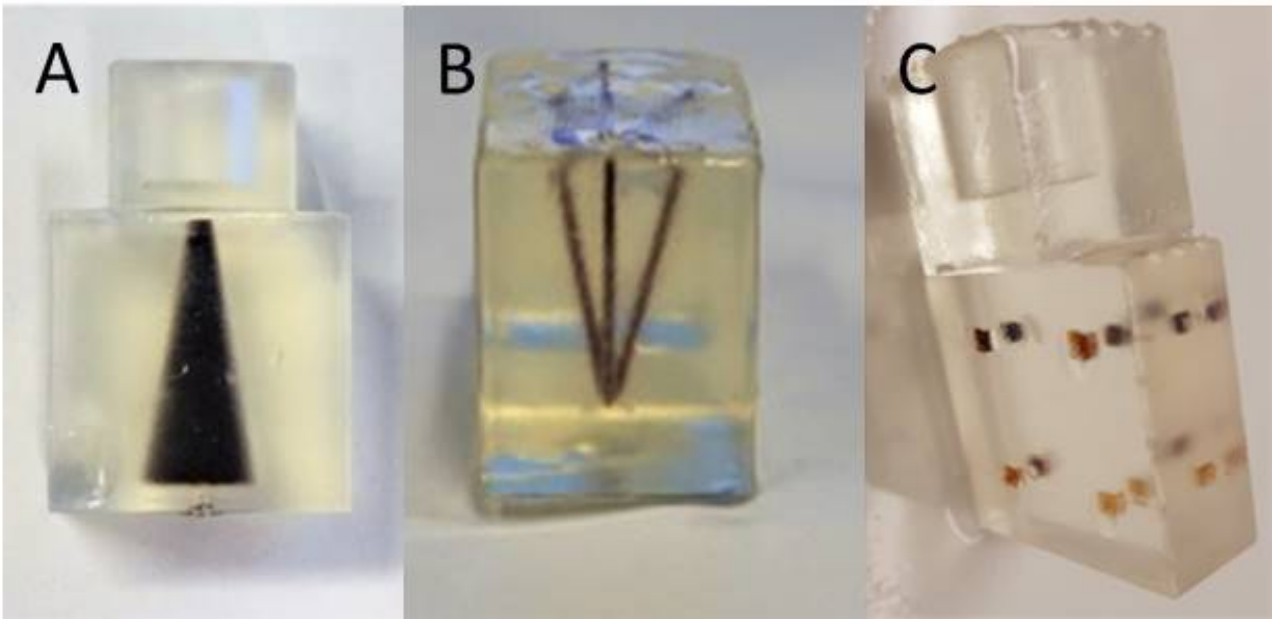


Figure 8.1: Visualization of phantoms used for MPI measurements. A: shape phantom, B: resolution phantom and C: concentration phantom

8.3.1. 1D Cartesian Sequence

During the measurement of the 1D sequence the phantom has been moved step-wise in the z-dimension and afterwards in the y-dimension with the robot. Each patch represents one robot position, in total 19x19 = 361 1D patches per position.

8.3.2. 2D Lissajous Sequence

During the measurement of the 2D sequence the phantom has been moved step-wise in z-dimension with the robot. Each patch represents one robot position, in total 19 2D pateches per position.

Bibliography

[1] J. Weizenecker, B. Gleich, J. Rahmer, H. Dahnke, and J. Borgert, "Three-dimensional real-time in vivo magnetic particle imaging," *Physics in Medicine and Biology*, vol. 54, no. 5, p. L1, 2009.

[2] A. Khandhar, P. Keselman, S. Kemp, R. Ferguson, P. Goodwill, S. Conolly, and K. Krishnan, "Evaluation of peg-coated iron oxide nanoparticles as blood pool tracers for preclinical magnetic particle imaging," *Nanoscale*, vol. 9, no. 3, pp. 1299–1306, 2017.

[3] J. Franke, R. Lacroix, H. Lehr, M. Heidenreich, U. Heinen, and V. Schulz, "Mpi flow analysis toolbox exploiting pulsed tracer information – an aneurysm phantom proof," *International Journal on Magnetic Particle Imaging*, vol. 3, no. 1, 2017. [Online]. Available: https://journal.iwmpi.org/index.php/iwmpi/article/view/36

[4] J. Haegele, J. Rahmer, B. Gleich, J. Borgert, H. Wojtczyk, N. Panagiotopoulos, T. Buzug, J. Barkhausen, and F. Vogt, "Magnetic particle imaging: visualization of instruments for cardiovascular intervention," *Radiology*, vol. 265, no. 3, pp. 933–938, 2012.

[5] J. Salamon, M. Hofmann, C. Jung, M. G. Kaul, F. Werner, K. Them, R. Reimer, P. Nielsen, A. vom Scheidt, G. Adam, T. Knopp, and H. Ittrich, "Magnetic particle/magnetic resonance imaging: In-vitro

MPI-guided real time catheter tracking and 4D angioplasty using a road map and blood pool tracer approach," *PloS ONE*, vol. 11, no. 6, pp. e0 156 899–14, 2016.

[6] E. Y. Yu, M. Bishop, B. Zheng, R. M. Ferguson, A. P. Khandhar, S. J. Kemp, K. M. Krishnan, P. W. Goodwill, and S. M. Conolly, "Magnetic particle imaging: A novel in vivo imaging platform for cancer detection," *Nano Letters*, vol. 17, no. 3, pp. 1648–1654, 2017. [Online]. Available: http://dx.doi.org/10.1021/acs.nanolett.6b04865

[7] K. Murase, M. Aoki, N. Banura, K. Nishimoto, A. Mimura, T. Kuboyabu, and I. Yabata, "Usefulness of magnetic particle imaging for predicting the therapeutic effect of magnetic hyperthermia," *Open Journal of Medical Imaging*, vol. 5, no. 02, p. 85, 2015.

[8] K. H. Jin, M. T. McCann, E. Froustey, and M. Unser, "Deep convolutional neural network for inverse problems in imaging," *IEEE Transactions on Image Processing*, vol. 26, no. 9, pp. 4509–4522, 2017.

[9] S. Lunz, O. Öktem, and C.-B. Schönlieb, "Adversarial regularizers in inverse problems," *arXiv preprint arXiv:1805.11572*, 2018.

[10] J. Behrmann, C. Etmann, T. Boskamp, R. Casadonte, J. Kriegsmann, and P. Maaβ, "Deep learning for tumor classification in imaging mass spectrometry," *Bioinformatics*, vol. 34, no. 7, pp. 1215–1223, 2017.

[11] S. Wang, Z. Su, L. Ying, X. Peng, S. Zhu, F. Liang, D. Feng, and D. Liang, "Accelerating magnetic resonance imaging via deep learning," in *Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on*. IEEE, 2016, pp. 514–517.

[12] J. Schlemper, J. Caballero, J. V. Hajnal, A. N. Price, and D. Rueckert, "A deep cascade of convolutional neural networks for dynamic mr image reconstruction," *IEEE transactions on Medical Imaging*, vol. 37, no. 2, pp. 491–503, 2018.

[13] E. Kang, J. Min, and J. C. Ye, "A deep convolutional neural network using directional wavelets for low-dose x-ray ct reconstruction," *Medical physics*, vol. 44, no. 10, 2017.

[14] K. Gong, J. Guan, K. Kim, X. Zhang, G. E. Fakhri, J. Qi, and Q. Li, "Iterative pet image reconstruction using convolutional neural network representation," *arXiv preprint arXiv:1710.03344*, 2017.

[15] B. Gleich and J. Weizenecker, "Tomographic imaging using the nonlinear response of magnetic particles," *Nature*, vol. 435, no. 7046, pp. 1214–1217, June 2005.

[16] J. Weizenecker, B. Gleich, and J. Borgert, "Magnetic particle imaging using a field free line," *Journal of Physics D: Applied Physics*, vol. 41, no. 10, p. 105009, 2008. [Online]. Available: http://stacks.iop.org/0022-3727/41/i=10/a=105009

[17] T. Kluth, "Mathematical models for magnetic particle imaging," *Inverse Problems*, 2018, accepted manuscript online available at http://iopscience.iop.org/article/10.1088/1361-6420/aac535. [Online]. Available: http://iopscience.iop.org/article/10.1088/1361-6420/aac535

[18] T. Knopp and T. M. Buzug, *Magnetic Particle Imaging: An Introduction to Imaging Principles and Scanner Instrumentation*. Berlin/Heidelberg: Springer, 2012.

[19] S. Lunz, O. Öktem, and C.-B. Schönlieb, "Adversarial regularizers in inverse problems." [Online]. Available: https://arxiv.org/abs/1805.11572

[20] H. Li, J. Schwab, S. Antholzer, and M. Haltmeier, "Nett: Solving inverse problems with deep neural networks," *arXiv preprint arXiv:1803.00092*, 2018.

[21] S. V. Venkatakrishnan, C. A. Bouman, and B. Wohlberg, "Plug-and-play priors for model based reconstruction," in *Global Conference on Signal and Information Processing (GlobalSIP), 2013 IEEE*. IEEE, 2013, pp. 945–948.

[22] J. R. Chang, C. Li, B. Póczos, B. V. K. V. Kumar, and A. C. Sankaranarayanan, "One network to solve them all - solving linear inverse problems using deep projection models," *CoRR*, vol. abs/1703.09912, 2017. [Online]. Available: http://arxiv.org/abs/1703.09912

[23] T. Meinhardt, M. Möller, C. Hazirbas, and D. Cremers, "Learning proximal operators: Using denoising

networks for regularizing inverse imaging problems," *CoRR*, vol. abs/1704.03488, 2017. [Online]. Available: http://arxiv.org/abs/1704.03488

[24] K. Gregor and Y. LeCun, "Learning fast approximations of sparse coding," in *Proceedings of the 27th International Conference on International Conference on Machine Learning*. Omnipress, 2010, pp. 399–406.

[25] M. Andrychowicz, M. Denil, S. Gomez, M. W. Hoffman, D. Pfau, T. Schaul, B. Shillingford, and N. De Freitas, "Learning to learn by gradient descent by gradient descent," in *Advances in Neural Information Processing Systems*, 2016, pp. 3981–3989.

[26] P. Putzky and M. Welling, "Recurrent inference machines for solving inverse problems," *arXiv preprint arXiv:1706.04008*, 2017.

[27] J. Adler and O. Öktem, "Solving ill-posed inverse problems using iterative deep neural networks," 2017. [Online]. Available: https://arxiv.org/abs/1704.04058

[28] B. Kelly, T. P. Matthews, and M. A. Anastasio, "Deep learning-guided image reconstruction from incomplete data," *arXiv preprint arXiv:1709.00584*, 2017.

[29] J. Adler, A. Ringh, O. Öktem, and J. Karlsson, "Learning to solve inverse problems using wasserstein loss," *CoRR*, vol. abs/1710.10898, 2017. [Online]. Available: http://arxiv.org/abs/1710.10898

[30] J. Adler and O. Öktem, "Learned primal-dual reconstruction," 2017. [Online]. Available: https://arxiv.org/abs/1707.06474

[31] A. Hauptmann, F. Lucka, M. M. Betcke, N. Huynh, B. T. Cox, P. C. Beard, S. Ourselin, and S. R. Arridge, "Model based learning for accelerated, limited-view 3d photoacoustic tomography," *CoRR*, vol. abs/1708.09832, 2017. [Online]. Available: http://arxiv.org/abs/1708.09832

[32] K. Hammernik, T. Klatzer, E. Kobler, M. P. Recht, D. K. Sodickson, T. Pock, and F. Knoll, "Learning a variational network for reconstruction of accelerated mri data," *Magnetic resonance in medicine*, vol. 79, no. 6, pp. 3055–3071, 2018.

[33] D. Van Veen, A. Jalal, E. Price, S. Vishwanath, and A. G. Dimakis, "Compressed sensing with deep image prior and learned regularization," *arXiv preprint arXiv:1806.06438*, 2018.

[34] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep image prior," *arXiv preprint arXiv:1711.10925*, 2017.

[35] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 184–199.

[36] L. Xu, J. S. Ren, C. Liu, and J. Jia, "Deep convolutional neural network for image deconvolution," in *Advances in Neural Information Processing Systems*, 2014, pp. 1790–1798.

[37] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 2, pp. 295–307, 2016.

[38] H. C. Burger, C. J. Schuler, and S. Harmeling, "Image denoising: Can plain neural networks compete with bm3d?" in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2392–2399.

[39] J. Xie, L. Xu, and E. Chen, "Image denoising and inpainting with deep neural networks," in *Advances in neural information processing systems*, 2012, pp. 341–349.

[40] X. Tao, H. Gao, X. Shen, J. Wang, and J. Jia, "Scale-recurrent network for deep image deblurring," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8174–8182.

[41] C. Chen, Q. Chen, J. Xu, and V. Koltun, "Learning to see in the dark," *arXiv preprint arXiv:1805.01934*, 2018.

[42] D. V. Veen, A. Jalal, E. Price, S. Vishwanath, and A. G. Dimakis, "Compressed sensing with deep image prior and learned regularization," *CoRR*, vol. abs/1806.06438, 2018. [Online]. Available: https://arxiv.org/abs/1806.06438

9. Initial benchmark case for integrated optimization of cross-border railway traffic

Andreas Bärmann[1], Jonasz Staszek[1]

[1] *Friedrich-Alexander-Universität Erlangen-Nürnberg*

Abstract. In this submission we introduce our problem – devising an optimization model used for a simultaneous attribution of locomotives and drivers to transportation tasks in the international context. We quickly discuss the novelty of the suggested approach. We also describe the industrial partner – DB Cargo Polska. A major part of the submission is devoted to describing the integer model devised so far. It is an integer model, with seven groups of binary (decision) variables and thirty constraints, reflecting a wide variety of legal, technical and organizational constraints in the problem studied (e.g. drivers' license requirements to locomotives and routes, country-specific working time requirements for drivers, maintenance needs of locomotives etc.). At this point, we optimize for the total number of trains performed (i.e. we maximize it). It is followed by the description of the current status of the model implementation, together with the description of the data which will be made publicly available. A review of the literature in the subject is left to a future submission for this deliverable.

Keywords: Rail freight transport, discrete optimization, international rail freight, rail driver assignment, loco routing, benchmark case, Poland.

9.1. Introduction

9.1.1. Information about the industry partner

Our industrial partner – DB Cargo Polska S.A. – belongs to DB Cargo AG, which connects railway freight carriers from across the Europe and itself belongs to Deutsche Bahn – one of the largest logistics groups in the world.
DB Cargo AG is present in almost all European national markets and from the year 2013, also in the United Arab Emirates. The company is involved in transport organization, rail transport itself in many sectors (e. g. automotive, coal and metals, chemistry, intermodal transport, etc.) and siding operation.
The Company (through its subsidiaries) is also involved in the construction and repair of railway rolling stock. It also manages several railway lines in Upper Silesia, as well as the Szczecin port terminal and transshipment terminals in Kedzierzyn-Kozle and Slawkow at the end of a broad-gauge track connecting Poland with Ukraine. In 2016, the Company owned 243 locomotives and 3112 railway wagons. During the same period, the company transported 57.6 million tons of freight and carried out a transport work of 2.8 billion ton-kilometers.

9.1.2. Relevance and novelty of the industrial application

In our project, we intend to schedule driver and locomotives simultaneously to the existing transportation tasks. This approach is novel, as it:

1. considers two highly interrelated groups of assets simultaneously and hence unlocks new potential optimization choices
2. it has never been treated in the literature before for the railway carriers.

It will help our industrial partner to better manage their crews as well as locomotive fleet. Hence they will be able to serve more trains using existing resources, which shall have both a sound economic (through increased efficiency of both the personnel and the expensive assets) as well as environmental effect (more transportation

work can be performed by the trains, which are believed to be more eco-friendly tran trucks).

9.2. Mathematical description of the model under development

9.2.1. Description of the model

The following sections provide information about the current formulation of the model. It would be relevant to define the primary sets - e.g. the data provided to us by the industry partner. They are namely:

1. Trains set T - including all the client orders for transportation service in a given time range (usually a monthly one), together with departure/arrival times and places.
2. Locomotives set L - comprising of all the locomotives owned or rented by the industrial partner, together with the timing and location of maintenance periods (whenever appropriate)
3. Drivers set D - describing the locomotive drivers employed by the industrial partner, together with their licensing to both locomotive types and route (e.g. section of a network between two stations).

Out of these three sets (as well as additional information, e.g. about suitability of locomotives to trains), the variables and the secondary sets are built. Basing on those, we construct an integer model to optimize the allocation of locomotives and drivers to trains. Its current objective function is to make sure that all (or a maximal possible number) of the client orders are realized.

9.2.1.1. Variables

Variables are summarized in Table 1

Table 9.1: Summary of variables

Name	Description	Type
$x^t_{d,l}$	Train t is served by a d, l combination	binary
y^t_d	Train t is the first job of a driver d in their shift	binary
v^t_d	Train t is the last job of a driver d before 12h break	binary
z^t_d	Train t is the last job of a driver d before 35h break	binary
u^t_l	Train t is the last job of a loco l before a maintenance period	binary
f^{t_1,t_2}_d	Loco l shall serve train t_2 after serving t_1	binary
h^w_d	Driver d has worked on the Sunday of the week w	binary

9.2.1.2. Secondary sets

The sets are summarized in Table 2.

9.2.1.3. Model formulation

Objective function

$$\max \sum_{t\in T, l\in L, d\in D} x^t_{d,l}$$

Time compatibility of trains

$$\sum_{d,l\in D,L} x^t_{d,l} \leq 1 \quad \forall t \in T \tag{9.1}$$

(At most one driver-loco combination can serve each train)

Table 9.2: Summary of secondary sets

Name	Description
T_t^{+12h}	Used to determine which trains t_1 cannot be assigned to a driver d if t is his first job in a working day
T_t^{-12h}	Used to determine which trains t_2 could be the first job assigned to a driver d if t is to be performed by him
T_t^{B+}	Used to determine which job t_3 cannot be assigned to a driver d if t is his last shift before a daily break
T_t^{35h}	Used to determine trains t_4 which cannot be assigned to a driver d if t is his last last job before a 35h break
T_w^{week}	Used to determine which trains t_5 belong to a calculation week w
T_w^{Sunday}	Used to determine which trains t_6 are scheduled for the Sunday h_w
$T_{l,m}^{pre-maint}$	Used to determine which trains t_7 are scheduled to run shortly before a loco l is due for maintenance
T_m^{depo}	Used to determine which trains $t_8 \in T_{l,m}^{pre-maint}$ are heading to the depot station
$T_{t,l,m}^{block}$	Used to determine which $t_9 \in T_{l,m}^{pre-maint}$ depart from p_{m^l} shortly before m^l, but after the train t has arrived
T_l^{next}	Used to determine which train pairs t_1, t_2 could be served by a loco l sequentially
T_t^{time}	Used to determine which trains $t_1 1 \in T$ are in time conflict with the train t

$$\sum_{d \in D} x_{d,l}^{t} + \sum_{d \in D} x_{d,l}^{t_1} \leq 1 \quad \forall l \in L, \quad \forall t_1 \in T_t^{time} \quad \forall t \in T \tag{9.2}$$

$$\sum_{l \in L} x_{d,l}^{t} + \sum_{l \in L} x_{d,l}^{t_1} \leq 1 \quad \forall d \in D, \quad \forall t_1 \in T_t^{time} \quad \forall t \in T \tag{9.3}$$

(At most one of the trains in time conflict can be served with on d, l combination)

Space compatibility of trains

$$\sum_{d \in D} x_{d,l}^{t_2} \leq \sum_{t_2 \in T_{t1}^{next}} f_l^{t_1,t_2} \quad \forall l \in L, \quad \forall t_1 \in T \tag{9.4}$$

(If a train t is served by a d, l pair, i.e. $x_{d,l}^{t}$ is 1, then the next train served by that loco needs to start from the same station (these are stored in f variables))

$$\sum_{t_2 \in T_{t1}^{next}} f_l^{t_1,t_2} \leq \sum_{d \in D} x_{d,l}^{t_2} \quad \forall l \in L, \quad \forall t_1 \in T \tag{9.5}$$

(If a train t is served by a d, l pair, i.e. $x_{d,l}^{t}$ is 1, then the next train served by that loco needs to start from the same station (these are stored in f variables))

$$f_l^{t_1,t_2} \quad + \sum_{\substack{\forall d \in D \\ \forall t_3 \in T_{t_1}^{next}}} x_{d,l}^{t_3} \quad \leq \quad 1 \tag{9.6}$$

$$\forall t_1 \in T, \quad \forall t_2 \in T_{t_1}^{next}, \quad \forall t_3 \in T_{t_1}^{next}, \quad departure(t_2) > departure(t_3) > arrival(t_1)$$

(If t_1 and t_2 are selected to be performed by the same locomotive l, then no train t_3 departing earlier than t_2 can be assigned to the loco l)

$$\sum_{l \in L} f_l^{t_1,t_2} \leq 1 \quad \forall t_1 \in T, \quad \forall t_2 \in T_{t_1}^{next} \tag{9.7}$$

(A loco can be attributed to a maximum of one successor to each t_1 train)

$$\sum_{d \in D} x_{d,l}^{t_2} \leq \sum_{t_2 \in T_{t1}^{next}} g_d^{t_1,t_2} \quad \forall d \in D, \quad \forall t_1 \in T \tag{9.8}$$

(If a train t is served by a driver d, i.e. $x_{d,l}^t$ is 1, then the next train served by that loco needs to start from the same station (these are stored in g variables))

$$\sum_{t_2 \in T_{t1}^{next}} g_d^{t_1,t_2} \leq \sum_{l \in L} x_{d,l}^{t_2} \quad \forall d \in D, \quad \forall t_1 \in T \tag{9.9}$$

(If a train t is served by a d, i.e. $x_{d,l}^t$ is 1, then the next train served by that loco needs to start from the same station (these are stored in f variables))

$$g_d^{t_1,t_2} \; + \sum_{\substack{\forall l \in L \\ \forall t_3 \in T_{t_1}^{next}}} x_{d,l}^{t_3} \; \leq \; 1 \tag{9.10}$$

$$\forall t_1 \in T, \quad \forall t_2 \in T_{t_1}^{next}, \quad \forall t_3 \in T_{t_1}^{next}, \quad departure(t_2) > departure(t_3) > arrival(t_1)$$

(If t_1 and t_2 are selected to be performed by the same driver d, then no train t_3 departing earlier than t_2 can be assigned to the driver d)

$$\sum_{l \in L} f_l^{t_1,t_2} \leq 1 \quad \forall t_1 \in T, \quad \forall t_2 \in T_{t_1}^{next} \tag{9.11}$$

(A driver can be attributed to a maximum of one successor to each t_1 train)

Drivers' working time requirements

$$y_d^t \leq \sum_{l \in L} x_{d,l}^t \quad \forall l \in L \tag{9.12}$$

(If a train t is a first train for a driver d, then it needs to have a loco assigned by x variable)

$$y_d^t + \sum_{d \in D, l \in L} x_{d,l}^{t_1} \leq 1 \quad \forall t \in T, \forall t_1 \in T_t^{+12h} \tag{9.13}$$

(If a job t is selected to be the first train of a driver d in a given working day, then the driver cannot drive trains which end later than $s_t + 12$h, which are stored in T_t^{+12h})

$$x_{d,l}^t \leq \sum_{t \in T^{-12h}} y_d^t \leq 1 \quad \forall d \in D, \forall l \in L \tag{9.14}$$

(If a job t is selected to be carried out by a driver d, then the driver must have started working with a job starting at $e_t - 12h$ or later. These jobs are stored in T_t^{-12h}))

$$y_d^t + \sum_{t \in T_t^{-12h}} x_{d,l}^t \leq 1 \quad \forall d \in D, \forall l \in L \tag{9.15}$$

(If the job t is a first job of driver d, then the driver undertake have undertaken any jobs during the last 12h - these jobs are listed in T_t^{-12h})

$$v_d^t + \sum_{t \in T_t^{B+}} x_{d,l}^t \leq 1 \quad \forall d \in D, \forall l \in L \tag{9.16}$$

(If the job t is a final job of driver d before the daily break, then the driver cannot undertake any jobs within the next 12h - these jobs are listed in T_t^{B+})

$$v_d^t \leq \sum_{t \in T_t^{-12h}} y_d^t \quad \forall d \in D, \forall l \in L \tag{9.17}$$

(If the job t is a final job of driver d before the daily break, then the driver must have begun working within the $[e_t - 12h, e_t]$ - these jobs are listed in T_t^{-12h})

$$z_d^t \leq v_d^t \leq 1 \quad \forall d \in D, \forall t \in T \tag{9.18}$$

(If the job t is a final job to driver d before the weekly break, then it is also the last job before the daily break of that working day. On the other hand, not every daily break is also a weekly break.

$$z_d^t + \sum_{t \in T_t^{35h}} x_{d,l}^t \leq 1 \quad \forall d \in D, \forall l \in L \tag{9.19}$$

(If a job t is selected to be the last job of a driver d before a weekly break, then no jobs starting within $[e_t, e_t + 35h]$ can be attributed to the driver. These jobs are stored in T_t^{35h})

$$1 \leq \sum_{t \in T_w^{week}} z_d^t \qquad \forall d \in D \tag{9.20}$$

(At least one 35h break has to be scheduled per calculation week. All jobs in one calculation week are stored in T_w^{week})

$$x_{d,l}^t \leq h_{w_n,d}^{work} \qquad \forall d \in D, \forall t \in T_{w_n}^{sunday} \tag{9.21}$$

(If at least one job was carried out by the driver d on Sunday, then the corresponding h variable has to equal to 1.)

$$h^{work}_{w_n,d} + h^{work}_{w_{n+1},d} + h^{work}_{w_{n+2},d} + h^{work}_{w_{n+3},d} \leq 3 \qquad \forall d \in D \quad (9.22)$$

(At least every fourth Sunday needs to be off. Information whether the driver has no jobs planned on a Sunday is stored in the h^{work} parameter.)

Locomotive maintenance needs

$$u^{l,m}_t \leq \sum_{d \in D} x^t_{d,l} \leq 1 \qquad \forall t \in T^{pre-maint}_{l,m}, \quad \forall m \in M^l, \qquad \forall l \in L \quad (9.23)$$

(If job t is to be carried out as the last one of the loco l before a maintenance period, then it needs a driver d assigned to it.)

$$u^{l,m}_t + \sum_{d \in D} x^{t_1}_{d,l} \leq 1 \qquad \forall t_1 \in T^{block}_{t,l,m}, \quad \forall m \in M^l, \qquad \forall l \in L \quad (9.24)$$

(If, by performing the job t as the last one before maintenance period m the loco reaches the depot station p_m, it should not perform any further jobs up until the maintenance period begins.)

$$\sum_{t \in T^{depo}_m} u^{l,m}_t = 1 \qquad \forall m \in M^l, \quad \forall l \in L \quad (9.25)$$

(For each maintenance period m the loco should - in its final job $u^{l,m}_t$ - arrive at the depot station.)

Nature of variables

$$x^t_{d,l} \in \{0,1\} \ \ \forall t \in T, \ \ \forall d \in D, \ \ \forall l \in L \quad (9.26)$$

$$y^t_d \in \{0,1\} \ \ \forall t \in T, \ \ \forall d \in D \quad (9.27)$$

$$v^t_d \in \{0,1\} \ \ \forall t \in T, \ \ \forall d \in D \quad (9.28)$$

$$z^t_d \in \{0,1\} \ \ \forall t \in T, \ \ \forall d \in D \quad (9.29)$$

$$u^{l,m}_t \in \{0,1\} \ \ \forall t \in T^{pre-maint}_{l,m}, \ \ \forall t \in T, \ \ \forall m \in M^l, \forall l \in L \quad (9.30)$$

(Finally, we need to make sure that the decision variables are binary.)

9.2.2. Current status of the implementation of the model

The implementation of the model described above is currently developed in Python, using gurobi optimization suite, as well as a number of freeware libraries, such as numpy or pandas. At the current stage, the tasks the code will perform could be summarized in the following way:

1. Transform the data provided by the industrial partner to the sets described in the previous section
2. Construct an integer model basing on the sets generated in step 1, as described in the previous section.
3. Optimize that model.
4. Convert the results of the optimization into a visualization (or other format, agreed with DB Cargo Polska).

9.3. Description of the public available data provided in the next deliverable of this work package WP5.

The data which will be made available will comprise:

1. An anonymized list of transportation tasks ordered by the clients to the industrial partner, together with an indication of suitable locomotive type.
2. An anonymized list of locomotives available for assignment, together with their maintenance periods and an indication where should each maintenance be carried out.
3. An anonymized list of drivers, together with their licenses both to locomotive types as well as to the routes travelled by particular trains.

10. Benchmark case for optimal shape design of air ducts in combustion engines

Naomi Auer[1], Michael Hintermüller[1], Karl Knall[2]

[1] *WIAS Berlin*

[2] *MathTec*

Abstract. In order to optimize the shape design of air ducts in combustion engines, we consider a free form shape optimization problem subject to the Navier-Stokes equations. The associated numerical solution requires an efficient computation and yet accurate approximation of an adjoint-based shape gradient and also elements of higher order in a shape-gradient-related descent method for high Reynolds number flows. Moreover, geometric constraints need to be taken care of. As a benchmark case for the described problem, we propose a curved duct geometry with one inlet and one outlet which is inscribed in a restrictive design space. The geometry has been developed together with engineers from the automotive industry.

Keywords: Shape optimization, Navier-Stokes equations.

10.1. Introduction

Many optimal design problems in engineering or biomedical sciences require to determine an optimal shape of a region of interest in order to minimize a number of suitable objectives subject to fluid flow. For performance optimization of combustion engines, in particular in the automobile industry, the optimal shape design of several components of the engine, such as air ducts, is crucial. Often additional geometric constraints impose further restrictions on the possible design. Mathematically, this problem results in a constrained multi-objective free form shape optimization problem subject to the Navier-Stokes system. In this context, for computer based rapid prototyping, constrained free form shape optimization techniques are typically superior to parameterization based techniques due to their geometric flexibility.

The goal of this project is the derivation of adjoint-based representations of shape gradient-related descent directions and the numerical realization of associated minimization schemes. For this purpose and for reasons of efficiency, reduced order models (e.g. based on shape-aware adaptive discretization) need to be developed. Furthermore, a finite volume or finite element based optimization tool which involves appropriate primal and dual turbulence models for high Reynolds numbers flows needs to be implemented and analyzed and proper preconditioning of saddle point problems has to be developed. The newly developed solver has to be tested for industry relevant use cases in a parallel computing environment.

In this context, applications in the automotive industry and quantitative biomedicine will be addressed.

10.2. The industrial partner

Math.Tec, a Vienna based mathematical consulting company, pursues the strategic goal of providing mathematics-based solutions to real-world challenges in the areas of warehouse logistics optimization, production logistics optimization, transport logistics optimization and industry optimization. Based on the development of mathematical models, Math.Tec provides innovative impulses and customer-centric approaches in order to identify the customer's challenges which are then solved by developing and employing custom-tailored software technology. Math.Tec's core competencies are in the use of modern, up-to-date mathematical techniques, optimization algorithms and software realizations.

Many tasks in the area of industry optimization have a technical and mathematical (physical, electromagnetic, fluid-mechanical or mechanical) background. The numerical simulation of fluids (CFD Computational Fluid Dynamics) is a methodical focus in this area of industry optimization. This is a technique which has increas-

ingly complemented the traditional mainstays of fluid mechanics in recent years. The use of these techniques to solve engineering problems has, along with the further development of simulation techniques and optimising algorithms, come increasingly to the fore. Math.Tec supports many customers, for example in the area of engine building, in realising innovative new mathematical processes in the area of numerical fluid mechanics. The most widely used solution methods in fluid mechanics (which Math.Tec runs on high performance computers) are finite difference methods (FDM), finite volume methods (FVM) and the finite element method (FEM).

10.3. Previous scientific literature

Shape optimization problems governed by the Navier-Stokes equations have already been considered in the literature. We mention the articles [1], [2] and [3], which will be relevant for our project.

In [1], Boisgérault and Zolésio propose regularity assumptions which guarantee the shape differentiability of certain cost functionals. For this purpose, they use the so-called speed method which allows to do the computations in the reference domain instead of a perturbed domain, without loosing the divergence-free property.

In [2], Laurain and Sturm opt for a domain representation of the shape derivative, the so-called distributed shape derivative, as it is more general than a boundary expression and easy to compute. They propose a Lagrangian approach to compute the distributed shape derivative and they show how the level set method can be used in this context.

The authors of [3] consider the shape analysis of the unsteady Navier-Stokes equations. They discuss the Fréchet-differentiability of the flow with respect to the transformation of the problem to a reference domain, in an abstract setting as well as for the case of the unsteady Navier-Stokes equations. With these results, the Fréchet differentiability of certain objective functions is proved.

10.4. Mathematical description

10.4.1. Navier-Stokes equations

In order to model the flow in the considered shape, we use the stationary regime of the Navier-Stokes equations for the velocity $\boldsymbol{u}$ and the kinematic pressure p

$$\begin{aligned}
(\boldsymbol{u}\cdot\nabla)\boldsymbol{u} - \nu\Delta\boldsymbol{u} + \nabla p &= 0 \ \text{ in } \Omega,\\
\nabla\cdot\boldsymbol{u} &= 0 \ \text{ in } \Omega,\\
\boldsymbol{u} &= \boldsymbol{g} \ \text{ on } \Gamma_i,\\
\boldsymbol{u} &= \boldsymbol{0} \ \text{ on } \Gamma_w,\\
-\nu\partial_n\boldsymbol{u} + p\boldsymbol{n} &= \boldsymbol{0} \ \text{ on } \Gamma_o,
\end{aligned}$$

where $\boldsymbol{g}$ is the inflow profile, ν is the kinematic viscosity and $\boldsymbol{n}$ is the outer normal vector. Figure 10.1 shows a schematic geometry and its boundary (left) and the solution $\boldsymbol{u}$ of the Navier-Stokes equations (right).

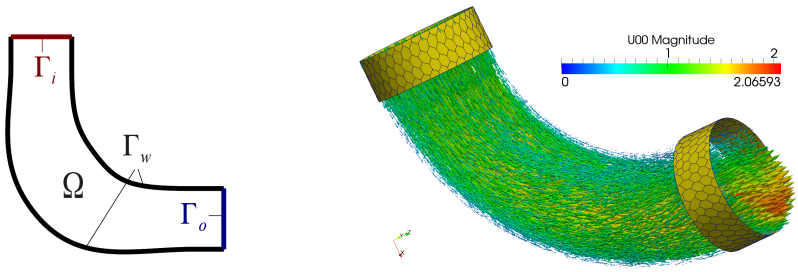


Figure 10.1: Left: sketch of a geometry and its boundary. Right: solution $\boldsymbol{u}$ of the Navier-Stokes equations.

10.4.2. Cost functional

The cost functional depends on the geometry Ω. In order to achieve uniform outflow at the outlet, the cost functional

$$\mathcal{J}_1(\boldsymbol{u}(\Omega)) = \frac{1}{2}\int_{\Gamma_o}(\boldsymbol{u}\cdot\boldsymbol{n}-\bar{u})^2 \quad \text{with } \bar{u} = \frac{1}{|\Gamma_o|}\int_{\Gamma_i} -\boldsymbol{g}\cdot\boldsymbol{n}$$

is used. To minimize the total pressure loss, we use

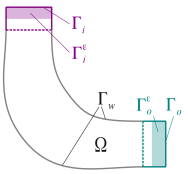
$$\mathcal{J}_2(\boldsymbol{u}(\Omega)) = -\frac{|\Gamma_i|}{|\Gamma_i^\epsilon|}\int_{\Gamma_i^\epsilon}\left(p+\frac{1}{2}|\boldsymbol{u}|^2\right)\boldsymbol{u}\cdot\boldsymbol{n} - \frac{|\Gamma_o|}{|\Gamma_o^\epsilon|}\int_{\Gamma_o^\epsilon}\left(p+\frac{1}{2}|\boldsymbol{u}|^2\right)\boldsymbol{u}\cdot\boldsymbol{n}.$$

The mixed cost functional then has the form

$$\mathcal{J}_{12}(\boldsymbol{u}(\Omega)) = (1-\gamma)\mathcal{J}_1(\boldsymbol{u}(\Omega)) + \gamma\rho\mathcal{J}_2(\boldsymbol{u}(\Omega)),$$

with weighting parameter $\gamma \in [0,1]$ and

$$\rho = \begin{cases} \frac{\|\partial\mathcal{J}_1(\boldsymbol{u}(\Omega^0))\|_{L^2(\Gamma_w^0)}}{\|\partial\mathcal{J}_2(\boldsymbol{u}(\Omega^0))\|_{L^2(\Gamma_w^0)}} & \text{if } \gamma\in(0,1), \\ 1 & \text{if } \gamma\in\{0,1\}. \end{cases}$$



10.4.3. Adjoint equation

The adjoint method allows us to obtain the shape derivative without computing the material derivative of the state. The adjoint state and pressure $(\boldsymbol{v}, q)$ are calculated as the solution of the adjoint equations

$$\begin{aligned} -\nu\Delta\boldsymbol{v} - (\nabla\boldsymbol{v})^T\cdot\boldsymbol{u} - \nabla\boldsymbol{v}\cdot\boldsymbol{u} + \nabla q &= \gamma\nu k_\epsilon\left[(\boldsymbol{u}\cdot\boldsymbol{n})\boldsymbol{u} + \left(p+\frac{1}{2}|\boldsymbol{u}|^2\right)\boldsymbol{n}\right] \text{ in } \Omega, \\ \nabla\cdot\boldsymbol{v} &= -\gamma\nu k_\epsilon\boldsymbol{u}\cdot\boldsymbol{n} \quad \text{in } \Omega, \\ \boldsymbol{v} &= 0 \quad \text{on } \Gamma_i\cup\Gamma_w, \\ -\nu\partial_n\boldsymbol{v} - \boldsymbol{n}(\boldsymbol{u}\cdot\boldsymbol{v}) - (\boldsymbol{u}\cdot\boldsymbol{n})\boldsymbol{v} + q\boldsymbol{n} &= -(1-\gamma)\nu(\boldsymbol{u}\cdot\boldsymbol{n}-\bar{u})\boldsymbol{n} \quad \text{on } \Gamma_o \end{aligned}$$

with $k_\epsilon(x) = \begin{cases} -\frac{|\Gamma_i|}{|\Gamma_i^\epsilon|} & \text{if } x\in\Gamma_i^\epsilon, \\ -\frac{|\Gamma_o|}{|\Gamma_o^\epsilon|} & \text{if } x\in\Gamma_o^\epsilon, \\ 0 & \text{else.} \end{cases}$

The Euler-semiderivative is then obtained by

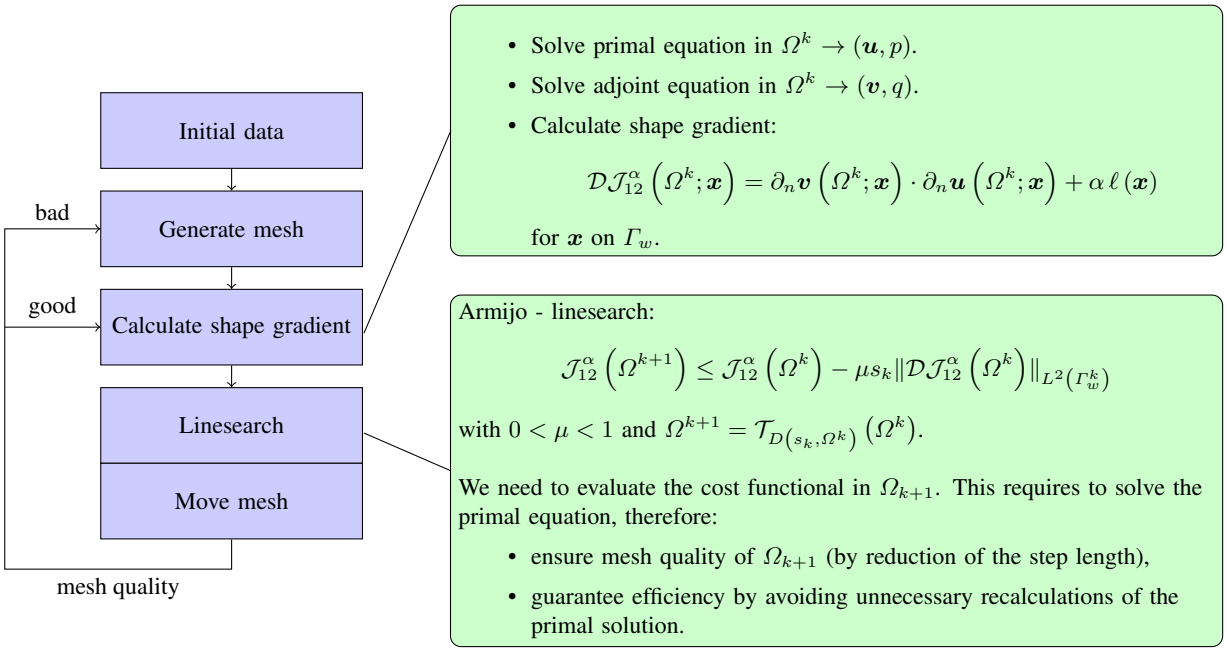
$$\partial\mathcal{J}_{12}^\gamma(\Omega;V) = \int_{\Gamma_w}[\partial_n\boldsymbol{v}\cdot\partial_n\boldsymbol{u}]\,V\cdot\boldsymbol{n}$$

and the shape gradient by

$$D\mathcal{J}_{12}^\gamma(\Omega) = (\partial_n\boldsymbol{v}\cdot\partial_n\boldsymbol{u})\,|_{\Gamma_w}.$$

10.4.4. Descent algorithm

The optimization is done with a gradient descent algorithm which uses an Armijo linesearch. This algorithm is illustrated in the following diagram.



10.4.5. Geometrical constraints

Further restrictions on the possible design are imposed by geometrical constraints which are included via a barrier or a penalty method. Figure 10.2 illustrates the constraint imposed by the design space.

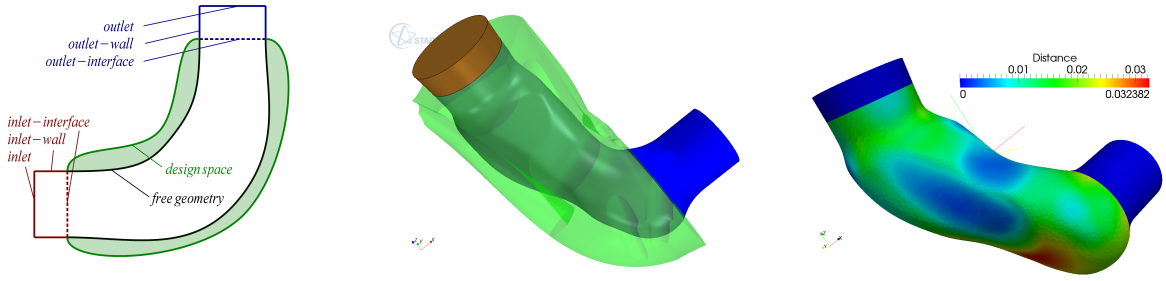


Figure 10.2: Left: sketch of geometry and design space. Middle: geometry with transparent design space. Right: geometry with distance values to the design space.

The minimization problem with geometrical constraint

$$\text{minimize } \mathcal{J}_{12}\left(\boldsymbol{u}\left(\Omega\right)\right) \qquad \text{s.t.} \quad \Omega \subset K$$

can be reformulated as

$$\text{minimize } \mathcal{J}_{12}\left(\boldsymbol{u}\left(\Omega\right)\right) + \alpha\mathcal{L}\left(\Omega\right)$$

with $\alpha > 0$ and $\mathcal{L}\left(\Omega\right) = \int_{\Omega} \ell\left(\Omega\right)$. In the case of the barrier method, we are using

$$\ell\left(\Omega\right) = \left|\ln d\left(x, K^c\right)\right|,$$

and in the case of the penalty method

$$\ell\left(\Omega\right) = \left(d\left(x, K\right)\right)^{\beta},$$

with $\beta \geq 1$, $K^c = \mathbb{R}^n \setminus K$ and the distance function $d\left(x, K\right) = \min\limits_{y \in K} |x - y|$.

The sensitivity is calculated by

$$\partial\mathcal{L}(\Omega, V) = \int_{\Omega} \ell'(\Omega, V)\, dx + \int_{\Gamma} \ell(\Omega) \langle V(0), \boldsymbol{n} \rangle = \int_{\Gamma} \ell(\Omega) \langle V(0), \boldsymbol{n} \rangle .$$

To avoid kinks between the fixed geometry and the geometry which needs to be optimized (see Figure 10.3), we add a penalty functional (green) to the cost functional

$$\mathcal{J}_{12}^{\alpha,\varphi}(\boldsymbol{u}(\Omega)) = \mathcal{J}_{12}(\boldsymbol{u}(\Omega)) + \alpha\mathcal{L}(\Omega) + \varphi\mathcal{L}_F(\Omega)$$

with $\mathcal{L}_F(\Omega) = \int_{\Omega}(d(x, K_F))^2$, weight φ, distance function d and valid domain K_F.
The sensitivity is then calculated by

$$\partial\mathcal{L}_F(\Omega, V) = \int_{\Gamma} \ell_F(\Omega) \langle V(0), \boldsymbol{n} \rangle \quad \text{with } \ell_F = (d(x, K_F))^2.$$

The displacement of the mesh is based on the negative shape gradient, which is shown in the left of Figure 10.4. It consists of the part which is related to the achievement of a uniform outflow and the minimization of the total pressure loss (Figure 10.4, middle) and the part which belongs to the penalty function of the geometrical constraints (Figure 10.4, right).

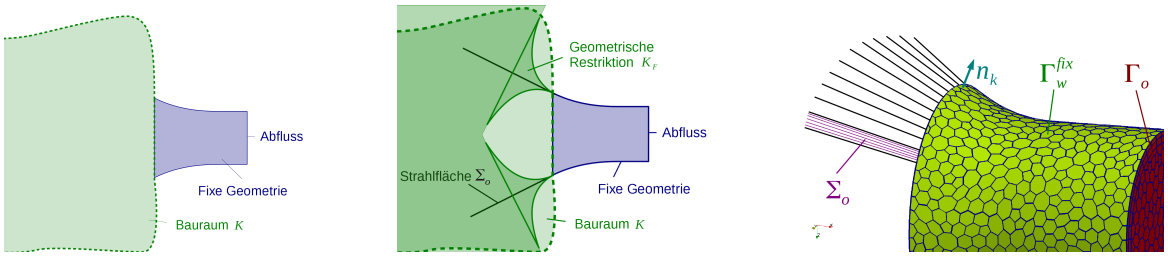


Figure 10.3: Left: fixed geometry (blue) and design space (green). Middle: additional valid domain. Right: continuation of the fixed geometry.

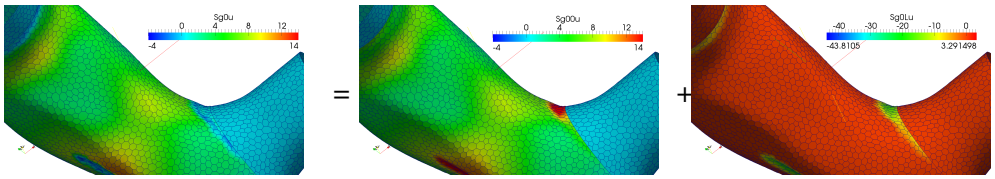


Figure 10.4: Left: negative shape gradient after 20 iterations. Middle: part of the shape gradient related to uniform outflow and minimization of total pressure loss. Right: part of the shape gradient related to geometrical constraints.

10.4.6. Turbulence modeling for high Reynolds numbers

The turbulent flow in applications with high Reynolds numbers (Re $= 200,000$) is modeled with a Reynolds-averaged Navier–Stokes (RANS) approach. We use a k-ϵ turbulence model which describes the evolution of the turbulent kinetic energy k and the rate of dissipation of the turbulent energy ϵ.

The kinematic viscosity ν in the Navier-Stokes equations is replaced by the effective viscosity, which is the sum of kinematic and turbulent viscosity:

$$\nu_{eff} = \nu + \nu_t.$$

Instead of the convection equation

$$\boldsymbol{u}_t + \nabla \cdot \left(\boldsymbol{u}\boldsymbol{u}^T\right) - \nu \Delta \boldsymbol{u} + \nabla p = \boldsymbol{z},$$

the equation

$$\boldsymbol{u}_t + \nabla \cdot \left(\boldsymbol{u}\boldsymbol{u}^T\right) - \nabla \cdot \left(\left(\nu + \nu_t\right) \nabla \boldsymbol{u}\right) + \nabla \left(p + \frac{2}{3}k\right) = \boldsymbol{z}$$

is solved with $\nu_t = C_\nu \frac{k^2}{\varepsilon}$. The kinetic energy k and the dissipation ε are solutions of

$$\begin{aligned}
\partial_t k &= -\left(\boldsymbol{u} \cdot \nabla\right) k + \nabla \cdot \left(\left(\nu + \nu_t\right) \nabla k\right) + \nu_t \left|\nabla \boldsymbol{u} + \nabla \boldsymbol{u}^T\right|^2 - \varepsilon \\
\partial_t \varepsilon &= -\left(\boldsymbol{u} \cdot \nabla\right) \varepsilon + \nabla \cdot \left(\left(\nu + \frac{\nu_t}{\sigma_\varepsilon}\right) \nabla \varepsilon\right) + C_1 \frac{\varepsilon}{k} \nu_t \left|\nabla \boldsymbol{u} + \nabla \boldsymbol{u}^T\right|^2 - C_2 \frac{\varepsilon^2}{k}
\end{aligned}$$

with standard model constants $C_\nu = 0.09$, $C_1 = 1.44$, $C_2 = 1.92$, $\sigma_\varepsilon = 1.3$.

Figure 10.5 illustrates the change of geometry during the shape optimization for a turbulent flow with Reynolds number $\mathrm{Re} = 200,000$.

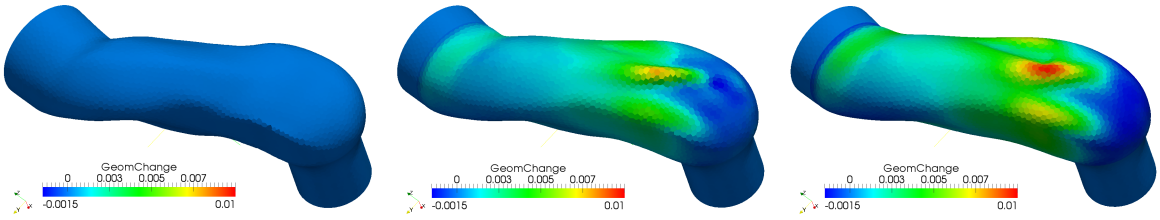


Figure 10.5: Change of geometry in the beginning (left), after 20 (middle) and after 70 iterations (right).

10.4.7. Laplace-Beltrami smoothing

A Laplace-Beltrami smoothing is applied by using the following equation on the surface Γ:

$$\begin{aligned}
-\varepsilon \Delta_\Gamma w + w &= -g_{\mathcal{J}_{12}} \qquad \text{on} \quad \Gamma, \\
w &= 0 \qquad\qquad\quad \text{on} \quad \partial\Gamma.
\end{aligned}$$

The Laplace-Beltrami smoothing realizes a preconditioning of the gradient method. It permits a higher regularity of each grid movement while keeping the property of a descent direction. Figure 10.6 shows results without and with the Laplace-Beltrami smoothing for two different geometries.

10.4.8. Shape optimization tested on various geometries

Figure 10.7 shows the change of geometry after the shape optimization for four different examples.

10.5. Description of the publicly available data

The developed shape optimization tool should be tested on the benchmark geometry which will be introduced in the following. The proposed geometry has been developed together with engineers from the automotive industry and presents an example which is relevant for real-life applications in this field.

The geometry describes an air duct in a combustion engine and is shown from different perspectives in Figure 10.8. It consists of one inlet Γ_i and one outlet Γ_o. The parts next to the inlet and to the outlet, which are called inlet-wall and outlet-wall, are fixed. These parts are connected with the free geometry that shall be optimized.

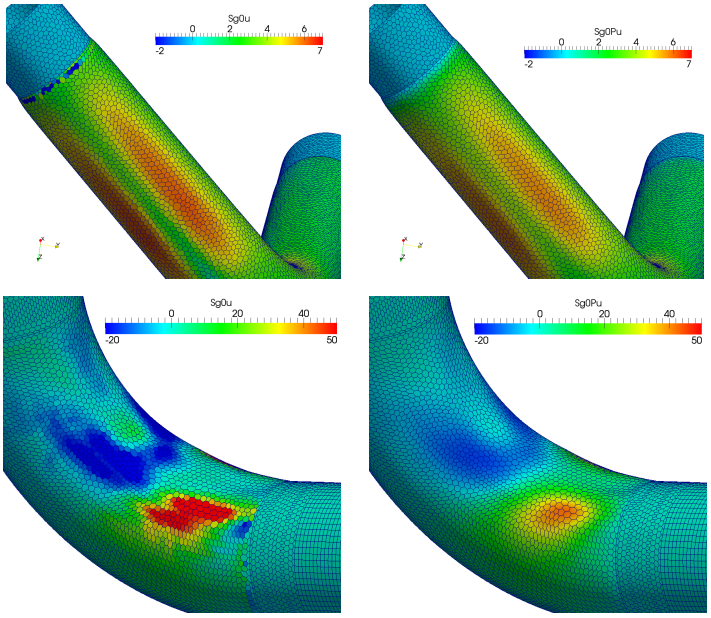


Figure 10.6: Left: Shape gradient without LB-smoothing. Right: Shape gradient with LB-smoothing.

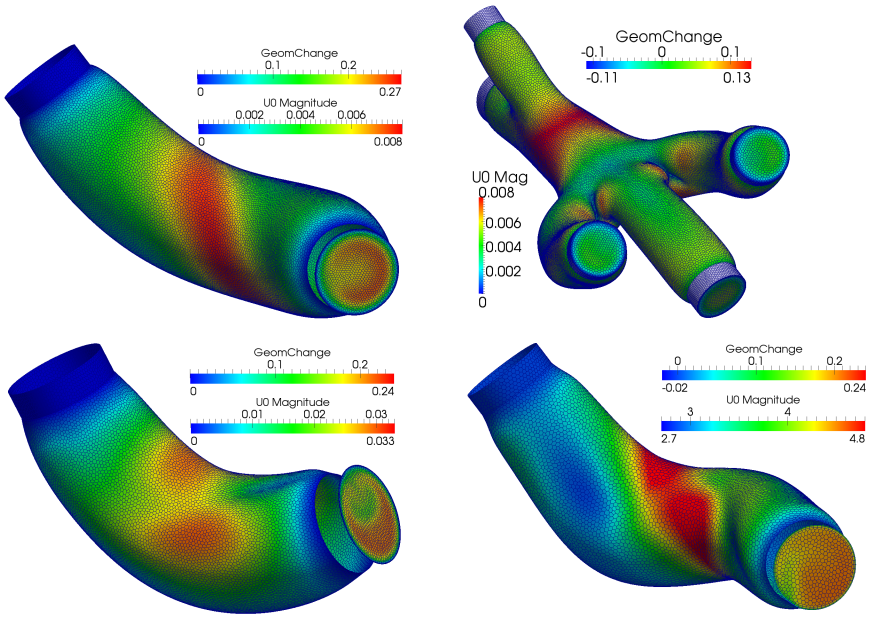


Figure 10.7: Final geometries: Upper: Re = 200, Re = 400; Lower: Re = 1000, Re = 200 000

The caption of Figure 10.8 identifies the constituent parts of the geometry.

The whole geometry is inscribed in a design space which should be respected as an outer limit of the shape variation. The design space is illustrated in Figure 10.9. Figure 10.10 shows the geometry with transparent design space.

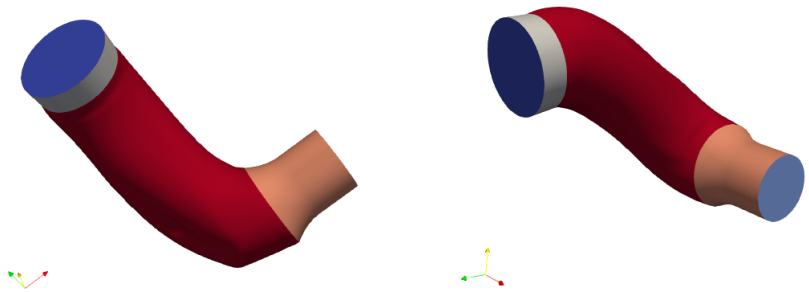


Figure 10.8: The proposed benchmark geometry (from two different perspectives) consists of an inlet (dark blue), a fixed inlet-wall (gray), an outlet (light blue), a fixed outlet-wall (orange) and a wall whose shape shall be optimized (red).

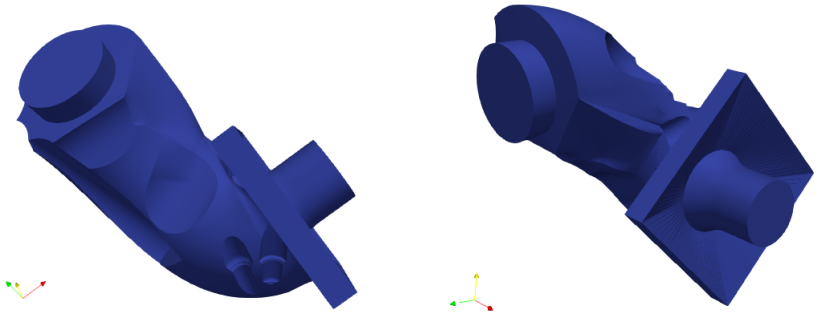


Figure 10.9: The design space which constrains the possible shape of the air duct (from two different perspectives).

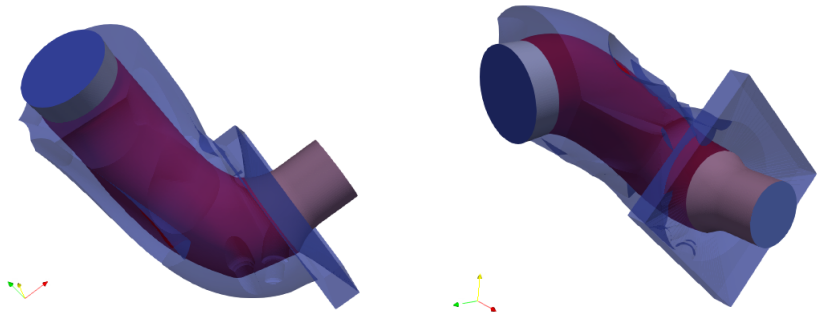


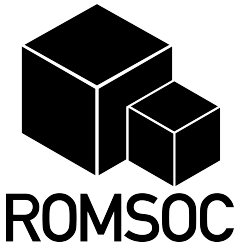
Figure 10.10: The geometry of the air duct inscribed in the design space (from two different perspectives).

Bibliography

[1] S. Boisgérault and J.-P. Zolésio, "Shape derivative of sharp functionals governed by navier-stokes flow," *Partial differential equations (Praha, 1998)*, vol. 406, pp. 49–63, 1999.

[2] A. Laurain and K. Sturm, "Distributed shape derivative via averaged adjoint method and applications," *ESAIM: Mathematical Modelling and Numerical Analysis*, vol. 50, no. 4, pp. 1241–1267, 2016.

[3] M. Fischer, F. Lindemann, M. Ulbrich, and S. Ulbrich, "Fréchet differentiability of unsteady incompressible navier–stokes flow with respect to domain variations of low regularity by using a general analytical framework," *SIAM Journal on Control and Optimization*, vol. 55, no. 5, pp. 3226–3257, 2017.



The ROMSOC project

June 8, 2020

ROMSOC-D5.1-3.0

Horizon 2020