U. PORTO

FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Master's Thesis in Mechanical Engineering on

Modeling and analysis of rough contact by computational homogenization

António Manuel Couto Carneiro

Advisor:
Prof. Francisco Manuel Andrade Pires, Ph.D.

Co-advisor:
Rodrigo Pinto Carvalho, M.Sc.

Dissertation submitted to the Department of Mechanical Engineering of the Faculty of Engineering the University of Porto in fulfillment of the requirements for the degree of Master in Mechanical Engineering

Porto, July 2019

CM2S

Computational Multi-Scale
Modeling of Solids and Structures

Para o meu irmão Rafael

Page intentionally left blank

Agradecimentos

A realização da presente dissertação contou com diversas contribuições que direta, ou indiretamente, permitiram alcançar os respetivos objetivos, traçados no início deste período. Refiro-me não só àqueles que colaboraram de uma forma muito próxima no trabalho, mas também a todos os que me apoiaram ao longo de todo o meu percurso académico e pessoal. Espero que de alguma forma tenha já expressado o meu louvor a todos eles, que seguramente não se cingem aos que são mencionados aqui.

Em primeiro lugar, uma gratificação muito especial é devida ao Prof. Francisco Manuel Andrade Pires, orientador desta dissertação, pelo rigor científico, organização e visão na sua condução, e ainda por toda a motivação e reconhecimento do meu esforço. Estarei para sempre grato pela a oportunidade de trabalhar com um profissional de excelência, que tem sido uma fonte de inspiração tanto a nível profissional como pessoal.

Em segundo lugar, ao coorientador da presente dissertação, o Eng. Rodrigo Pinto Carvalho, pela imensa disponibilidade e paciência dedicadas a este trabalho, bem como pelas sugestões e ajuda inestimáveis. A sua contribuição foi certamente insubstituível para a conclusão da dissertação no tempo estabelecido. Ainda, um agradecimento sentido lhe é devido pela sua amizade e simpatia constantes.

Uma palavra de apreço a toda o grupo CM2S, por proporcionarem um ambiente de investigação estimulante, com elevado nível científico e espírito de equipa. Quero também agradecer ao meu colega de curso e amigo Rui Coelho, que me acompanhou durante toda a dissertação. Todo o seu apoio contribuiu, seguramente, para facilitar a realização deste trabalho.

Ao CETRIB, em particular, ao Dr. Carlos Miguel da Costa Gomes Fernandes, pelo tempo e dedicação no esclarecimento de tópicos sobre a medição experimental de rugosidade, bem como pelos resultados experimentais cedidos.

Quero também reconhecer o projeto de investigação *NORTE-01-0145-FEDER-000022 SciTech - Science and Technology for Competitive and Sustainable Industries - Research Line 1: Advanced Materials & Structures*, financiado pelo Portugal 2020, por assegurar o financiamento deste trabalho.

A toda a minha família, um obrigado de todo o coração. Aos meus avós, António e Maria Emília, por me terem ensinado o valor do trabalho árduo, e pelo amor e simpatia mais puros. Aos meus pais, Manuel e Maria Ângela, pela minha educação enquanto pessoa, e pelo suporte inesgotável durante toda a minha vida. Ao meu irmão Rafael, por estar sempre presente quando preciso e para o que for preciso. A todos eles, um muito obrigado por todos os sacrifícios, e por me ajudarem constantemente a crescer.

À Cláudia, o amor da minha vida. Pela paciência nestes últimos meses, pelo seu carinho e amor incessáveis, e por representar tudo o que me motiva a trabalhar e evoluir, para tentar ser melhor a cada dia.

O meu sincero apreço a todos,

António Manuel Couto Carneiro

God made the bulk;
the surface was invented by the devil.

Wolfgang Pauli

Page intentionally left blank

Abstract

Modeling and analysis of rough contact by computational homogenization

Keywords: Rough contact; Rough surface generation; Computational contact homogenization; Multiscale modeling.

In typical theoretical and computational contact mechanics problems, it is often tacitly assumed that the boundaries of contacting bodies are smooth. However, it can be regarded as today's engineering common-sense that all surfaces are rough at some length scale. A numerical discretization accounting for all the details involved in a rough surface, usually spanning several length scales, would quickly render the numerical model excessively heavy. Multiscale approaches, based on contact homogenization techniques, have been proposed in the last years, in order to model several roughness length scales, while reducing the total computation time, in comparison with the direct numerical approach.

In this work, the elastic, non-adhesive and frictionless contact between a Gaussian self-affine topography and a rigid and flat plane is modeled within a single and multiscale finite element method framework, coupled with the dual mortar contact discretization. The numerical framework starts with the generation of randomly rough topographies, that reproduces any given input Power Spectral Density (PSD). Single scale 2D simulations are first performed, in order to define a statistically Representative Contact Element (RCE). Rules of thumb for mesh spacing, length and height of the RCE, and also the number of topography realizations are established. These conditions are, then, embedded within a multiscale framework, defining a contact problem for each involved scale. The original topography is divided into distinct scales, by introducing several splitting frequencies in the PSD. The multiscale solution for the real contact area fraction is computed in a multiplicative homogenization step, by multiplying the results of each scale, which are obtained independently of each other.

In comparison with the single scale response, the multiscale solution provides similar results over a wide roughness spectra, while benefiting from a very attractive computational cost. Moreover, an improved homogenization scheme is proposed, which aims at better capturing the influence of the contact pressures distribution throughout the entire load range. This new approach relies on a weighted average multiplicative scheme, showing to provide accurate results for both light (low pressure) and nearly full contact, independently of the number of scales considered in the PSD splitting. By employing this technique, novel results for 2D problems with extremely wide spectra and complex 3D rough contact problems are obtained, with short simulation time and memory usage.

Page intentionally left blank

Resumo

Modelação e análise de contacto rugoso através de homogeneização computacional

Palavras-Chave: Contacto rugoso; Geração de superfícies rugosas; Homogeneização computacional de contacto; Modelos multi-escala.

Em problemas típicos de mecânica do contacto teórica e computacional, é comum assumir implicitamente que as fronteiras dos corpos em contacta são lisas. Contudo, hoje em dia, é senso-comum em engenharia que todas as superfícies são rugosas a alguma escala de observação. Uma discretização numérica que inclua todo os detalhes de uma superfície rugosa, que com frequência existem ao longo de várias ampliações, rapidamente torna o modelo numérico excessivamente pesado. Várias estratégias multi-escala baseadas em homogeneização de contacto têm vindo a ser propostas nos últimos anos, com o intuito de modelar várias escalas de rugosidade, consumindo menos tempo de cálculo em comparação com a via numérica direta.

Neste trabalho, o contacto elástico, não adesivo e sem atrito entre uma topografia Gaussiana auto-similar (*self-affine*) e um plano rígido é modelado num contexto de estratégias multi-escala e a uma única escala, através do método dos elementos finitos, com a formulação dual mortar para a discretização do contacto. A estratégia numérica parte da geração de topografias rugosas aleatórias, que verificam qualquer Densidade de Potência Espectral (DPE) (*Power Spectral Density*) pretendida, de modo a se definir um Elemento de Contacto Representativo (ECR). Regras de ouro para a malha, comprimento e altura do ECR, e também para o número de ECRs são estabelecidas. Após, estas condições são incluídas na estratégia multi-escala, para a definição do problema de contacto a cada escala. A topografia original é separada em diferentes escalas, através da introdução de várias frequências de divisão na DPE (PSD). A solução multi-escala para a fração da área real de contacto é calculada num passo de homogeneização multiplicativa, através da multiplicação dos resultados a cada escala, que são obtidos de forma independente.

Em comparação com a estratégia a uma única escala, a solução multi-escala é idêntica para espetros de rugosidade amplos, beneficiando de tempos de cálculo muito atrativos. Adicionalmente, é proposta uma estratégia de homogeneização melhorada, com o objetivo de incorporar a influência da distribuição das pressões de contacto ao longo de todo o carregamento com mais informação. Esta nova abordagem é baseada num esquema multiplicativo com médias ponderadas, produzindo resultados precisos tanto para contacto infinitesimal como completo, independentemente do número de escalas consideradas. Utilizando esta técnica, resultados inovadores para problemas em 2D com espetros extremamente largos e ainda problemas complexos de contacto rugoso em 3D são obtidos, com tempos de simulação curtos e baixa utilização de memória.

Page intentionally left blank

Contents

List of Figures

Chapter 1
Introduction

Chapter 2
On rough surface characterization

Chapter 3
Numerical generation of randomly rough topography

Chapter 4
Micromechanical elastic contact: analytical models

Chapter 5
Single scale dual mortar finite element modeling of rough contact

Chapter 6
Multiscale finite element modeling of rough contact by contact homogenization

Chapter A
Notes on Fourier transforms

Page intentionally left blank

List of Tables

Chapter 3
Numerical generation of randomly rough topography

Chapter 5
Single scale dual mortar finite element modeling of rough contact

Chapter 6
Multiscale finite element modeling of rough contact by contact homogenization

Page intentionally left blank

Nomenclature

General abbreviations

2D	**Two D**imensional
3D	**Three D**imensional
ACF	**Autoc**orrelation **F**unction
ACL	**Autoc**orrelation **L**ength
AR	**Autor**egressive
ARMA	**Autor**egressive **M**oving **A**verage
API	**A**pplication **P**rogramming **I**nterface
BEM	**B**oundary **E**lement **M**ethod
BGT	**B**ush-**G**ibson-**T**homas (model)
BLT	**B**ond **L**ine **T**hickness
CDF	**C**umulative **D**istribution **F**unction
DFT	**D**iscrete **F**ourier **T**ransform
DNS	**D**irect **N**umerical **S**imulation
DTFT	**D**iscrete-**T**ime **F**ourier **T**ransform
EDM	**E**letrical **D**ischarge **M**achining
FE	**F**inite **E**lement
FEM	**F**inite **E**lement **M**ethod
FFT	**F**ast **F**ourier **T**ransform
GFMD	**G**reen's **F**unction **M**olecular **D**ynamics
GW	**G**reenwood-**W**illiamson (model)
GW-SE	**G**reenwood-**W**illiamson **S**implified **E**lliptic (model)
HDC	**H**eight **D**ifference **C**orrelation
HSM	**H**ertz-**S**ignorini-**M**oreau
IBVP	**I**nitial **V**alue **B**oundary **P**roblem
IDFT	**I**nverse **D**iscrete **F**ourier **T**ransform
IDTFT	**I**nverse **D**iscrete-**T**ime **F**ourier **T**ransform
IFFT	**I**nverse **F**ast **F**ourier **T**ransform
KKT	**K**arush-**K**uhn-**T**ucker
LINKS	Large Strain Implicit Nonlinear Analysis of Solids Linking Structures

MA	**M**oving **A**verage
MS	**M**ulti**s**cale
NCGM	**N**onlinear **C**onjugate **G**radient **M**ethod
NCP	**N**onlinear **C**om**p**lementarity
NTS	**N**ode-**t**o-**S**egment
PDASS	**P**rimal-**D**ual **A**ctive **S**et **S**trategy
PDF	**P**robability **D**ensity **F**unction
PSD	**P**ower **S**pectral **D**ensity
PVW	**P**rinciple of **V**irtual **W**ork
RAM	**R**andom **A**ccess **M**emory
RCE	**R**epresentative **C**ontact **E**lement
RMS	**R**oot **M**ean **S**quare
RSSE	**R**epresentative **S**elf-affine **S**urface **E**lement
RVE	**R**epresentative **V**olume **E**lement
STS	**S**egment-**t**o-**S**egment
TIM	**T**hermal **I**nterface **M**aterials

General notation

a, A	Scalar
$\boldsymbol{a}$	Vector
$\boldsymbol{A}$	Second-order tensor
$\mathbf{A}$	Matrix
$\mathscr{A}$	Fourth-order tensor
$f(\bullet)$	Continuous function
$f[\bullet]$	Discrete function

Operators

$\frac{\partial^n(\bullet)}{\partial a^n}$	Partial derivative of order n relative to a
$\frac{\mathrm{d}^n(\bullet)}{\mathrm{d}a^n}$	Total derivative of order n relative to a
det	Determinant of a second-order tensor
$\mathrm{div}(\bullet)$	Divergence of a tensor
$\mathrm{erf}(\bullet)$	Error function
$\mathrm{erfc}(\bullet)$	Complementary error function
$\mathrm{DFT}(\bullet)$	Discrete Fourier transform
$\mathcal{F}\{\bullet\}$	Fourier transform
$\mathcal{F}^{-1}\{\bullet\}$	Inverse Fourier transform
$\mathrm{FFT}(\bullet)$	Fast Fourier transform
$\boldsymbol{I}$	Second-order identity tensor
i	Imaginary number ($\mathrm{i} = \sqrt{-1}$)

$\mathrm{IDFT}(\bullet)$	Inverse discrete Fourier transform
$\mathrm{IFFT}(\bullet)$	Inverse fast Fourier transform
$\max\{\bullet,\bullet\}$	Maximum operator
$\Pr(\bullet)$	Probability
$\delta(\bullet)$	Dirac Delta function
δ_{ij}	Kronecker delta
$\nabla(\bullet)$	Gradient operator
$\nabla^2(\bullet)$	Laplacian operator
$\nabla_x(\bullet)$	Spatial gradient operator
$\emptyset$	Empty set
$\mathbf{0}$	Zero tensor
$(\bullet)^*$	Complex conjugate
$(\bullet) * (\bullet)$	Linear convolution
$(\bullet) \circledast (\bullet)$	Circular convolution
$(\bullet) \star (\bullet)$	Linear correlation
$(\bullet) \circledstar (\bullet)$	Circular correlation
$\lVert \bullet \rVert$	Euclidean vector norm
$\overline{(\bullet)}$	Spatial average
$\langle \bullet \rangle$	Ensemble average
$\angle(\bullet)$	Argument of complex number
$\lvert \bullet \rvert$	Magnitude of complex number / Absolute value operator
$(\bullet) \cdot (\bullet)$	Dot product
$(\bullet) : (\bullet)$	Tensor double contraction
$(\bullet)^{\mathrm{T}}$	Transpose of a tensor
$(\bullet)^{-1}$	Inverse of a tensor
$(\bullet)^{-\mathrm{T}}$	Inverse of the transpose of a tensor
$(\dot{\bullet})$	Total time derivative
$(\bullet) \otimes (\bullet)$	Dyadic product
$(\bullet) \setminus (\bullet)$	Set difference
$(\bullet) \times (\bullet)$	Cartesian product

Subscripts

$(\bullet)_{\mathrm{c}}$	Contact
$(\bullet)_{\mathrm{ext}}$	Exterior
$(\bullet)_{\mathrm{ext}}$	External
$(\bullet)_{\mathrm{fit}}$	Numerical fit
$(\bullet)_{\mathrm{fix}}$	Fixed
$(\bullet)_{\mathrm{int}}$	Internal
$(\bullet)_{\mathrm{max}}$	Maximum value
$(\bullet)_{\mathrm{nyq}}$	Nyquist (frequency)

$(\bullet)_{\mathrm{rms}}$	Root mean square
$(\bullet)_{\mathrm{split}}$	Split
$(\bullet)_x$	Relative to the x direction
$(\bullet)_y$	Relative to the y direction
$(\bullet)_+$	Quantity in the positive periodic boundary
$(\bullet)_-$	Quantity in the negative periodic boundary

Superscripts

$(\bullet)^h$	Discretized version of $(\bullet)$ with the FEM
$(\bullet)^{\mathrm{m}}$	Mortar
$(\bullet)^{\mathrm{MS}}$	Multiscale
$(\bullet)^{\mathrm{s}}$	Non-mortar
$(\bullet)^{\eta}$	Normal direction
$(\bullet)^{\tau}$	Tangential direction

Domains and boundaries

$\mathbb{Z}$	Integer set
$\mathbb{R}$	Real set
Ω_e	Finite element sub-domain
Ω_0	Domain of a body in the reference configuration
Ω_t	Domain of a body in the current configuration
$\partial\Omega_0$	Boundary of a body in the reference configuration
$\partial\Omega_t$	Boundary of a body in the current configuration
Γ_σ	Neumann boundary in the reference configuration
Γ_u	Dirichlet boundary in the reference configuration
γ_σ	Neumann boundary in the current configuration
γ_u	Dirichlet boundary in the current configuration
Γ_{c}	Potential contact boundary in the reference configuration
γ_{c}	Potential contact boundary in the current configuration
Γ_{a}	Active contact boundary in the reference configuration
$\partial\Omega_{\mathsf{ext}}$	Exterior boundary
$\partial\Omega_{\mathsf{fix}}$	Fixed boundary
$\partial\Omega_+$	Positive periodic boundary
$\partial\Omega_-$	Negative periodic boundary

Roughness Parameters

$z_{\mathrm{rms},x}$	Root mean square roughness/height of a continuous profile
$z_{\mathrm{rms},xy}$	Root mean square roughness/height of a continuous surface

$z'_{\mathrm{rms},x}$	Root mean square slope of a continuous profile
$z'_{\mathrm{rms},xy}$	Root mean square slope of a continuous surface
$z''_{\mathrm{rms},x}$	Root mean square curvature of a continuous profile
$z''_{\mathrm{rms},xy}$	Root mean square mean curvature of a continuous surface
R_q	Root mean square roughness of a discrete profile
S_q	Root mean square roughness of a discrete surface
$R_{\Delta q}$	Root mean square slope of a discrete profile
$S_{\Delta q}$	Root mean square slope of a discrete surface
$R_{\Delta^2 q}$	Root mean square curvature of a discrete profile
$S_{\Delta^2 q}$	Root mean square curvature of a discrete surface
R_{sk}	Discrete profile skewness
S_{sk}	Discrete surface skewness
R_{ku}	Discrete profile kurtosis
S_{ku}	Discrete surface kurtosis

Roughness characterization and random processes

b	Frequency scale factor in the weak anisotropy power spectrum
C_0	Scale factor of the PSD of a self-affine surface
$\hat{C}_0$	Scale factor of the PSD of a discrete self-affine surface
C'_0	Scale factor of the PSD of a self-affine profile
$\hat{C}'_0$	Scale factor of the PSD of a discrete self-affine surface
D_p	Profile fractal dimension
D_s	Surface fractal dimension
e	Error of form
f_Z	Probability density function of topography heights
F_Z	Cumulative distribution function of topography heights
G	Profile fractal scale factor
g	Surface fractal scale factor
H	Hurst roughness exponent
h	Topography height
J	Profile-surface fractal scale factor in the strong anisotropy power spectrum
k	Wavenumber/spatial frequency
$\boldsymbol{k}$	Wavevector
k_l	Low frequency cut-off
k_r	Roll-off frequency
k_s	High frequency cut-off
Ω_s	Sampling frequency
L	Domain length
l_s	Sampling wavelength, or sampling interval
$m_{\theta n}$	Profile spectral moment of order n

m_n	Profile spectral moment of order n, from an isotropic rough surface
m_{mn}	Surface spectral moment of order mn
R	Autocorrelation function
$\widetilde{R}$	Discrete circular autocorrelation function
$\hat{R}$	Discrete estimate of autocorrelation function
w	Waviness profile
z	Roughness profile
$\mathscr{Z}$	Random variable associated with roughness height
α	Nayak's parameter or spectrum breadth
β	Autocorrelation length
β_2	Kurtosis of the heights distribution
γ_1	Skewness of the heights distribution
γ_2	Excess of kurtosis of the heights distribution
ζ	Ratio between the roll-off and short cut-off wavelength
λ	Wavelength
λ_l	Long wavelength cut-off
λ_r	Roll-off wavelength
λ_s	Short wavelength cut-off
ξ	Ratio between the roll-off and large cut-off wavelength
μ_i	Central moment of order i of the probability density function
μ_z	Mean height
σ_z	Standard deviation of heights
τ	Position shift
Φ	Surface power spectral density
Φ_θ	Profile power spectral density
$\hat{\Phi}$	Discrete estimate of surface PSD
$\hat{\Phi}^\theta$	Discrete estimate of profile PSD
ϕ	Phase

Numerical generation of rough topography

A	DFT of the input white noise
a, b	Coefficients of the general ARMA model
H	DFT of the filter coefficients / Transfer function of the system
h	Filter coefficients
L	Periodic length of the generated topography
l_s	Sampling length
M, N	Number of sampling points in y and x direction, respectively
Z	DFT of the topography height
z	Surface/profile height
η	White noise signal

η'	White noise transformed by Hill's algorithm
η''	White noise with phases imposed to ϕ'
χ	Auxiliary variable
ϕ	Random phase
ϕ'	Random phase associated with η'

Micromechanical contact

A	Nominal contact area
A_c	Real contact area
a,b	Semi-axis of the elliptical contact area
$\mathscr{D}_{\mathrm{sum}}$	Density of summits per unit area
d	Separation
$\hat{d}$	Separation non-dimensionalized by σ_s
E	Young modulus
E^*	Effective Young modulus
F	Normal load
g	Dimensionless curvature
$ƙ$	Coefficient of proportionality between real contact area fraction and nominal pressure non-dimensionalized by the RMS slope and E^*
N_{sum}	Number of summits in the surface
R	Radius of curvature of summits
$P(p,\zeta)$	Probability density function of contact pressure p at magnification ζ
p	Contact pressure
p_0	Nominal exterior pressure
p_m	Mean contact pressure
t	Dimensionless separation non-dimensionalized by σ_z
z_s	Summit height
$\hat{z}_s$	Summit height non-dimensionalized by σ_s
$\mathscr{Z}_s$	Random variable associated with summit height
δ	Penetration
κ	Curvature of summits
ν	Poisson's ratio
ζ	Magnification
σ_s	Standard deviation of summit heights
σ_p^2	Variance of contact pressures
Υ	Joint probability density function of summit height and principal curvature, for an isotropic Gaussian surface
φ_{sum}	Probability density function of summit heights
$\hat{\varphi}_{\mathrm{sum}}$	Probability density function of summit heights non-dimensionalized by σ_s

Kinematics

A	Area of the current configuration
A_0	Area of the reference configuration
d	Number of spatial dimensions
$\boldsymbol{e}_1$, $\boldsymbol{e}_2$, $\boldsymbol{e}_3$	Generic base vector of a Cartesian coordinate system
$\boldsymbol{F}$	Deformation gradient
J	Jacobian of the deformation gradient
$\boldsymbol{n}$	Outward unit normal vector in the current configuration
$\boldsymbol{N}$	Outward unit normal vector in the reference configuration
t	Time
$\boldsymbol{u}$	Displacement vector
V	Volume of the current configuration
V_0	Volume of the reference configuration
$\boldsymbol{X}$	Position of a point in the reference configuration
$\boldsymbol{x}$	Position of a point in the current configuration
φ	Deformation map between the reference and current configurations

Strain, stress and constitutive laws

$\boldsymbol{C}$	Right Cauchy-Green strain tensor
$\mathscr{C}$	Fourth-order constitutive tensor
$\boldsymbol{E}$	Green-Lagrange strain tensor
$\boldsymbol{f}$	Force in the current configuration
$\boldsymbol{P}$	First Piola-Kirchoff stress tensor
$\boldsymbol{S}$	Second Piola-Kirchoff stress tensor
$\boldsymbol{\sigma}$	Cauchy stress tensor
Ψ	Strain energy function

Governing equations

$\boldsymbol{b}$	Body forces in the current configuration
H^1	Sobolev vector space
$\mathcal{M}$	Solution space for the Lagrange multiplier vector
m	Mass of a body
T	Total simulation time
t	Time
$\bar{\boldsymbol{t}}$	Prescribed surface tractions in the current configuration
$\mathcal{U}$	Solution space for the displacement field
$\bar{\boldsymbol{u}}$	Prescribed displacements in the current configuration
$\mathcal{V}$	Solution space for the virtual displacement field

$\boldsymbol{t}$	Surface traction in the current configuration
$\delta\boldsymbol{u}$	Virtual displacement field
$\delta\boldsymbol{\lambda}$	Trial surface traction
$\delta\Pi$	Virtual work
$\boldsymbol{\lambda}$	Lagrange multiplier vector
λ^{η}	Normal projection of the Lagrange multiplier vector
$\boldsymbol{\lambda}^{\tau}$	Tangential component Lagrange multiplier vector
ρ	Density in the current configuration
ρ_0	Density in the reference configuration

Contact mechanics

g	Gap function
$\boldsymbol{g}$	Gap vector
$\hat{\boldsymbol{x}}$	Projected point
p^{η}	Contact normal pressure
$\boldsymbol{v}^{\tau}$	Tangential relative velocity
$\boldsymbol{t}_{\mathrm{c}}$	Contact traction vector in the current configuration
$\boldsymbol{t}^{\tau}$	Tangential contact traction in the current configuration
β	Coulomb's friction law parameter
$\boldsymbol{\eta}$	Outward unit normal vector to the non-mortar contact surface
μ	Coefficient of friction
ψ	Coulomb's friction law slip function

Mortar finite element approximation and solution algorithm

C_j^{η}	Nonlinear complementarity function in the normal direction at the non-mortar node j
c^{η}	Complementarity parameter in the normal direction
$\mathbf{D}$	First mortar coupling matrix
D_{jk}	Element from the first mortar coupling matrix
d	Nodal displacement
$\tilde{g}$	Weighted gap
$\mathbf{M}$	Second mortar coupling matrix
M_{jk}	Element from the second mortar coupling matrix
N	Shape function for the displacements and geometry
n^{e}	Number of finite elements
n^{s}	Number of non-mortar nodes
n^{m}	Number of mortar nodes
n^{λ}	Number of non-mortar nodes carrying Lagrange multipliers
x	Nodal coordinates in the current configuration

z	Nodal Lagrange multiplier vector
z	Component of the nodal Lagrange multiplier vector
$\boldsymbol{\xi}$	Point in the element parameter space
Φ	Shape function for the Lagrange multiplier

Numerical model and multiscale approach

$f_{i,n}^{\{j\}}$	Contact pressure probability density of the n-th contact pressure value in scale j for the load increment i
H_{sub}	Height of the rough block
H_{ref}	Height of the refined mesh of the rough block
k_{split}	Splitting frequency
n^{p}	Number of values discretized from the contact pressure distribution
n^{s}	Number of scales
$\bar{p}_i^{\eta\{j\}}$	Mean contact normal pressure at load increment i, in scale j
$p_{i,n}^{\eta\{j\}}$	n-th discrete value of the contact normal pressure at load increment i in scale j
$p_0^{\{j\},i}$	Nominal exterior pressure at load increment i in scale j
$\Delta p_{i,n}^{\eta\{j\}}$	Width of the pressure bins relative to the n-th discrete contact pressure value at scale j and load increment i
Δx	Mesh spacing

Chapter 1

Introduction

This introductory chapter provides a context for the current dissertation, and sets its main objectives. The work was developed on the numerical modeling of rough contact with the *Finite Element Method* (FEM), within a dual mortar contact formulation. Both single scale modeling by *Direct Numerical Simulation* (DNS) and multiscale modeling by contact homogenization were investigated. The problem analyzed in this contribution can be described as the elastic, non-adhesive and frictionless contact between a deformable rough block and a flat rigid surface. The outline of the dissertation is provided at the end of the chapter, in order to facilitate the navigation within the document.

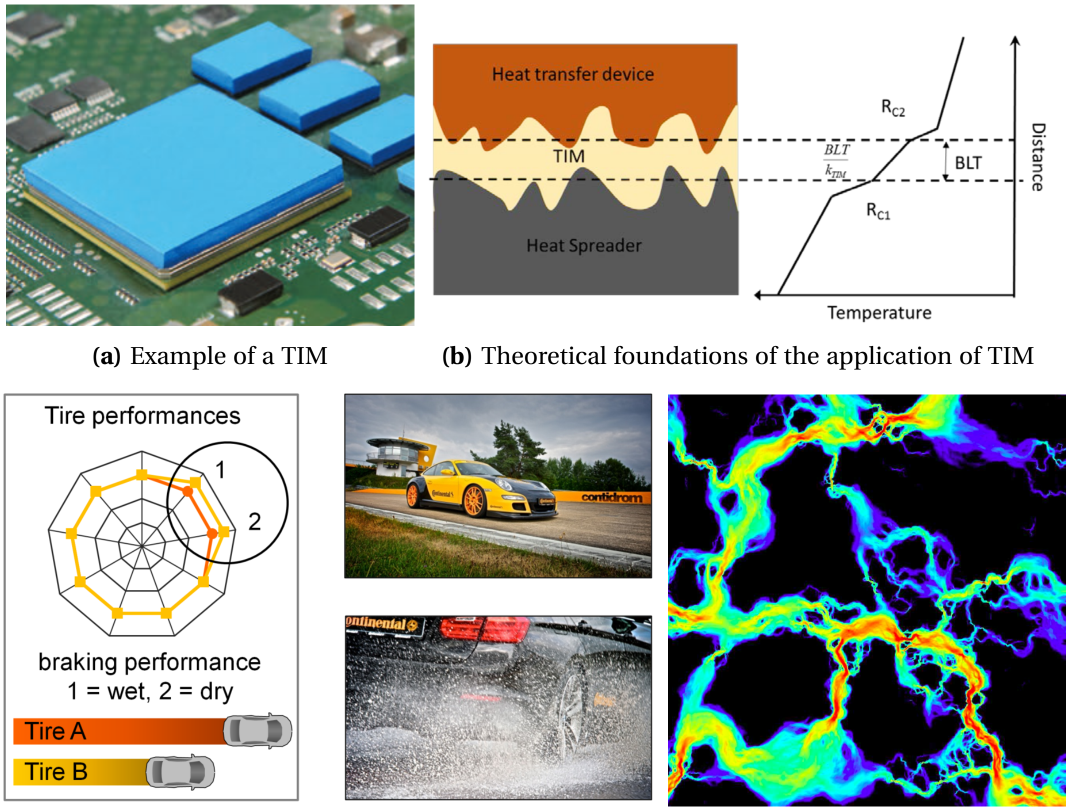
1.1 Motivation

Contact is one of the fundamental ways by which bodies interact. Mechanical loads and energy are generally transfered by contact, together with other physical quantities such as electric current and thermal energy. Contact mechanics problems can be regarded as classical continuum mechanics problems for deformable bodies, with the additional difficulties associated with the complex contact boundary conditions. These specify, e.g., that solid bodies shall not penetrate each other. Apart from the classical geometry and material nonlinearities, contact introduces boundary nonlinearities, since the contact interface, i.e., the partition of the domain where the contact conditions apply, is not known beforehand. Furthermore, in addition to the material bulk properties required for the modeling of solid mechanics problems, contact mechanics requires interface properties, such as the coefficient of friction. These are extremely difficult to predict and depend on several conditions, like the local contact pressure, relative tangential velocity and temperature, just to name a few. Frictional contact plays a major role in the technological and economic landscape of today's society. Tzanakis *et al.* (2012) reports that 1% of the gross national product in several nations comes from frictional loses. Another example is the power dissipation at the tire-road interaction, which makes about 20-30% of the total fuel consumption (Nitsche, 2011). In engineering, the study of contacting interfaces, typically provided with relative motion, is commonly designated *tribology*. The word was introduced by Jost (1966), standing for *the study of* ("logy") *rubbing* ("tribo").

Awareness to contact and friction related issues can be traced back to the ancient Egypt, regarding the transportation of large stone blocks to the construction of the pyramids (≈2500 BC). Notorious developments are documented in the personal scripts of Leonard da Vinci (1452-1519) on the phenomenological laws of friction. However, it was only through Guillaume Amontons (1663-1705) and later by Charles-Augustin Coulomb that these laws became widely accepted by the scientific community. Back in that time, only empirical formulation of contact mechanics was known, specially in the context on friction. The first analytical formulation of elastic and frictionless contact was provided in the seminal work by Hertz (1882). Several analytical models regarding distinct situations have been proposed since then, e.g. in reference textbook by K. L. Johnson (1987).

Hertz contact theory and the vast majority of contact models, however, are restricted to very simple situations, such as the elastic frictionless contact under small deformations. In the last decades, developments in computer hardware boosted the utilization of numerical methods to find the solution of several engineering problems. In particular, the finite element method became the gold standard of the numerical methods for industrial applications. Due to its versatility in modeling large deformations and arbitrary material laws, together with frictional contact formulation (where the dual mortar methods can be regarded as the current state of the art methodology), the FEM is capable of answering the ever-growing industry pressure for shorter development periods. This requires a systematic approach based on numerical solutions, rather than trial-and-error processes. The employment of numerical techniques is further substantiated by the excessive cost, or even impossibility, of performing experimental research on some subjects. Typical engineering applications where contact mechanics is paramount, and computational contact mechanics earns its leaving, are sheet metal deep drawing, slip between reinforcing steel and concrete and crashworthiness assessment tests, for the automobile industry.

The aforementioned interface properties required for the application of FEM to frictional contact problems result from micromechanical features of the contact interface, namely, from the rough character of the boundaries. If fact, roughness influences several features of the contact, namely, the real contact area, which is substantially smaller than the apparent contact area. The real contact area has a considerable influence in several physical phenomena, regarding purely normal frictionless elastic (and eventually elasto-plastic) contact. It has been verified from experimental and numerical investigations that the friction force is proportional do the real contact area. Also, electrical and thermal contact resistances are drastically impacted by the real contact area, rather than the apparent counterpart. For the latter, examples of applications where a precise control of thermal contact resistance is required are aircraft joint subjected to aerodynamic heating and structural joints in machine tools (Madhusudana, 2014). Heat dissipation in microelectronics through the application of *Thermal Interface Materials* (TIM), which maximize the heat transfer from the components to the heat sinker, is also a field of growing interest. The sealing of valves in nuclear power plants is assured by steel to steel contact, owing to the high pressure of the involved fluids (Yastrebov, Durand, *et al.*, 2011). In order to avoid micro-leakage, the free volume between two contacting rough surface must verify some conditions, which can be established heuristically, but also predicted numerically. Figure 1.1 illustrates some engineering applications where rough contact is paramount.



(a) Example of a TIM

(b) Theoretical foundations of the application of TIM

(c) Tire performance in wet and dry conditions

(d) Fluid flow in the free volume of two contacting rough surfaces

Figure 1.1: Influence of rough contact in engineering problems. Thermal Interface Materials (the blue paste in **(a)**, adapted from HENKEL (2017)) are applied to the interface of microelectronics devices at heat dissipation components, in order to promote maximum heat transfer, by decreasing the inherently high contact thermal resistance, as illustrated in **(b)** (adapted from Suh *et al.* (2015)). Roughness has a crucial influence on tire performance in several conditions, to the hysteretic effect of rubber materials, at different scales— see figure **(c)**, adapted from Wagner (2018). Roughness also plays a central role in sealing, since the contact area resulting from the contact of rough surfaces is usually composed by several isolated *islands*, i.e., disconnected regions, therefore fluids can flow through the free volume of the contact interface. This is illustrated in **(d)**, where the dark regions represent the contact spots and the reddish colors suggest higher flow rates (adapted from Dapp, Lücke, *et al.* (2012)).

There is only a small number of cases where frictionless rough contact can be assumed. The frictional rough contact, which occurs more often, impacts, for example, tire-road interactions. This topic has been continuously studied by tires manufacturers, in order to optimize tire profiles and, consequently, safety and rolling noise.

Roughness has been observed to behave like a fractal, meaning that its features extend across several length scales—the so called self-affine rough surfaces. The multiscale nature of roughness turns impracticable the direct application of finite element techniques, in order to model its mechanical response, due to the excessively large meshes that it would require. This motivates the formulation of numerical multiscale strategies, with the purpose of reducing the computational resources for modeling the frictionless and frictional rough contact. Commonly, these strategies are designated by *contact homogenization*, which establishes the analogy with typical homogenization techniques applied the bulk of heterogeneous materials (Stupkiewicz, 2007). In a contact homogenization approach, the rough surface is replaced with a smooth version having equivalent interface properties (real contact area and coefficient of friction). A review on multiscale modeling of rough contact was recently published in Vakis *et al.* (2018).

Other class of contact problem which are amenable for the treatment with contact homogenization schemes are the frictional contact with third bodies. In these cases, particles are assumed to exist between the two contacting media. This class of problems is also extremely important in engineering, e.g., in powder lubrication (Iordanoff *et al.*, 2002), wheel-rail contact with grains of sand (Berthier *et al.*, 2004) and also in biomedical application, such as in wear induced by particles in the knee joint, typically associated with osteoarthritis (Berthier *et al.*, 2004).

1.2 State of the art

A brief literature review on the main topics of this dissertation is presented next. In the respective chapter, the following state of the art is extended. Here, only the fundamental aspects and main references are discussed.

1.2.1 Roughness characterization

It is current practice in engineering to characterize rough surfaces by a given set of roughness parameters, such as the RMS height and slope, which condense in a single number all roughness features (Thomas, 1999; ISO 4287, 1997). Furthermore, it has been repeatedly verified experimentally that the RMS parameters are scale dependent, hence are not suited to describe the multiscale character of roughness (Sayles and Thomas, 1978). More complete descriptions of roughness have been attempted following the concept of *Autocorrelation Function* (ACF). Exponentially decaying ACF roughness models were initially proposed, but it has been realized that the correlation length (parameter of the exponential ACF) was scale dependent, as well (Thomas, 1999). The technique that allows a truly scale independent roughness characterization is the Fourier transform of the autocorrelation length, called the *Power Spectral Density* (PSD). It contains the spectral information of the rough topography, such that the surface can be synthesized from this function as the superposition of several harmonics (sinusoidal waves) with random phases—the amplitudes are extracted from the PSD itself. The power spectrum of real surfaces verify a power law (Sayles and Thomas, 1978), typical of fractal surfaces (Russ, 1994), and is experimentally verified to be scale independent (Persson, 2014). The PSD of fractal rough surface, usually termed self-affine rough surface, is expressed in terms of the Hurst roughness exponent H. In turn, this is a measure of the fractal dimension of the topography. The fractal characterization of roughness can be extended to model anisotropy in rough surfaces. In addition to the spectral characterization of rough topography, also the heights distribution is paramount, specifically for the distinction between Gaussian and non-Gaussian heights distributions. The characterization of non-Gaussian topographies is usually reduced to the identification of two parameters, namely the skewness and kurtosis.

1.2.2 Numerical generation of rough surfaces

The application of numerical methods to rough contact requires the discretization of the rough topography. One alternative for performing such discretization is by experimental measurements. While this may seem a safe and practical alternative for rough *profiles*, the measurement of 3D rough topography is cumbersome and costly in time. A methodology perfectly fit to be embedded in a numerical framework is the numerical generation of randomly rough topography. The main goal of these numerical techniques is the synthesis of rough profiles and surfaces verifying prescribed spectral properties and also features of the heights distribution—skewness and kurtosis. Initial algorithms were based on *time-series* concepts, such as *Autoregressive Moving Average models* (ARMA). The cornerstone of these methods were the establishment of coefficients of a recursive expression, which represents the transformation of a white noise signal (Staufert, G., 1979; DeVries, W. R., 1979). An important category of generation algorithms was started with the work of Patir (1978), based on the linear transformation of random matrices. Several extensions of this method have been proposed, namely in the work of Bakolas (2003), by employing the *Nonlinear Conjugate Gradient Method* NCGM to improve the efficiency in the solution of the system of nonlinear equations. Algorithms based on *Fast Fourier Transforms* (FFT) became very popular, since the seminal work of Hu and Tonder (1992). Several variations and improvements have been proposed, from which it is important to cite the contributions by J.-J. Wu (2000b, 2004). The generation of non-Gaussian topography, which is already incorporated in several of the previous references, is based on Johnson system of frequency curves (N. L. Johnson, 1949; Elderton and N. L. Johnson, 1969), and on the fitting algorithm by I. D. Hill, R. Hill, *et al.* (1976).

1.2.3 Single scale modeling of frictionless elastic rough contact

Two major approaches have been adopted for modeling frictionless elastic rough contact, namely, analytical and numerical models. The first widely accepted analytical model for rough contact was proposed by Greenwood J. A. and Williamson J. B. P. (1966), based on the hypothesis that the asperity heights (surface summits or profile peaks) have spherical caps with the same radius every height. Several multiasperity models have succeeded the original theory by Greenwood and Williamson, by incorporating features of the exact theory for isotropic Gaussian random rough surfaces, developed by Nayak (1971). Among these multiasperity theories, the BGT model proposed by Bush, Gibson, and Thomas (1975) and a simplified version by J. A. Greenwood (2006) have raised notoriety in the scientific community. A completely different contact theory was developed by Persson (2001a,b). This model is based on the probability distribution of the contact stresses, and does not concern topography characteristics directly. Analytical theories are mostly formulated in the realm of small deformations, and are limited to elastic contact (the only kind of material behavior concerned in this dissertation). Numerical methods allow the incorporation of complex effects, neglected in the aforementioned theories, such as contact spots coalescence. The *Boundary Element Method* (BEM) is a very popular strategy for numerical modeling of rough contact (Campañá, Müser, and Robbins, 2008; Yastrebov, Anciaux, *et al.*, 2015). Despite the computational advantages of this method, it is also restricted to simple cases, such as linear elasticity and small strains. The *Finite Element Method* (FEM) is the the most versatile tool for solving rough contact problems, allowing the modeling of large strains and nonlinear material laws. It has been successfully applied to frictionless rough contact with elastic and elasto-plastic material laws by Hyun, Pei, *et al.* (2004), Pei *et al.* (2005), and Hyun and Robbins (2007).

1.2.4 Multiscale modeling of rough contact

The major downside of the application of the FEM to rough contact is due to the multiscale character of rough surfaces. Since several length scales are involved in the model, prohibitively fine finite element meshes are required and, consequently, the computation time turns out to be excessive very rapidly. In order to circumvent this issue, several multiscale strategies have been employed, lately. *Computational homogenization*, usually applied to the bulk of heterogeneous materials, can also be extended to contact interfaces (Stupkiewicz, 2007; Temizer and Wriggers, 2008). It consists, primarily, on the replacement of a rough and complex boundary, with a smooth one, having equivalent contact properties. Significant effort has been put on multiscale strategies for rough contact within the context of rubber friction, within finite element frameworks (Reinelt, 2009; Nitsche, 2011; Wagner, 2018). In the recent works by Wagner, Wriggers, Klapproth, *et al.* (2015) and Wagner, Wriggers, Veltmaat, *et al.* (2017), and in opposition to several preceding strategies, the rough surface is split into scales based on its PSD, such that each scale covers a portion of the frequency range. These multiscale approaches, despite being formulated for frictional contact and, typically, viscoelastic materials, can often be readily restricted to frictionless contact and, thus, provide a multiscale solution for the real contact area prediction.

1.3 Objectives

The main objective of this dissertation is to analyze rough contact by means of a multiscale finite element approach based on computational homogenization. The scope is restricted to Signorini-type problems, comprising the elastic, non-adhesive and frictionless contact between a deformable self-affine rough block and a flat and rigid surface. Multiscale strategies are employed for predicting the real contact area evolution with pressure, using less computational resources and having similar precision in comparison with *Direct Numerical Simulation* (DNS) solutions. To accomplish this goal, a numerical framework is built from the ground up, wrapping around the in-house finite element code `LINKS`, equipped with the dual mortar contact discretization. The numerical tool encompasses preprocessing features for random rough topography generation, finite element mesh generation and input data files writing for `LINKS`, which have all been implemented from scratch, together with the post-processing routines required within the implementation of the multiscale strategies. In a first stage, the numerical tool is used to accurately characterize the micromechanical contact problem, in a single scale setup, such that a *Representative Contact Element* (RCE) can be defined. In a second stage, the implemented multiscale schemes are employed to predict the real contact area for different micromechanical contact problems, and the results are properly validated. By using the multiscale techniques, rough contact problems with features typically beyond the scope of application of DNS solutions are analyzed.

1.4 Outline

Following the presentation of the motivation for this dissertation, and a brief overview of the state of the art references, this introduction ends with a succinct outline of the remaining document.

Chapter 2 - On rough surface characterization

The fundamental concepts of multiscale roughness characterization are thoroughly explained in this chapter. The sequence on which the different subjects are presented intends to emphasize, in a sequential manner, the importance of each quantity for the complete and scale independent characterization of roughness. Focus is given to the spectral characterization of roughness via the Power Spectral Density and the respective spectral moments. With grounds on these ideas, the definition of self-affine rough surface is introduced and the mathematical treatment of fractal roughness is addressed. The description of non-Gaussian roughness is presented, as well. This chapters ends with a reformulation of all concepts for discrete topography, thus bridging the gap between pure analytical characterization and the future application of numerical methods.

Chapter 3 - Numerical generation of randomly rough topography

The algorithms for generating rough surfaces and profiles from prescribed spectral properties are introduced in this chapter. Both the Gaussian and non-Gaussian topography generators implemented in this work are discussed and the respective flowcharts are illustrated. Following the introduction of each algorithm, some characteristics are assessed in a series of numerical tests, both for the profile and surface generation versions. This chapter is closed by the application of the numerical generation algorithms to real cases of roller bearings and gear surfaces.

Chapter 4 - Micromechanical elastic contact: analytical models

The most relevant rough contact analytical models for frictionless elastic contact are discussed in this chapter. It covers several multiasperity models, namely, the Greenwood and Williamson, the BGT and the simplified elliptical model, and also the Persson contact theory. These models have been implemented and the results for each one, in particular, the contact area evolution, are plotted. A shallow comparison of these models is performed, and some inherent caveats are referred.

Chapter 5 - Single scale dual mortar finite element modeling of rough contact

In this chapter, a through presentation of the mathematical formulation of contact problems within a dual mortar formulation is provided. It comprises a concise presentation of fundamental aspects of general solid mechanics problems, and of the contact constraints for frictionless and frictional contact. The mortar based finite element formulation is introduced in the weak form, concerning only frictionless contact. This chapter also features a description of the numeral model and the embedding numerical framework. The single scale representativeness tests, performed in order to define a statistically Representative Contact Element (RCE) in 2D, are analyzed.

Chapter 6 - Multiscale finite element modeling of rough contact by contact homogenization

The multiscale approach to rough contact is explained in this chapter. Both the methodology and specific issues, often overlooked in the literature, are discussed. The results and computational requirements are assessed and compared with single scale results. An improved multiscale strategy is proposed, and compared with the initial scheme. The application of the multiscale approach to three dimensional contact, case where the available computational resources would quickly be exhausted, seals this chapter.

Chapter 7 - Concluding remarks and future work

The main conclusions and most relevant contributions of this work are summarized in this last chapter. In addition, suggestions for future continuation of the work developed during this dissertation are mentioned.

Appendix A - Notes on Fourier transforms

This appendix provides an overview of the fundamental aspects of Fourier analysis, namely, on Fourier transforms, required for a smooth understanding of all the dissertation. The numerical generation of rough surfaces and even the spectral characterization of roughness are fundamentally based on such concepts, hence, this appendix was written in order for the dissertation to be self-contained.

Appendix B - Recipe for BGT model computation

The BGT model for rough contact lies on a complex mathematical formulation. The authors, through extensive analytical transformations, proposed simplified expressions, prone to computer implementation. Due to misprints in the original publication, and for the sake of clarity, this appendix resumes the sequence of operations required for implementing this model.

Appendix C - Determination of RMS parameters from spectral properties

The derivation of the analytical relations between RMS parameters in two and three dimensions, with the Autocorrelation Function and Power Spectral Density is presented in this chapter. These results are paramount in roughness analysis, yet their origin is often hidden in the literature.

Chapter 2

On rough surface characterization

Surfaces are boundaries between two media. Intrinsic attributes such as color and hardness can intuitively be attributed to surfaces. Yet, within contact mechanics, the most fundamental property is the geometrical features of these boundaries. In fact, surface geometry is not only fundamental, but poses major obstacles to the analysis of physical phenomena involving roughness. Frequently, engineering surfaces are represented as nominally smooth and, thus, described by an analytical function. The rolling elements from roller bearings, gear tooths with involute profile, are examples of such cases. Nevertheless, real surfaces do not match their nominal shape due to several inevitable abnormalities that take place in their *genesis*. Roughness contributes to the geometrical deviations relative to the real surface shape, however, there are other sources of error which must be identified and isolated.

2.1 Fundamental concepts

As a thought experiment, one can breakdown the production of an hypothetic surface into several sub-processes, and then study the individual influence of each one in the final surface shape. Consider the top face of a cube, which is ideally flat and parallel to the bottom face, as shown in Figure 2.1. This cube is to be produced by milling, from a block of metal. First, consider that, during the machining process, there will be a misalignment of the tool, while theoretically controlling every other variable. This can be visualized by a vibration free operation, in which the material removal process results in continuous chip formation, producing a perfectly clean cut on the metal. As a result, the top face is made flat, but having a slight slope. Nevertheless, this surface is smooth, i.e., there is no roughness, even though it already presents some error relative to the nominal surface. Next, consider the equipment vibration. This will superimpose some kind of waviness over the flat (inclined) top surface. Regardless the wavy appearance of the surface, one can still conceive it as smooth, yet it is no longer parallel to the bottom surface nor flat. The surface is said to be smooth, because the wavelength is sufficiently long, which does not conflict with our perception of roughness—it is possible to imagine sliding a finger through the surface with a virtual sensation of smoothness, as long as wavelength is kept large enough. Indeed, one can conceive that by progressively reducing this wavelength,

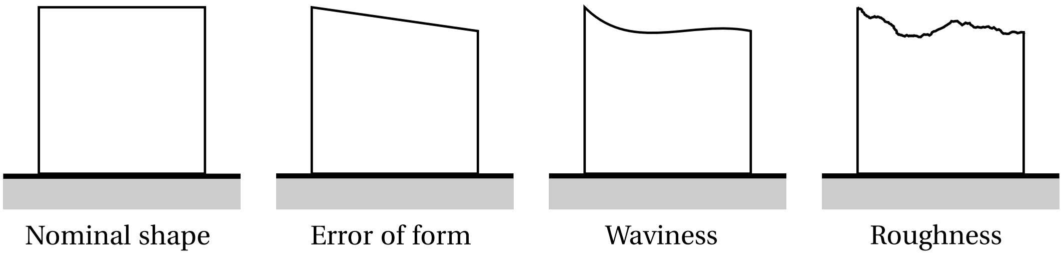


Figure 2.1: Nominal surface and sources of error. The nominal geometry of surfaces differs from the real geometry due to errors of different nature. Errors of form are of very large wavelength, usually represented by a trend surface. For shorter wavelengths, errors are said to belong to the waviness profile. Deviations due to roughness are attributed to the shortest wavelengths. The definition of waviness and roughness is scale dependent.

the surface would cease to be smooth, at a certain point. Finally, the mechanics of material removal are included in such virtual experiment. This generates random errors of high spatial frequency, or short wavelength, that may even go down to the atomic scale, creating a high number of protuberances (cf. Figure 2.1). At this point, one can comfortably call this a rough surface—roughness corresponds to the high frequency variations. The final surface height $h(x,y)$ is reconstructed, roughly speaking, by adding the roughness contribution $z(x,y)$ to the waviness $w(x,y)$ and error of form $e(x,y)$

$$h(x,y) = e(x,y) + w(x,y) + z(x,y)\,. \tag{2.1}$$

Typically, each of these variations, error of form, waviness and roughness, are increasing in frequency and decreasing in amplitude

This thought experiment leads to the vague definition of roughness as the high frequency contribution of deviations. However, this statement implies a scale dependency, as a high frequency variation in a country road can be interpreted as low frequency in a small gearbox shaft. In fact, when using a finger to access whether a surface is rough or smooth, the finger is a sensor itself, which can detect variations in height with a particular resolution. The multiscale nature of roughness can also be assessed from the impact of surface geometry in the performance of physical systems. For example, roughness of amplitude around 1 mm may increase drag on ships due to hull friction, while variations of 1 μm in amplitude in gear tooths will affect friction and wear (Thomas, 1999)—in this case, errors around 1 mm would, mosts likely, be classified as errors of form. Another interesting example is the interaction between car tires and rough road surfaces. In this case, the long wavelengths rule the mechanics of the car's suspension system, while shorter wavelengths are responsible for the frictional behavior between the rubber tires and the road. The smallest wavelengths may fit into the atomic scale, and are related with atom packing and dislocations, which have emerged and form step shapes on the surface (Einax *et al.*, 2013; Misbah *et al.*, 2010). Certainly, a finger could not distinguish between a perfectly smooth surface and other which presented roughness only at atomic scales. From this perspective, the definition of roughness and its separation from the other sources of error, namely, from waviness, seems quite arbitrary. To go into further detail, it is convenient to dig down to surface topography measurement.

2.1.1 Topics on roughness measurement and filtering

The concept of *topography* is responsible for the characterization of surface geometry. For instance, a surface may be described by a height function $h(x, y)$, which holds all information of its topography. If $h(x, y)$ is known, the surface slope at every point $\nabla h(x, y)$ and the mean curvature $\frac{1}{2}\nabla^2 h(x, y)$ would be completely defined.[1] Equation (2.1) emphasizes that surface topography is not the same as roughness, unless there is zero error of form and waviness. Unfortunately, it is not physically possible to perform a continuous description of real surfaces. Instead, surfaces are sampled at a finite number of points, and every point provides some information. In general, each sampled point describes the local height, but has no information neither about the slope, nor curvature, which must be computed *a posteriori*, e.g., by using finite difference techniques. The sampling procedure may be performed along a line, called profile measurement, or on a area, termed areal measurement.[2]

When the topography of any surface is obtained from experimental measurements, a filtering procedure must be employed, in order to remove the unwanted frequency contributions, such that the roughness profile can be isolated. The filtered frequencies are either low frequency deviations, associated with errors of form and waviness, or even high frequencies, that are not relevant for the physical phenomena under study—called *functional filtering* (Thomas, 1999). Regarding the removal of the error of form, it is common practice to fit a trend line (or trend plane, for areal measurements), to the experimental data. Eventually, the fitting function may be a quadratic surface, when one wants to capture some curvature that is known to exist *a priori*—for example, the involute profile of gear tooths. The fitted surface is then subtracted to the data, resulting in a zero mean set, usually— in fact, $w(x, y)$ and $z(x, y)$ in Equation (2.1) are zero mean functions, by definition. The transformed data is filtered, in order to remove the waviness profile. Filtering techniques can be applied either by computer software or by electronic circuits. This operation detaches the waviness from the roughness profile, by defining a separation frequency called *cut-off frequency*—or the respective *cut-off wavelength*. The set up of this parameters is crucial, since roughness measurements are comparable as long as the filtering is performed with the same cut-off between measurements. This issue is addressed by standards, such as ISO 4287 (1997), that suggest practical values for the cut-off as a function of the sample length and expected roughness.

The extraction of the roughness profile from experimental data is always accompanied by unintentional filtering, specially if the measurement instrument is a stylus device (profilometer). Very high frequency variations cannot be measured due to the finite radius of the stylus tip. This also leads to distortion of the measured surface, relative to

[1] The operator $\nabla(\cdot)$ denotes the gradient of a function. When applied to the surface height function, which is a scalar field, it represents a vector whose magnitude equals the maximum surface slope. In Cartesian coordinates, and for a two variable function, it writes $\left(\frac{\partial h}{\partial x}, \frac{\partial h}{\partial y}\right)$. The mean curvature at a point (x, y) is defined as the average between the curvatures at two perpendicular directions, which is related to the second partial derivative on each direction, separately. It can also be written in Cartesian coordinates as $\frac{1}{2}\left(\frac{\partial^2 h}{\partial x^2} + \frac{\partial^2 h}{\partial y^2}\right)$, or by using the Laplacian operator, $\frac{1}{2}\nabla^2 h(x, y)$. Note that this is not the real, physical, curvature at every point, but a quantity which closely relates to it.

[2] Area or surface measurement are also common designations.

the real one—the measurement output is termed the *traced profile*, with the purpose of distinguishing it from the real topography. However, it has been verified that for common engineering surfaces, the error due to this distortion is negligible (Thomas, 1999). The sampling process also limits the measured frequency bandwidth, by setting the upper limit equal to the Nyquist frequency—half the sampling frequency (see Appendix A.3.1, in page 212). Wavelengths larger than the sample length cannot be reliably represented, hence low frequency frequency is also inevitable

2.1.2 Classification of rough surfaces

Surfaces can be classified according to several criteria. Figure 2.2 shows the common typology of engineering surfaces. The first criteria concerns the variation of properties along the surface. If several surface patches have similar properties, the surface is said to be homogeneous, otherwise, it is inhomogeneous. For instance, a surface which was sandblasted in a certain region and lapped in other neighboring patch is inhomogeneous. Based on their stochastic nature, surface can be classified as either deterministic or random. Rough surface are random, but, in contrast, some waviness profiles can be predicted deterministically, e.g., in surfaces produced by turning (Thomas, 1999). Following the characteristics of the height distribution, one can further subdivide surfaces as Gaussian and non-Gaussian. If heights are normally distributed, i.e., follow a Gaussian distribution, the surface is called Gaussian, naturally; if not, they are termed non-Gaussian. The last criterion concerns the directionality of roughness properties. An isotropic surface shows similar properties in every direction. As an example, if profile measurements were carried over lines with arbitrary orientation, they would look statistically similar. In contrast, an anisotropic surface shows different roughness properties in profiles measured along different directions. Usually, the visual identification of anisotropy is immediate, through the presence of scratches along a preferential direction (roughness lay), or even from a non-uniform stretch of the topography properties. The mathematical characterization of anisotropy is not obvious, though.

2.1.3 Roughness in today’s engineering and further readings

Functional performance of surfaces depends greatly on their roughness properties. Reported work on the influence of roughness on several physical phenomena is extremely vast, and interest on this field has been growing, in the last years (Bruzzone *et al.*, 2008; Persson, Albohr, *et al.*, 2005). Topics like surface optimization for contact resistance, sealing, adhesion, friction and manufacturing of extremely small equipment drags attention to rough surface analysis. A good understanding on the mechanisms by which roughness influences physical systems must be preceded by a solid background on rough surface characterization. The following sections present a brief review on the main aspects of random rough surface characterization. This will serve as the basis for the development of numerical tools, aiming at the generation of random rough topography. For further details on roughness measurement the user is referred to the work of Thomas (1999), Mainsah *et al.*, 2013 and Mummery (1992). The textbook by Bhushan (1998) lays an extensive presentation of roughness measurement and characterization at micro and nanoscale, and highlights the importance of these concepts for several engineering applications. Rough-

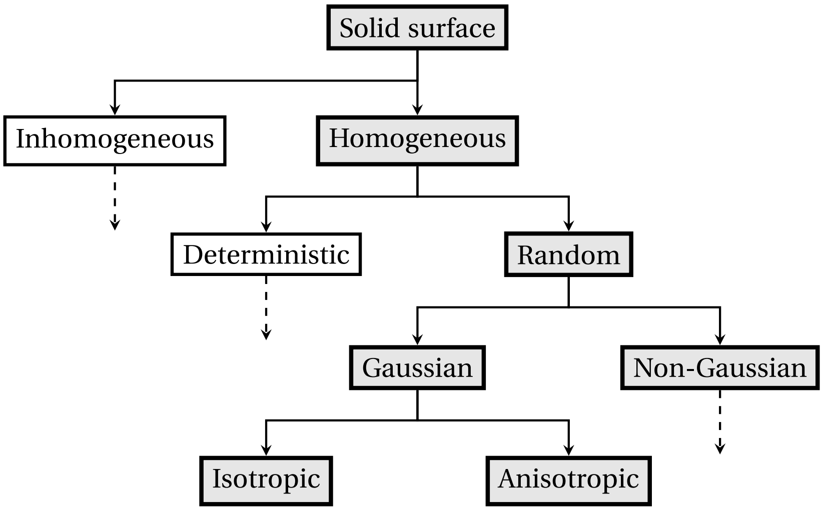


Figure 2.2: Types of surfaces, adapted from Nayak (1971). The dashed lines suggest that subdivision continues following the same rules of the other branches and are not represented here for simplicity. Colored boxes highlight the types of surface of interest for this work.

ness measurement, and respective parameter computation, are widely standardized. The interested reader shall find extremely useful and practical information on the standards ISO 4287 (1997), ISO 25178 (2016) and ASME B46 (2009).

2.2 Roughness parameters

In engineering, rough surfaces are characterized by roughness parameters, traditionally. This practice tries to condense all relevant information about the rough topography in a given set of parameters, which are to be easily computed from measured data. Reducing the rough topography to a numerical value is convenient for comparative studies. They allow for an easy assessment of the evolution of roughness before and after some loading condition. Roughness parameters can be calculated both from profiles and surface measurements. The symbol $R_{(\cdot)}$, along with a specific subscript, denotes profile parameters, and similarly, surface data is referred by $S_{(\cdot)}$. Even though profile measurements are much easier, faster and cheaper to perform, surface parameters provide a more meaningful information about the topography. In fact, surface parameters are computed from all profiles contained in the surface. Moreover, physical phenomena involving rough contact, intrinsically being a three-dimensional problem, depend on surface, rather than on profile geometry. For example, a local maximum of a roughness profile is termed a *peak*, which is defined at points where

$$\frac{\partial z}{\partial x} = 0 \quad \text{or} \quad \frac{\partial z}{\partial y} = 0\,, \tag{2.2}$$

depending on the measurement direction. On the other hand, a local maximum in a rough surface, called a *summit*, must prove

$$\|\nabla z\| = 0\,. \tag{2.3}$$

Since roughness profiles result from intercepting the rough surface with a measurement plane, their peaks are very unlikely to occur at the same position of the summits. The mathematical definitions in Equation (2.2) and Equation (2.3) do also suggest the previous statement: for a tortuous surface, there are several points where only one of the partial derivatives vanish and, thus, peaks will show up at points where there are no summits.

Nowadays, there are a large number of roughness parameters proposed in the literature, and referred in international standards. Whitehouse (1982) even used the parlance *parameter rash*, satirizing the overwhelming number of roughness parameters available. Actually, it happens that some are known not to bring new information, when compared to other existing parameters. The current work does not aim at presenting an extensive review on roughness parameters, but simply at commenting on their utility for roughness characterization. For a more thorough discussion on roughness parameters, their interpretation and application, the reader is encouraged to take look at the work of Thomas (1999), Mainsah *et al.* (2013) and Mummery (1992).

Remark 2.1 on the graphical representation of rough topography.
In the following sections, whenever a rough topography is illustrated, being either a profile or surface, the material is assumed to extend along the negative direction of z.

2.2.1 Root Mean Square parameters

Among the large collection of roughness parameters, some are noteworthy for rough surface analysis, due to their importance on micromechanical contact models, and owing to their relation with other roughness characterization techniques (K. L. Johnson, 1987; McCool, 1986; Nayak, 1971). These are the *root mean square parameters* (abbreviated RMS, for *root mean square*). In particular, one shall be interested in the RMS height (or roughness) slope and curvature.

RMS height

For a continuous roughness profile $z(x)$, or surface $z(x, y)$, one defines RMS height, alternatively designated RMS roughness, as

$$z_{\mathrm{rms},x} = \sqrt{\frac{1}{L}\int_0^L z^2(x)\;\mathrm{d}x} = \sqrt{\overline{z^2(x)}}\,; \tag{2.4a}$$

$$z_{\mathrm{rms},xy} = \sqrt{\frac{1}{L_x L_y}\int_0^{L_x}\int_0^{L_y} z^2(x,y)\;\mathrm{d}y\mathrm{d}x} = \sqrt{\overline{z^2(x,y)}}\,. \tag{2.4b}$$

Here the notation $\overline{(\bullet)}$ is introduced as the spatial average operator. RMS roughness is a measure of the surface height relative to the mean plane. It is computed as the average of squared heights, thus making this parameter sensible to extreme values, being either peaks or valleys. Although it cannot distinguish between these two types of extrema, because it only considers the square real values, which are, by definition, positive. Since one works with sampled surfaces most of the time, Expressions (2.4) are rarely applicable. Sampling the profile $z(x)$ over $x \in [0, L]$ results, typically, in N equally spaced points. If the sampling is to be performed on a surface $z(x, y)$ over the region defined by $(x, y) \in [0, L_x] \times [0, L_y]$, it will then result in a grid of N by M equally spaced points, in the x and y directions, respectively.[3] For discrete data, RMS roughness writes

$$z_{\mathrm{rms},x} \approx R_q = \sqrt{\frac{1}{N}\sum_{n=0}^{N-1} z_n^2}\,; \tag{2.5a}$$

$$z_{\mathrm{rms},xy} \approx S_q = \sqrt{\frac{1}{MN}\sum_{m=0}^{M-1}\sum_{n=0}^{N-1} z_{m,n}^2}\,. \tag{2.5b}$$

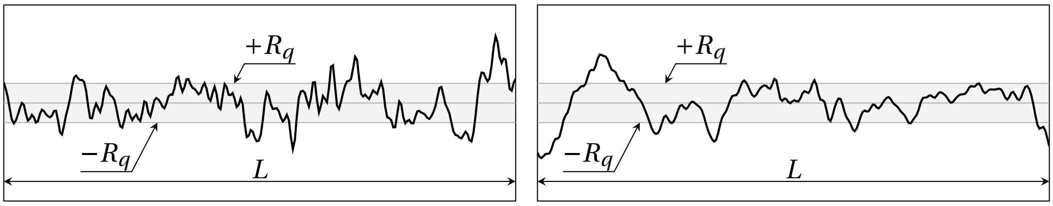
RMS roughness contains information on the global height variation. Higher values of RMS height suggest surfaces with higher contribution of extreme values, thus higher deviations from the mean value (zero). However, this mean variation might occur over any distance, which is not considered in the computation of R_q or S_q. One can readily conceive two rough profiles with the same value of R_q, but having considerably different topography. To illustrate this concept, two different profiles with equal length and R_q are plotted in Figure 2.3. In average, the profile displayed in Figure 2.3a has larger slopes, in magnitude, than the one in Figure 2.3b, even though they share the same value of RMS roughness. Thus, RMS roughness alone is unable to capture all information of rough topography.

Remark 2.2 on the notation for discrete variables indices.
When a continuous function $z(x, y)$ is sampled on a grid of N equally spaced points in the x direction and M in the y direction, the result will be denoted by the discrete variable $z_{m,n}$, with $m = 0, \ldots, M-1$ and $n = 0, \ldots, N-1$. While the traditional argument order (x, y) is kept for the continuous function, when the sampling is introduced, the first index m refers to the y coordinate, while the second index n refers to the x coordinate. This way, the position of matrix element (m, n) resembles the physical position of the respective point on plane xOy, where the x axis is, traditionally, horizontal (number of column), and the y axis is vertical (number of line). Since this convention was used in programming the surface generators (for personal convenience), it will be used through the text, to keep the notation consistent.

RMS slope

The previous discussion suggests the usage of slope related parameters to distinguish between surfaces with the same average amplitude characteristics. Analogously to RMS

[3] Here, the operator $[\bullet,\bullet] \times [\bullet,\bullet]$ denotes the Cartesian product.



(a) Profile with high average slope magnitude **(b)** Profile with low average slope magnitude

Figure 2.3: Comparison between different profiles with same value of R_q. Both profiles have the same sampling length and roughness RMS, yet visual differences are quite evident. The profile in Figure 2.3b is *smoother* than of Figure 2.3a. This is, in average, the magnitude of the profile slope and curvature in Figure 2.3b are smaller, when compared to Figure 2.3b.

roughness, one defines RMS slope of continuous topographies as

$$z'_{\text{rms},x} = \sqrt{\overline{\left(\frac{\mathrm{d}z(x)}{\mathrm{d}x}\right)^2}}\;; \tag{2.6a}$$

$$z'_{\text{rms},xy} = \sqrt{\overline{\|\nabla z(x,y)\|^2}}\,. \tag{2.6b}$$

Application of Expressions (2.6) to discrete surfaces requires numerical computation of derivatives. Several strategies can be adopted to discretize derivate operations, and, typically, finite-differences formulae are used. Namely, a first-order finite difference stencil is recommended by ISO 25178 (2016), while ASME B46 (2009) suggests a sixth-order stencil. Both of these alternatives may lead to inaccuracies due to smoothing and sharp corners (Jacobs *et al.*, 2017). Following a simpler alternative, based on a forward finite-difference scheme (Bhushan, 1998), RMS slope simply comes

$$z'_{\text{rms},x} \approx R_{\Delta q} = \sqrt{\frac{1}{(N-1)}\sum_{n=0}^{N-2}\left(\frac{z_{n+1}-z_n}{\Delta x}\right)^2}\;; \tag{2.7a}$$

$$z'_{\text{rms},xy} \approx S_{\Delta q} = \sqrt{\frac{1}{(M-1)(N-1)}\sum_{m=0}^{M-2}\sum_{n=0}^{N-2}\left(\frac{z_{m,n+1}-z_{m,n}}{\Delta x}\right)^2+\left(\frac{z_{m+1,n}-z_{m,n}}{\Delta y}\right)^2}\,. \tag{2.7b}$$

RMS curvature

Just like two surfaces with different RMS slope can share the same value of RMS roughness, it may happen that two surfaces have the same value of RMS slope, yet different curvature properties. Therefore, a curvature parameter shall also be useful for a consistent characterization of rough surfaces. Curvature relies on the computation of second derivatives, then the discretization shall follow the same mindset behind slope calculation, i.e., based on finite-difference schemes. Starting with the continuous scenario, the

RMS curvature comes

$$z''_{\mathrm{rms},x} = \sqrt{\overline{\left(\frac{\mathrm{d}^2 z(x)}{\mathrm{d}x^2}\right)^2}}\ ; \tag{2.8a}$$

$$z''_{\mathrm{rms},xy} = \frac{1}{2}\sqrt{\overline{\left[\nabla^2 z(x,y)\right]^2}}\ . \tag{2.8b}$$

It is important to emphasize that Expressions (2.8) are not numerically equal to the geometrical curvature κ of a profile or surface. Concerning profile curvature, the classical result follows

$$\kappa = \frac{z''(x)}{\left(1+[z'(x)]^2\right)^{\frac{3}{2}}}\ , \tag{2.9}$$

where $z'(x)$ and $z''(x)$ denote the first and second derivative of $z(x)$ in order to x. From Equation (2.9), one sees that curvature equals the second derivative of z, only when the slope vanishes, i.e., at profile peaks and valleys.[4] Most of the time, the interest is on these extrema, hence it is common, in rough surface analysis, to address $z''_{\mathrm{rms},x}$ as profile RMS curvature. (Bhushan, 1998; Nayak, 1971; Greenwood J. A. and Langstreth J. K., 1984). Furthermore, this quantity can be used in rough contact theories, in order to approximate the mean curvature of summits and peaks (McCool, 1986). These reasons support the designation of curvature for $z''_{\mathrm{rms},x}$ and $z''_{\mathrm{rms},xy}$.

Identically, the actual surface curvature is not equal to the Laplacian of surface height at every point. In addition, depending on the measurement direction, one will find different curvatures at the same point. For a smooth surface, one can always find two orthogonal directions where curvature is maximum and minimum—principal curvatures, associated with the principal directions of curvature. To account for the variation of curvature with direction, the average between the principal curvature is taken. It can be proved that the average between principal curvatures is equal to the average of curvatures between any two orthogonal directions (Sokolnikoff, 1951). Hence, one can work with the mean curvature between x and y direction, case where the mean summit curvature equals half the Laplacian, at the summits. For the discrete case, adopting a centered finite difference scheme (Bhushan, 1998), the curvature is approximated by

$$z''_{\mathrm{rms},x} \approx R_{\Delta^2 q} = \sqrt{\frac{1}{(N-1)}\sum_{n=0}^{N-3}\left(\frac{z_{n+1}-2z_n+z_{n-1}}{\Delta x^2}\right)^2}\ ; \tag{2.10a}$$

$$z''_{\mathrm{rms},xy} \approx S_{\Delta^2 q} = \left[\frac{1}{(M-2)(N-2)}\sum_{m=0}^{M-3}\sum_{n=0}^{N-3}\frac{1}{4}\left(\frac{z_{m,n+1}-2z_{m,n}+z_{m,n-1}}{\Delta x^2}+\right.\right. \tag{2.10b}$$

$$\left.\left. + \ \frac{z_{m+1,n}-2z_{m,n}+z_{m-1,n}}{\Delta y^2}\right)^2\right]^{\frac{1}{2}}\ .$$

[4] Also at inflection points, where the second derivate also vanishes. These are less relevant cases.

2.2.2 Scale dependency and representativeness

This analysis could be continued by including new RMS parameters, since it is certain that two surfaces sharing the same value of RMS curvature can have different RMS parameter involving third-order derivatives of the height function. Nevertheless, new complications arise, related to derivative computation, such as the decreasing number of points available for the application finite-differences, and also to the loss of physical meaning of such parameters.

A key point to mention is that these three parameters, and a great number of roughness parameters currently in use, are average quantities, which express the global behavior of the surface, but contain no information of the local geometry, e.g., about the shape of the summits. Moreover, to state that these parameters are characteristics of the surface, one needs to have experimental evidence on the invariance of such quantities under variations of sampling length, discretization and measurement equipment. In fact, it is verified that RMS roughness is very sensible to low cut-off wavelength, typically decreasing for shorter low cut-off wavelengths (Bhushan, 1998; Sayles and Thomas, 1978). On the other hand, it has been observed that RMS slope and curvature increase for decreasingly smaller high cut-off wavelengths (P. I. Oden *et al.*, 1992). These results suggest that RMS parameters are scale dependent. RMS roughness is sensible to large scales, while curvature and slope are more sensible to smaller scales. RMS slope and curvature do not even seem to converge, and tend to increase indefinitely with decreasing sampling interval.

As a final remark, it is noteworthy to mention that this discussion does not intend to drive out attractiveness from roughness parameters. Repeating a previous statement in this text, roughness parameters are paramount in engineering due to their convenient measurement and computation. This provides an efficient workflow for assessing roughness impact on certain phenomena, regardless of their scale and instrument dependence. In fact, all these parameters are verified to affect surface functional performance, yet, usually only one of these parameters dominate in a particular instance. For example, RMS height dominates contact stiffness (Campañá, Persson, *et al.*, 2011), while RMS slope and curvature are mostly responsible for adhesion properties (Pastewka and Robbins, 2014). Nonetheless, it is fair to state they are not properties of a rough surface, therefore ill-suited for the characterization of roughness across scales.

2.3 Random process

Attending to the random nature of rough surfaces (cf. Figure 2.2), it is logical to proceed their analysis in view of *random* or *stochastic processes*. The surface height at every point can be interpreted as a rough topography is a *random variable*, denoted $\mathscr{Z}$, which can have several realizations $z^{(k)}(x,y)$—or $z^{(k)}(x)$, for profiles. Stochastic analysis is a vast field of mathematics, from which just very basic concepts are referred in the following section. A rigorous mathematical description of random processes and random variables is unnecessary for the level of understanding required for its application to rough surface analysis. Hence, precise mathematical definitions will be avoided, and focus will be given to physical interpretation of the concepts. Meirovitch (2001) gives a tangible presentation on random vibrations, that can be directly applied to random surface description.

2.3.1 Autocorrelation function

A random variable is characterized by all the possible surfaces $z^{(k)}(x, y)$ that can be obtained from a measurement, termed the *ensemble*. There is a probability measure associated with each realization, i.e., the probability of the output surface being similar to a given reference. One shall attempt to describe the random surface by the statistics computed across the collection of realizations, termed *ensemble averages*. Consider that there are K realizations. The ensemble mean value of $\mathscr{Z}$ at some point i of coordinates (x_i, y_i) is given by

$$\mu_z(x_i) = \frac{1}{K}\sum_{k=0}^{K-1} z^{(k)}(x_i) = \langle z^{(k)}(x_i)\rangle \; ; \tag{2.11a}$$

$$\mu_z(x_i, y_i) = \langle z^{(k)}(x_i, y_i)\rangle \; . \tag{2.11b}$$

The average in Equations (2.11) is taken not over the the sample function, but over all realizations, at the specific point (x_i, y_i). It does not represent the mean height of the surface, but the mean height of the ensemble at that particular point. Observe that $\langle \bullet \rangle$ denotes the ensemble average operator, and that μ_z is a function of (x_i, y_i). Another important ensemble average, called the *autocorrelation function* (ACF), follows

$$R(x_i, x_i + \tau) = \langle z^{(k)}(x_i) z^{(k)}(x_i + \tau)\rangle \; ; \tag{2.12a}$$

$$R(x_i, y_i, x_i + \tau_x, y_i + \tau_y) = \langle z^{(k)}(x_i, y_i) z^{(k)}(x_i + \tau_x, y_i + \tau_y)\rangle \; . \tag{2.12b}$$

Note that R is a function of the position (x_i, y_i) and also of the shift (τ_x, τ_y). Figure 2.4 illustrates the points involved in the ACF calculation. This function is defined as the ensemble average of the product between the surface height at two points, whose relative position in given by the vector (τ_x, τ_y). Thus, such function reports to the joint probability density function of surface height at point (x_i, y_i) and $(x_i + \tau_x, y_i + \tau_y)$. This is, it can be intuitively associated with the probability of simultaneously existing a point with height $z_i^{(k)}$, occurring at point (x_i, y_i), and another with height $z_{i+\tau}^{(k)}$, at point $(x_i + \tau_x, y_i + \tau_y)$, in the same realization. Ultimately, if the ensemble averages involving the product of more points is specified, it would be possible to evaluate the probability of that particular realization to occur—or more precisely, of realizations infinitesimally similar it.

Regarding the characterization of some topography, a description based on ensemble averages is quite unsatisfactory, since it is not able to aggregate the characteristics of each realization. In addition, all previous ensemble averages are a function of the position, which is also inconvenient, due to the stochastic behavior of surface height. Nonetheless, rough surfaces are commonly assumed to be *stationary*, which implies that all ensemble averages are independent of the position (x_i, y_i). Usually, it suffices to consider that rough surfaces are *weak-sense stationary*, i.e., that the ensemble mean and ACF are position independent. For a stationary rough surface, increasing the sampling domain beyond a certain threshold does not introduce new information. Moreover, rough surfaces are also considered *ergodic* processes, which implies that a sample is representative of all

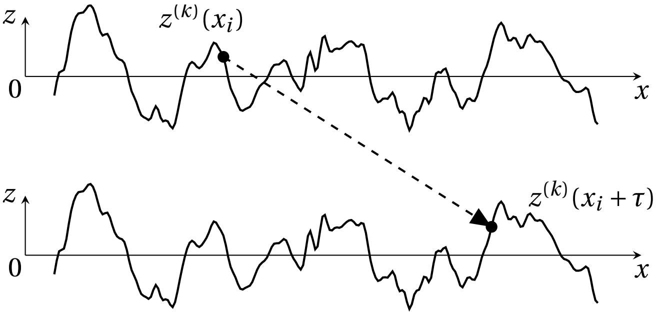


Figure 2.4: Points involved in the computation of the ACF. By taking the ensemble average of the height product at these two points, one gets a result related with the probability of both heights occurring separated by τ in a same profile. For a rough surface, the idea is similar, yet the displacement between both points is a two dimensional vector.

realizations—a single profile contains characteristics which are identical to every other profile measured on the surface.

By using the stationarity property, one can remove the position dependency from the ensemble averages—μ_z turns into a constant and ACF is left as a function of the shift (τ_x, τ_y). Besides this, from the ergodicity property, ensemble averages can be replaced by sample averages.[5] In this scenario, one can drop the superscript (k) referring to a particular realization, and rewrite Expressions (2.11) as

$$\mu_z = \overline{z(x)}\,, \tag{2.13a}$$

$$\mu_z = \overline{z(x,y)}\,. \tag{2.13b}$$

As for the redefinition of ACF in Expressions (2.12), it is important to mention that the number of points (x, y) having a corresponding point $(x+\tau_x, y+\tau_y)$ decreases with increasing τ_x and τ_y, unless the surface is infinite, or periodic. Thus, if surface length is finite and non-periodic in each direction, the matching region reduces as it slides over its clone. Based on the previous argument, the ACF for an ergodic finite-length topography becomes

$$R(\tau) = \frac{1}{L-\tau}\int_0^{L-\tau} z(x)z(x+\tau)\,\mathrm{d}x\,, \tag{2.14a}$$

$$R(\tau_x,\tau_y) = \frac{1}{(L_x-\tau_x)(L_y-\tau_y)}\int_0^{L_x-\tau_x}\int_0^{L_y-\tau_y} z(x,y)z(x+\tau_x,y+\tau_y)\,\mathrm{d}y\mathrm{d}x\,. \tag{2.14b}$$

Note that, for a profile, $R(0) = (z_{\mathrm{rms},x})^2$ and for a surface $R(0,0) = (z_{\mathrm{rms},xy})^2$, cf. Expressions (2.4). In statistical terms, and for zero mean functions (μ_z=0), which happens to be the case of rough topography, this quantity is also equal to the sample variance of heights σ_z^2, which is defined as the sample average of the squared heights. The standard

[5] Ergodic surfaces are necessarily stationary, yet stationary surfaces may not be ergodic. Both properties are referred in the text to show their separate effect.

deviation of the sample heights is defined as the square root of variance

$$\sigma_z = \sqrt{\overline{(z(x))^2}} = z_{\mathrm{rms},x} = \sqrt{R(0)}\,, \tag{2.15a}$$

$$\sigma_z = \sqrt{\overline{\big(z(x,y)\big)^2}} = z_{\mathrm{rms},xy} = \sqrt{R(0,0)}\,. \tag{2.15b}$$

In contrast with all roughness parameters previously referred, the ACF is a function, rather than a value, so it intrinsically holds a larger set of information about surface topography than a simple parameter. This quantity measures how similar the surface looks with a shifted copy of itself, i.e., its level of periodicity. Inspecting Expressions (2.14), it can be seen that when no shift is applied ($\tau = 0$), the integral is reduced to the squares of local heights, which are all positive quantities. If a non-zero shift is considered ($\tau \neq 0$), it is expected that positive values of height will be multiplied by negative ones, thus leading to a negative contribution from these points, and, ultimately, causing a decay on the ACF. It is expected that the tendency will be of decreasing ACF with increasing shift magnitude. Actually, it can be proven that the global maximum value of the autocorrelation function does occur at the origin. The rate at which ACF decreases depends on the shape of the profile: if a profile is nearly smooth with gentle slopes, the decay will be slower than a profile with high frequency oscillations. From this argument, it is seen that ACF is related to the local shape of the profile. Furthermore, if a profile is periodic, the its ACF will also be periodic, with the same period.

2.3.2 Exponentially decaying ACF roughness model

Experimental evidence suggests that several surfaces have profiles verifying an exponentially decaying ACF (Whitehouse and Archard, 1970; Greenwood J. A. and Langstreth J. K., 1984; Panda *et al.*, 2016). This roughness model is usually stated by

$$R(\tau) = R_q^2 \, \exp\big(-|\tau|\beta\big)\ . \tag{2.16}$$

Here, β is termed the autocorrelation length (ACL), and is equal to the shift that results in a decrease of the initial ACF value by a factor of e^{-1}, which is about 37% of R_q^2.[6] In general, this distance is less that 10% of the sample length (Panda *et al.*, 2016). The influence of the ACL on the profile topography can be seen in Figure 2.5. This figure displays two profiles with same RMS roughness and same length, yet with different ACL. The profile with longer ACL shows gentler slopes and wider peaks.

Remark 2.3 on the definition of autocorrelation function and length.
The autocorrelation function was defined as the spatial average of the product of surface heights with a shifted copy of itself, after ergodicity and stationarity considerations. In other bibliographic sources, namely in Thomas (1999) and Mainsah et al. *(2013), this function is called the autocovariance function, while the autocorrelation functions is considered as the autocovariance normalized by the square of RMS roughness. In this*

[6] In literature, autocorrelation function and length are commonly referred under the designation of *Texture Parameters*,

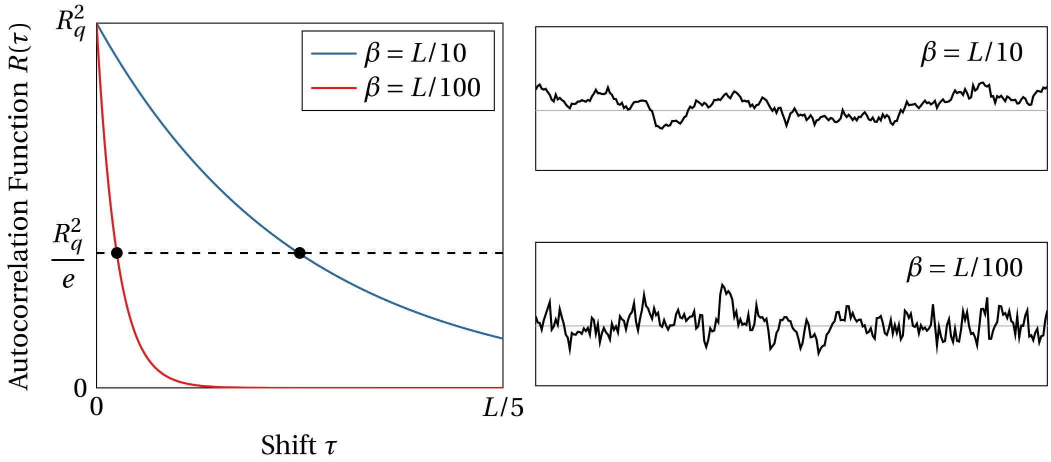


Figure 2.5: Effect of autocorrelation length on profile topography. Both profiles have the same length L and roughness RMS. The profile having longer ACL is smoother than the one with shorter ACL. These quantities can described local geometrical features of rough topography.

work, the definition given by Equation (2.14) *is adopted. Regarding the definition of autocorrelation length, in this work it is considered the distance from the origin where the ACF decays to* $1/e$ *of its initial value, with* $e = \exp(1)$. *Another common definition in literature uses the distance on which ACF decays to 10% of the value at the origin. If this definition is adopted, all previous expressions related with exponential autocorrelation need to be changed, accordingly.*

The exponentially decaying ACF, as expressed in (2.16), does not allow the existence of curvature in the profile. This was pointed out by Nayak (1971), who justified the consistency of results from Whitehouse and Archard (1970) based on the applied finite sampling length. Nayak suggested that real profiles may exhibit exponential ACF only at a certain distance from the origin, so that it can be fourth-order smooth at the origin. This result was later confirmed by Whitehouse (1978) and Greenwood J. A. and Langstreth J. K. (1984). Based on exponential ACF model, Whitehouse and Archard (1970) developed a theory characterizing the geometry and the statistics of profile peaks. Namely, theoretical predictions were derived for the areal density of peaks, mean peak height, variance of peak height, mean peak curvature, variance of peak curvature and mean slope. Focusing the theory on profile peaks is essential, since contact is primarily established on peaks (actually, on summits, but peaks are a good first approximation). These results predicted that the mean slope and curvature increase with decreasing sampling length, which was already mentioned earlier in Section 2.2—yet, from different arguments.

The concept of exponentially decaying ACF can easily be extended for two dimensions. The existence of two directions opens the possibility of existing directional properties, i.e., it allows anisotropy to exist—anisotropic rough profiles cannot exist, since only one direction is concerned. Anisotropic behavior can be incorporated within the exponential

ACF model by setting different autocorrelation lengths for each direction. In general, the exponential ACF for a surface reads

$$R(\tau_x, \tau_y) = S_q^2 \, \exp\left(-\sqrt{(\tau_x/\beta_x)^2 + \left(\tau_y/\beta_y\right)^2}\right) . \tag{2.17}$$

If the autocorrelation length is equal in both directions, the surface is isotropic and, hence, the ACF becomes circularly symmetric relative to the origin—it is only dependent on the distance relative to origin. When different autocorrelation lengths are specified, curves of constant ACF are elliptical whose aspect ratio is β_x/β_y. Surface summits also have approximate elliptical shape with same aspect ratio. The effect of ACL in both directions is represented in Figure 2.6. In particular, two isotropic surfaces with different ACL along with an anisotropic surface are shown. Comparing both isotropic surfaces one reaches the foregoing conclusions regarding profile analysis, where it was observed that increasing ACL leads to a overall smoother surface. As for the anisotropic surface, roughness marks can be observed along the y direction, i.e. in the direction of larger correlation length.

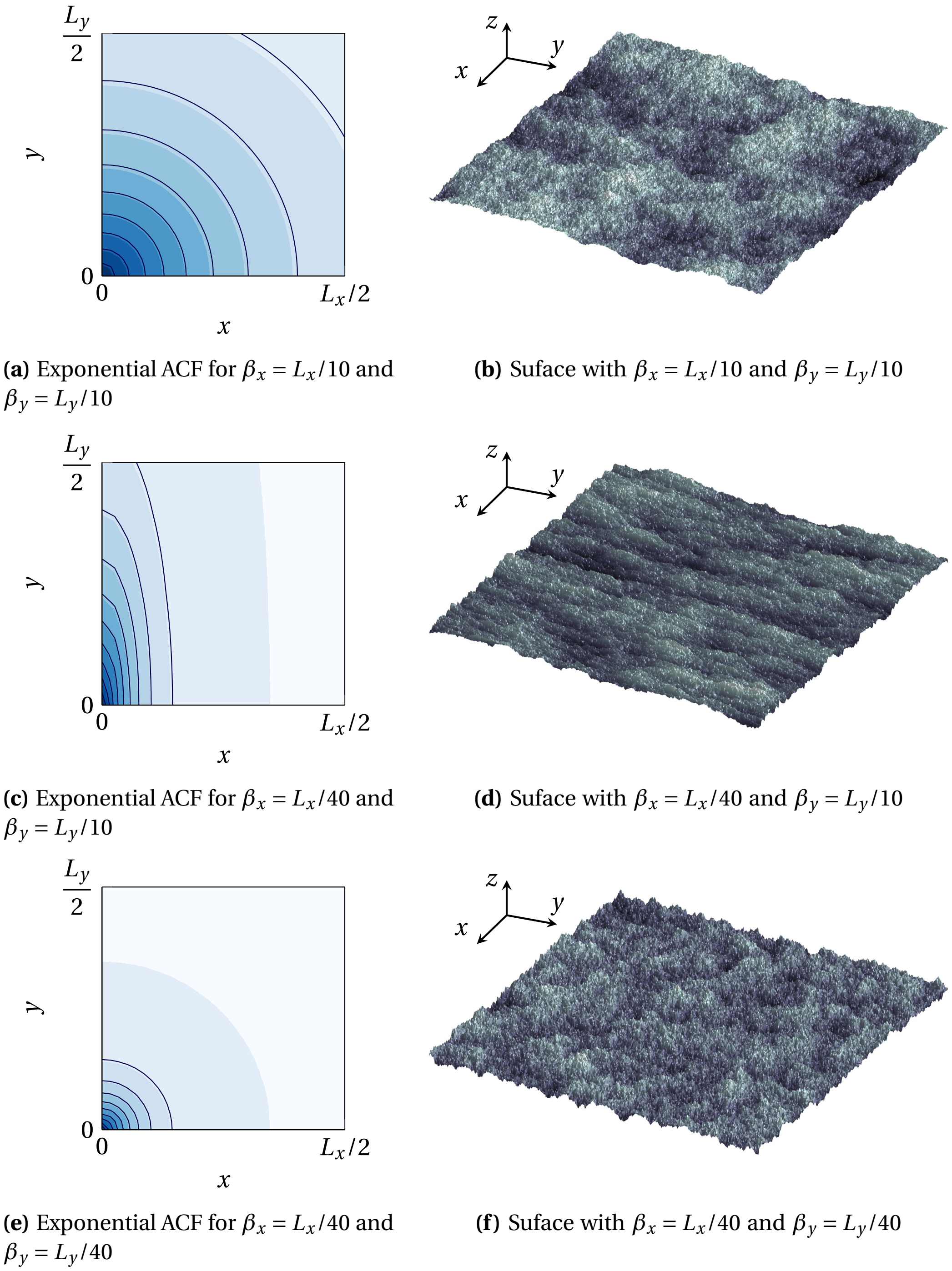
To summarize, it has been shown that roughness parameters are ill-suited to characterize rough topography, since they depended on sampling length, sampling frequency and instrument specifications. Furthermore, as they are average measures, they cannot describe local topography. The autocorrelation function and, in particular, the classical exponentially decaying ACF, solves this problem, since it holds information about the average amplitude and also on the local shape of peaks and summits. However, some limitations arise when using this technique. Zhang *et al.* (2014) and Panda *et al.* (2016) observed in real measurements that the autocorrelation length changes with sample size and with sampling frequency. Therefore, it suffers from the same drawbacks as roughness parameters: it is scale dependent. Thomas (1999) reached similar conclusions by observing that the ACF of rough profile is dependent on the cut-off length. Moreover, ACF alone is not able to uniquely characterize surface topography. As an illustration, Figure 2.7 shows profiles with similar ACF, yet considerable different topography. While profile in Figure 2.7a has peaks and valleys with similar height, profile in Figure 2.7b has very low peaks and deep valleys. This examples suggests that key to distinguish between these profiles lies on the height probability distribution.

Thus, there are two different problems that need to be tackled. First, a roughness model which is not scale dependent is sought, in order to incorporate an appropriate framework for studying roughness across scales. Second, the height probability distribution should be considered in order to model certain topographic features, cf. Figure 2.7. Before addressing these topics, however, the most notorious tool for roughness characterization, called the Power Spectral Density (PSD), is introduced and several aspects, such as the relations with previous techniques, are explored.

2.3.3 Power spectral density

An alternative tool for rough surface analysis, and arguably the most important one, is the Power Spectral Density (PSD).[7] Before stating the definition of PSD, one shall assume in

[7] The designation *power spectral density* comes from the signal processing theory, which usually concerns a value of electric voltage instead of surface height, and time instead of spatial coordinates. For a clearer



(a) Exponential ACF for $\beta_x = L_x/10$ and $\beta_y = L_y/10$

(b) Suface with $\beta_x = L_x/10$ and $\beta_y = L_y/10$

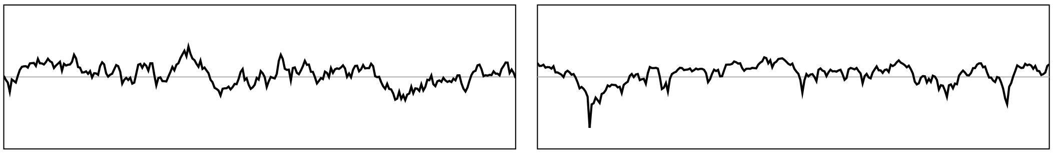
(c) Exponential ACF for $\beta_x = L_x/40$ and $\beta_y = L_y/10$

(d) Suface with $\beta_x = L_x/40$ and $\beta_y = L_y/10$

(e) Exponential ACF for $\beta_x = L_x/40$ and $\beta_y = L_y/40$

(f) Suface with $\beta_x = L_x/40$ and $\beta_y = L_y/40$

Figure 2.6: Effect of autocorrelation length in x and y directions in surface topography. Blue colors suggest higher values of ACF while white represents lower values (the maximum value is S_q^2 at the origin). In the same breadth with the observations on rough profile, it is observed that when the ACL decreases, the topography is less smooth. Anisotropy on surface topography results when different ACL are specified for each direction.



(a) Gaussian profile **(b)** Non-Gaussian profile

Figure 2.7: Profiles with different topography, yet having similar ACF. Profile plotted in Figure 2.7b shows deeper valleys and low peaks. Its topography can be characterized essentially by an high concentration of points neat the top . On the other hand, profile represented in Figure 2.7a shows valleys and peaks with approximately the same height, without any noticeable trend.

this section that $z(x,y)$ or $z(x)$ have infinite length, for simplification. In these conditions, one writes ACF as

$$R(\tau) = \lim_{L\to\infty} \frac{1}{L} \int_{-\frac{L}{2}}^{\frac{L}{2}} z(x)z(x+\tau)\ \mathrm{d}x\,, \tag{2.18a}$$

$$R(\tau_x,\tau_y) = \lim_{\substack{L_x\to\infty\\ L_y\to\infty}} \frac{1}{L_xL_y} \int_{-\frac{L_x}{2}}^{-\frac{L_x}{2}} \int_{-\frac{L_y}{2}}^{-\frac{L_y}{2}} z(x,y)z(x+\tau_x,y+\tau_y)\ \mathrm{d}y\mathrm{d}x\,. \tag{2.18b}$$

The PSD, denoted here by Φ for surfaces and Φ_θ for profiles, is formally defined by the Wiener–Khinchin theorem, which assumes stationarity for the topography height, as the Fourier transform of the ACF[8]

$$\Phi_\theta(k) = \mathcal{F}\{R(\tau)\} = \int_{-\infty}^{+\infty} R(\tau)e^{-\mathrm{i}k\tau}\ \mathrm{d}\tau\,; \tag{2.19a}$$

$$\Phi(k_x,k_y) = \mathcal{F}\left\{R(\tau_x,\tau_y)\right\} = \iint_{-\infty}^{+\infty} R(\tau_x,\tau_y)e^{-\mathrm{i}(k_x\tau_x+k_y\tau_y)}\ \mathrm{d}\tau_x\mathrm{d}\tau_y\,. \tag{2.19b}$$

The $\mathcal{F}\{\bullet\}$ stands for continuous Fourier transform, and $\mathrm{i}=\sqrt{-1}$ is the imaginary number. The ACF can be recovered from PSD by applying the inverse Fourier transform, here denoted by $\mathcal{F}^{-1}\{\bullet\}$, as

$$R(\tau) = \mathcal{F}^{-1}\{\Phi_\theta(k)\} = \frac{1}{2\pi}\int_{-\infty}^{+\infty} \Phi_\theta(k)e^{\mathrm{i}k\tau}\mathrm{d}k\,; \tag{2.20a}$$

$$R(\tau_x,\tau_y) = \mathcal{F}^{-1}\left\{\Phi(k_x,k_y)\right\} = \frac{1}{4\pi^2}\iint_{-\infty}^{+\infty} \Phi(k_x,k_y)e^{\mathrm{i}(k_x\tau_x+k_y\tau_y)}\ \mathrm{d}k_x\mathrm{d}k_y\,. \tag{2.20b}$$

In order to understand the notation used in Expressions (2.19) and (2.20), some aspects of Fourier analysis shall be mentioned in the following paragraphs. The reader is referred to Appendix A for a more detailed review of this topics. Only essential interpretation and results will be given in the present section.

view on this designation, the user is referred to Proakis and Salehi (2002).

[8] Fore a detailed proof of the theorem, the reader is referred to the book from Proakis and Salehi (2002).

In one dimension, Fourier transforms describe the decomposition of aperiodic functions as the sum of sinusoidal waves, spanning all frequencies/wavenumbers k. Each wave is characterized by its amplitude and phase, which are related to the magnitude and argument of the function's Fourier transform at that specific frequency. Notice that the Fourier transform of a real function is, in general, complex valued. Equation (2.20a) represents the synthesis of ACF by summing waves with all possible frequencies, with magnitude equal to $|\Phi_\theta(k)|$ and phase $\angle\Phi_\theta(k)$. Note that the ACF is a real valued function, while each contribution $e^{\mathrm{i}k\tau}$ is complex valued. For this reason, the integration includes negative frequencies (which are not physically meaningful) with equal amplitude, yet anti-symmetric phases, in order to cancel the imaginary contribution. This is called the conjugate symmetry property, and for the one dimensional transform is expressed by

$$\Phi_\theta(k) = \Phi_\theta(-k)^* \,, \tag{2.21}$$

where $(\bullet)^*$ denotes the complex conjugate. The two dimensional scenario follows the same ideas, but instead of summing one dimensional sinusoidal waves, two-dimensional waves are used as the basis of functions (see Figure A.6 on page 223). These waves are also characterized by amplitude, phase and frequency, but additionally also by a propagation direction. The information of frequency and direction is given by a wave vector $\boldsymbol{k} = (k_x, k_y)$, whose components are the frequency in each direction. The magnitude of the wave vector equals the frequency of the wave, and the direction gives its propagation direction. The ACF can be rebuilt by spanning all wave vectors, cf. Equation (2.20b). As a consequence of using complex notation, the conjugate symmetry property holds

$$\Phi(\boldsymbol{k}) = \Phi(-\boldsymbol{k})^* \,. \tag{2.22}$$

At this stage, additional details on the adopted notation are given carefully. When autocorrelation function was introduced in Section 2.3.1, the symbol R was used for both profiles and surfaces. Bearing in mind that the physical units associated with ACF are $[$*length*$]^2$, independently of the number of dimensions involved, it seems reasonable to keep the same symbol for both cases. However, this is not verified for PSD of profiles and surfaces. In Equations (2.20), spatial frequency k has units of [*radian per length*]. Therefore, Φ_θ must have units of $[$*length*$]^3$ *per* [*radian*], such that integral in Equation (2.20a) comes in $[$*length*$]^2$. For surfaces, a double integral on spatial frequency is involved, thus Φ must have units of $[$*length*$]^4$ *per* [*radian*]. The difference of physical units between both cases justifies this distinction. As a matter of fact, the PSD of a surface is related to the PSD of a profile taken along the direction defined by θ, so using the same symbol for both would not be consistent. Longuet-Higgins (1957b) related both PSD's by

$$\Phi_\theta(k) = \int_{-\infty}^{+\infty} \Phi(k_x, k_y)\, \mathrm{d}l \,, \tag{2.23}$$

with $l = \sqrt{k_x^2 + k_y^2 - k^2}$. In Equation (2.23), k_x and k_y denote the spatial frequency of the surface wave in each direction, or the projection of its wavevector onto the plane coordinate axis. The variable k denotes the frequency of a sinusoidal wave in a rough profile. The geometrical interpretation of this relation is based on the fact that two dimensional waves have a one dimensional projection along the line defined by θ, whose frequency differs from the wave frequency $\|\boldsymbol{k}\|$, unless θ is coincident with propagation direction of the wave. This means that waves of frequency different than k can contribute to the profile PSD Φ_θ at that frequency, hence one needs to sum the contribution of all waves whose projection along direction θ is a one dimensional wave of frequency k.

Remark 2.4 on the definition of Fourier transform.
In the current work, the following definition for the 1D Fourier transform was adopted

$$F(k) = \mathcal{F}\left\{f(x)\right\} = \int_{-\infty}^{\infty} f(x)e^{-\mathrm{i}kx}\,\mathrm{d}x\,,$$

and the respective inverse transform comes

$$f(x) = \mathcal{F}^{-1}\{F(k)\} = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(k)e^{\mathrm{i}kx}\,\mathrm{d}k\,.$$

These definitions are not rigid, and alternative formulations are found in the literature, by simply moving the factor $1/2\pi$ *from the inverse to the forward transform, and the same regarding the factor* $1/4\pi^2$ *in two dimensional transforms. As a consequence of adopting different definitions for the transform, several relations involving the power spectrum change accordingly to the underlying convention. One must pay attention to the definitions adopted in different works, before comparing certain results involving the surface, or profile, power spectrum. This will have consequences mainly on the topography synthesis from inverse Fourier transform, on the definition of spectral moments and their relation with RMS parameters.*

Relation with height spectrum

For the first time since the start of this chapter, frequency is introduced in the mathematical description of roughness, even though the term *frequency contribution* has already been referred. It is important to note that the PSD, from its definition, holds the frequency content of the ACF, not from the surface itself. The relation between PSD and a characteristics of a rough surface itself is given by the autocorrelation theorem, for Fourier transforms. Applying the 2D version of this theorem, it can be proved that

$$\Phi(\boldsymbol{k}) = \lim_{\substack{L_x\to\infty \\ L_y\to\infty}} \frac{\left|\mathcal{F}\left\{z(x,y)\right\}\right|^2}{L_x L_y}\,, \tag{2.24}$$

and in a similar fashion for rough profiles

$$\Phi_\theta(k) = \lim_{L\to\infty} \frac{|\mathcal{F}\{z(x)\}|^2}{L}\,. \tag{2.25}$$

Equations (2.24) and (2.25) relate the power spectral density of a rough surface (often termed as the surface's spectrum or power spectrum) to the squared magnitude of its Fourier transform. While the Fourier transform of a surface or profile is a complex valued function, its PSD is real valued and always non-negative—it equals the squared magnitude of a complex number. As consequence of being real valued, the PSD has zero phase for every frequency. From this perspective, it can be interpreted as a measure of the frequency content on a surface, describing what frequencies are present and their amplitude. Yet it does not provide any information on the phases associated with each frequency. In fact, the random behavior of rough surfaces is dictated by the randomness of phases. For a clearer understanding on the physical meaning of PSD, different profiles having increasingly complex PSD's are plotted in Figure 2.8. This figure also emphasizes the symmetry of PSD relative to the origin, as a consequence of the conjugate symmetry property.

One can apply the inverse Fourier transform directly to surface height to give

$$z(x,y) = \frac{1}{4\pi^2}\iint_{-\infty}^{+\infty} Z(k_x,k_y)e^{\mathrm{i}(k_x x+k_y y)}\,\mathrm{d}k_x\mathrm{d}k_y\,. \tag{2.26}$$

From the result expressed in Equation (2.24), it writes $|Z(k_x,k_y)| = \sqrt{\Phi(k_x,k_y)L_xL_y}$ and $\angle Z = \phi(k_x,k_y)$ is a random phase, associated with a specific wave vector. The surface height $z(x,y)$ can then be synthesized by its PSD and a random phases field

$$z(x,y) = \frac{1}{4\pi^2}\iint_{-\infty}^{+\infty} \sqrt{\Phi(k_x,k_y)L_xL_y}\; e^{\mathrm{i}(k_x x+k_y y+\phi(k_x,k_y))}\,\mathrm{d}k_x\mathrm{d}k_y\,. \tag{2.27}$$

In Equation (2.27), the limit of sample length to infinity was dropped for cleaner perception of the procedure, but without loss of generality. A similar expression can be written for a rough profile

$$z(x) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} \sqrt{\Phi_\theta(k)L}\; e^{\mathrm{i}(kx+\phi(k))}\,\mathrm{d}k\,. \tag{2.28}$$

Relation with ACF and RMS parameters. Spectral moments

It is interesting to note that some roughness parameters, in particular, the RMS parameters, can be obtained from the PSD. Starting with profile statistics, it can readily be seen from Equation (2.20a) that

$$R(0) = (z_{\mathrm{rms},x})^2 = \frac{1}{2\pi}\int_{-\infty}^{+\infty} \Phi_\theta(k)\,\mathrm{d}k\,, \tag{2.29}$$

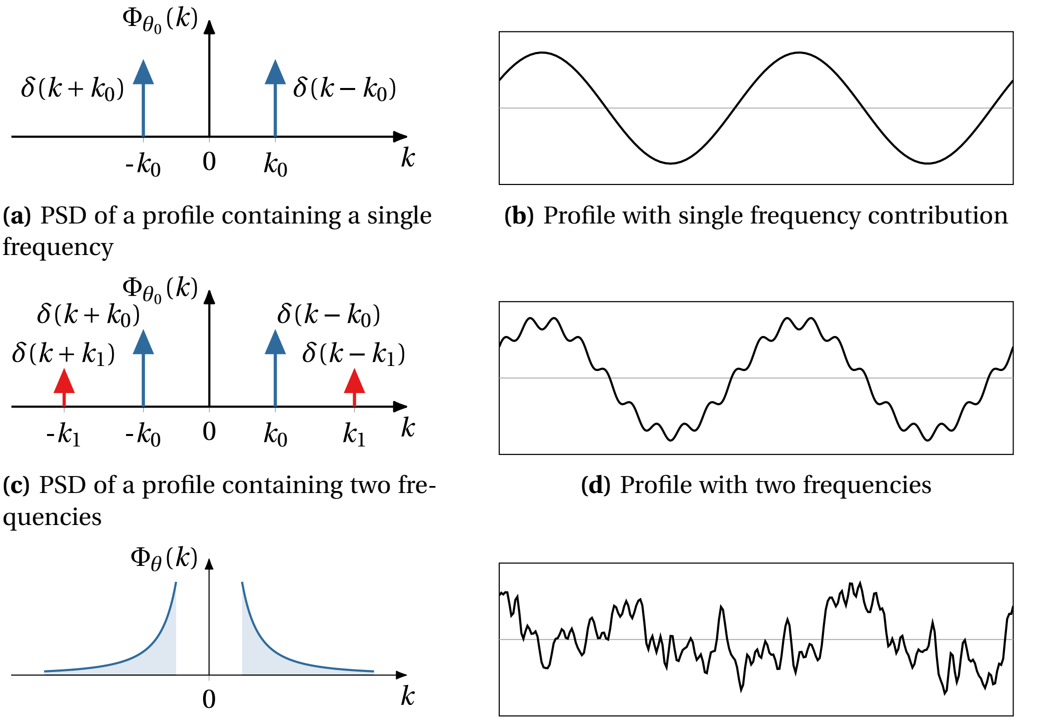
and similarly for surfaces

$$R(0,0) = (z_{\mathrm{rms},xy})^2 = \frac{1}{4\pi^2}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \Phi(k_x,k_y)\,\mathrm{d}k_x\mathrm{d}k_y\,. \tag{2.30}$$

The topography RMS roughness, which can be obtained from the autocorrelation function as the square of its value at the origin, can also be computed by integrating the power spectrum over all frequencies. In fact, similar relations linking the ACF with RMS slope and curvature can be established. For rough profiles, it can be shown that (see Appendix C, in page 229, for the derivation)

$$\left.\frac{\mathrm{d}^2R(\tau)}{\mathrm{d}\tau^2}\right|_{\tau=0} = -\left(z'_{\mathrm{rms},x}\right)^2\,; \tag{2.31}$$

$$\left.\frac{\mathrm{d}^4R(\tau)}{\mathrm{d}\tau^4}\right|_{\tau=0} = \left(z''_{\mathrm{rms},x}\right)^2\,. \tag{2.32}$$



(a) PSD of a profile containing a single frequency

(b) Profile with single frequency contribution

(c) PSD of a profile containing two frequencies

(d) Profile with two frequencies

(e) PSD of a profile containing a wide range of frequencies

(f) Profile having a continuous range of frequencies

Figure 2.8: PSD of different profiles with different frequency contributions. Each frequency in PSD represents the infinitesimal contribution of a single frequency. In Figure 2.8a and Figure 2.8c these contributions are represent by impulse functions $\delta(k)$, in order to reproduce a finite contribution in the function. The PSD is symmetric relative to zero frequency, as observed in Figures 2.8a, 2.8c and 2.8e. A single PSD can represent infinite surfaces, since the particular topography on a surface is ruled by its phases, which are not described by the power spectrum.

Thus, all relevant profile RMS parameters can be computed from the even order derivatives of the ACF at the origin. Regarding rough surfaces, the expressions for the RMS slope and curvature are cumbersome to derive, and shall not be presented here.

A more convenient approach to RMS parameters computation is through the so called spectral moments. These are defined differently for rough profiles and surfaces, respectively, as

$$m_{\theta p} = \frac{1}{2\pi}\int_{-\infty}^{+\infty} k^p \Phi_\theta(k)\,\mathrm{d}k\,; \tag{2.33}$$

$$m_{pq} = \frac{1}{4\pi^2}\iint_{-\infty}^{+\infty} k_x^p k_y^q \Phi(k_x,k_y)\,\mathrm{d}k_x\mathrm{d}k_y\,. \tag{2.34}$$

The spectral moments relate to profile and surface RMS parameters by

$$z_{\mathrm{rms},x} = \sqrt{m_{\theta 0}}\,, \tag{2.35a}$$

$$z'_{\mathrm{rms},x} = \sqrt{m_{\theta 2}}\,, \tag{2.35b}$$

$$z''_{\mathrm{rms},x} = \sqrt{m_{\theta 4}}\,; \tag{2.35c}$$

$$z_{\mathrm{rms},xy} = \sqrt{m_{00}}\,, \tag{2.36a}$$

$$z'_{\mathrm{rms},xy} = \sqrt{m_{20}+m_{02}}\,, \tag{2.36b}$$

$$z''_{\mathrm{rms},xy} = \sqrt{\frac{m_{40}+2m_{22}+m_{04}}{4}}\,. \tag{2.36c}$$

For the sake of completeness, the derivation of Expressions (2.35) and (2.36) is provided in Appendix C. The determination of RMS parameters both from an ACF and PSD, in some way, closes the framework cycle of surface description (with the exception of the height probability distribution, which will be discussed later). Yet, it should be emphasized the importance of both the ACF and PSD: by specifying one of these functions to model roughness, one determines, uniquely, the associated RMS parameters, while the converse is not true.

Remark 2.5 on the definition of spectral moments.
The definitions of spectral moments in Equations (2.33) *and* (2.34) *does not match the traditional expression, commonly found in the literature, where the coefficients* 2π *and* $4\pi^2$ *are not present. Following a previous remark, this is a consequence of the adopted definition for the Fourier transform, and in order to be able to write Expressions (2.35) and (2.36), it was necessary to alter the definition of spectral moment.*

For isotropic rough surfaces, case where the subscript θ can be dropped, since PSD depends only on the magnitude of wave-vector $\|\boldsymbol{k}\|$ and not on its direction, it is verified that (Nayak, 1971)

$$m_{00} = m_0\,; \tag{2.37a}$$

$$m_{20} = m_{02} = m_2\,; \tag{2.37b}$$

$$m_{11} = m_{13} = m_{31} = 0\,; \tag{2.37c}$$

$$3m_{22} = m_{40} = m_{04} = m_4\,. \tag{2.37d}$$

From the profile spectral moments of an isotropic surface, it is possible to compute an extremely important parameter for surface analysis, termed the spectral breadth or Nayak's parameter α, named after the seminal work by Nayak (1971). It is defined by

$$\alpha = \frac{m_0 m_4}{m_2^2}\,. \tag{2.38}$$

For anisotropic surfaces, Equation (2.37) are not verified and, hence, Nayak's parameter cannot be computed explicitly form profile moments. For such cases, Sayles and Thomas (1976) suggest the definition of equivalent profile spectral moments, such that

Equation (2.38) can be applied. The equivalent second and fourth order profile spectral moments come

$$m_{\theta 2} = \sqrt{m_{02} m_{20}}\,; \tag{2.39a}$$

$$m_{\theta 4} = \sqrt{m_{04} m_{40}}\,. \tag{2.39b}$$

Figure 2.9 illustrates the differences in profile topography caused by different profile PSD's. It reformulates Figure 2.5, replacing the ACF plot with the PSD. For an exponential ACF, the profile PSD is given by

$$\Phi_\theta(k) \propto \frac{2\beta}{\beta^2 + k^2}\,. \tag{2.40}$$

Recall that it was verified that the exponential ACF roughness model is scale dependent, thus it is not suited for roughness characterization. A larger autocorrelation length implies more contribution of long wavelengths, while a shorter ACL suggests that low and high frequencies have increasingly similar amplitudes, which is in line with the differences in topography observed earlier.

From the previous relations between spectral moments and derivatives of ACF, the requirement of a fourth order smooth function at the origin for the existence of slopes and curvatures, as referred in Section 2.3.1, is justified. Furthermore, inspecting Figure 2.9 and Equation (2.40), it can be verified that an exponentially decaying ACF implies a non-null zero frequency contribution. The zero frequency contribution can be interpreted as the mean value of function, since the integral of a sinusoidal function over any integer number of periods is zero. As a consequence, the mean value of z will not be zero. This contradicts the hypothesis that roughness z is measured relative to the mean plane, which in turn supports that the exponential ACF can not be verified in all domain.

The preceding discussion started with the definition of PSD, which does also depend on the definition of Fourier transform, in Equation (2.19). It happens that the definition of Fourier transform is not unique, since the coefficients 2π and $4\pi^2$ could be moved from the inverse transform to the forward transform, changing some of previous expressions—as remarked previously. For example, Nayak (1971) uses this coefficients on the forward transform, reaching slightly different relations, for instance between spectral moments and RMS parameters. A point often overlooked in the literature on rough surface description is the distinction between PSD of a profile and of a surface, specially regarding the difference of physical units of both quantities. Not only the definition of Fourier transform is not unique throughout literature, but also the definition of PSD itself does also show inconsistency. In fact, some authors apply the conjugate symmetry property implicitly, and instead of considering negative frequencies, they take only the positive frequencies and double their amplitudes. Jacobs *et al.* (2017) addresses the problem of incoherency of PSD definition, and the reconstruction of real PSD's from theoretical measurements as well. Practical aspects of PSD measurement are also discussed by Nayak (1973), which explores the effect of filtering the surface height data on the profile spectrum. Profile PSD measurement is currently standardized by SEMI MF1811 (2010), which also refers surface measurements, although briefly.

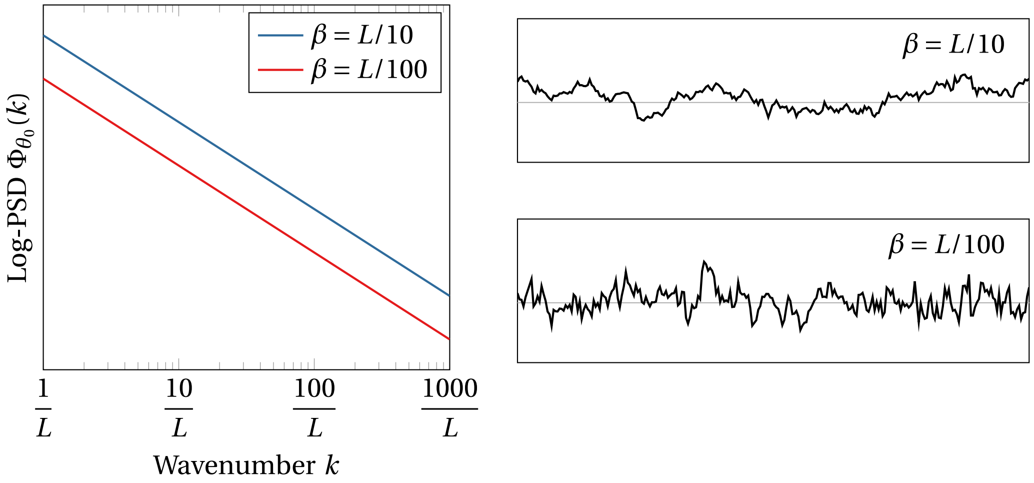


Figure 2.9: Effect of PSD on profile topography. It repeats Figure 2.6 based on the profile PSD, instead of ACF. For the longer ACL, low frequencies, i.e., long wavelengths show a larger contribution in power spectrum, while high frequencies have low amplitude, which explains the overall smoothness of this profile. Regarding the shorter ACL, amplitudes of long and short wavelengths are similar, which causes high frequencies to contribute more to the profile topography.

2.3.4 Gaussian and non-Gaussian surfaces

Alongside with ACF or PSD, the heights distribution allows for a complete description of rough topography (Persson, Albohr, *et al.*, 2005). Actually, this has been suggested earlier in Figure 2.2, when surfaces were classified according to their height probability distribution. The topographic features of profiles in Figure 2.7a and Figure 2.7b differ, since the heights of the former are normally distributed while in the latter they follow a non-Gaussian distribution. The height Probability Density Function (PDF) $f_Z(z)$ is defined as the probability per unit height of having points whose height belongs to an infinitesimal interval around a particular value

$$f_Z(z)\mathrm{d}z = \Pr\left(\mathcal{Z} \in [z,\ z+\mathrm{d}z]\right) . \tag{2.41}$$

Other quantity, the height cumulative distribution function (CDF), measures the probability of the surface height being equal or smaller than some value

$$F_Z(z) = \Pr\left(\mathcal{Z} \in [-\infty,\ z]\right) = \int_{-\infty}^{z} f_Z(t)\ \mathrm{d}t\,. \tag{2.42}$$

The PDF is usually described by its first four central moments, for practical purposes (Elderton and N. L. Johnson, 1969). The central moment of order i of a probability density function is defined as

$$\mu_i = \int_{-\infty}^{+\infty} f_Z(z)\left(z-\mu_z\right)^i\ \mathrm{d}z, \quad \text{for } i = 2,\dots,\infty\,. \tag{2.43}$$

Note that the first-order moment is not included in the central moments definition. The first-order *non-central* moment is the mean value, which is used in Equation (2.43)—the respective central moment is necessarily zero. The mean value is determined by

$$\mu_z = \int_{-\infty}^{+\infty} f_Z(z) z \, \mathrm{d}z \, . \tag{2.44}$$

From this definition and knowing that the integral of f_Z over all domain is one, it is immediate to see that the first-order central moment is null. The second-order central moment equals the variance of the distribution

$$\sigma_z^2 = \int_{-\infty}^{+\infty} f_Z(z) \left(z - \mu_z\right)^2 \, \mathrm{d}z \, . \tag{2.45}$$

By taking the square root of variance, one gets the standard deviation of the heights. In Expressions (2.15) the symbol σ_z was used to denote the sample standard deviation, which might not match exactly the standard deviation of the height distribution, yet it provides a good estimation, once ergodicity is assumed. A Gaussian distribution of surface heights is fully described by these two moments. Again, when dealing with rough surfaces, it is tacitly assumed that $\mu_z = 0$. The third-order moment normalized by the cube of standard deviation is called *skewness*, frequently denoted by γ_1 for a continuous variable, and writes

$$\gamma_1 = \frac{1}{\sigma_z^3} \int_{-\infty}^{+\infty} f_Z(z) \left(z - \mu_z\right)^3 \, \mathrm{d}z \, . \tag{2.46}$$

Finally, the fourth moment normalized by the fourth power of standard deviation is termed *kurtsosis*, β_2, which is given by

$$\beta_2 = \frac{1}{\sigma_z^4} \int_{-\infty}^{+\infty} f_Z(z) \left(z - \mu_z\right)^4 \, \mathrm{d}z \, . \tag{2.47}$$

For a Gaussian distribution, $\gamma_1 = 0$ and $\beta_2 = 3$. Sometimes, it is common to refer to the *excess of kurtosis* γ_2, which is the difference of kurtosis relative to normal distribution, i.e.

$$\gamma_2 = \beta_2 - 3 \, . \tag{2.48}$$

Distributions having positive excess of kurtosis are called *leptokurtic*, and the ones with negative excess of kurtosis are termed *platykurtic*. Skewness and kurtosis must verify the following inequality (Shohat, 1929)

$$\beta_2 \geq \gamma_1 + 1 \, . \tag{2.49}$$

Skewness depends on the odd power of the distance to the mean value, then surface summits and valleys have different impact on this parameter. In fact, skewness is a measure of the height distribution asymmetry. Positively skewed surfaces tend to show summits much taller than valleys, and the heights are concentrated on the base of the surface, while negatively skewed surfaces show deeper valleys when compared to summits, and heights are concentrated near the surface's top. This can be observed in the heights PDF, noting that positively skewed surfaces have a longer tail for positive heights, and the peak is on the left of the mean. The opposite happens for negatively skewed surface, which hold a longer tail for negative heights, and the distribution peak is on the right of the mean value. Regarding kurtosis, it involves a fourth power of heights, meaning that it mainly concerns the presence of outliers, i.e., points which are much taller (or deeper) than the average surface.

The effect of skewness and kurtosis on the height distribution can be seen in Figure 2.10. Skewness pushes the distribution to the left of right, depending whether it is positive of negative . Kurtosis most relevant effect is on the tails of the distribution, which are representative of the outliers. Larger values of kurtosis suggested longer and higher tails—Figure 2.10b is plotted in semi-logarithmic scale so that it emphasizes the impact of kurtosis on the extremes of the PDF. In Figure 2.11, several roughness profiles with same ACF, but varying values of skewness and kurtosis are depicted. Combining skewness and kurtosis, one can get very different topographies, either having deeper valleys, or taller peaks.

2.3.5 Statistical geometry of isotropic Gaussian surfaces

Gaussian surfaces have been the foundation of rough surface analysis, motivated by the extensive knowledge that has been built on the analytical expression for its PDF, and thus, is a trustworthy and simple tool for any investigation. The problem of characterizing rough surface statistics is drastically simplified if the surface is assumed to be Gaussian and isotropic. In fact, these two hypothesis make the problem amenable for analytical treatment. Surfaces satisfying, nearly, both criteria can be produced, e.g., by shot-peening, yet very few other examples are know to respect both properties (K. L. Johnson, 1987).

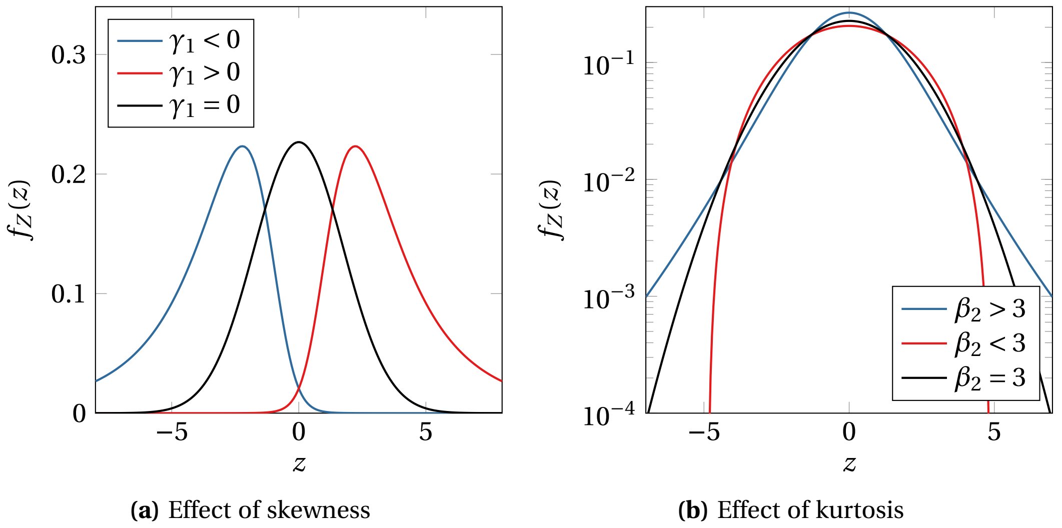
Following the work of Longuet-Higgins (1957b,a) on the statistical geometry of random moving surfaces, Nayak (1971) developed a similar theory concerning static, random, isotropic and Gaussian surfaces. Nayak's analysis is mostly based on the profile power spectrum and its spectral moments, from which he introduces the spectral breadth α—called bandwidth parameter in his work. By using this parameter, Nayak proceeds to compute profile and surface geometry statistics, which are only dependent on α, following a continuous description of surface height, i.e., with a vanishing small sampling interval. In particular, he derives analytical expressions for the probability distribution of summit heights (the probability of existing a summit with a certain height), spatial density of summits and the probability distribution of summit mean curvature. Nayak's theory was further complemented by Bush, Gibson, and Keogh (1976), who presented a closed-form solution for the joint probability density of summit height and curvature in two orthogonal directions, i.e., the probability of existing a summit with a particular value of height an curvature, in each direction.

The lower bound for α is 1.5, regarding isotropic surfaces, and can assume arbitrarily large values. From Nayak's theory, it can be concluded that the summit heights distribution of a Gaussian surface is non-Gaussian, yet it approximates a Gaussian curve for increasing values of α. Additionally, higher summits are sharper (largest curvature), in general, but increasing α reduces the difference between mean curvature at different summit heights—curvature is increasingly steadier with summit height, with increasing α. Nayak extended his analysis for profile statistics, and compared the results with the surface counterpart. It is found that the profile peak heights distribution distorts the summits distribution, and predicts peaks smaller than summits. This distortion is more aggressive for α closer to the limiting value 1.5. Furthermore, for $\alpha < 2.5$ profile statistics suggest lower peak mean curvature and slope, relative to the respective surface parameters. This theory gives good results even for slightly non-Gaussian heights distribution (Thomas, 1999; Sayles and Thomas, 1976).

Since Nayak's theory concerned continuous surfaces, it was unsuited for practical applications involving discrete topography measurements. Whitehouse David J. *et al.* (1978, 1982) extended Nayak's theory to account for a finite sampling interval, and provided results which could be related with real discrete profiles and surfaces. Discrete profile peaks were found from a three point scheme, such that a point which is higher than the previous and next point is considered a peak. For surface summits, a five point stencil was used. Greenwood J. A. and Langstreth J. K. (1984) reformulated Whitehouse and Phillps' ideas in a cleaner theory. They emphasized that a 5 point summit may not coincide with a real summit, but with ridges or settle points with particular orientations. In contrast, three point peaks are always peaks. In this new theory, discrete surface and profile statistics were related to RMS parameters (roughness, slope and curvature) as a function of sampling interval. The most relevant results are the weak dependence of summit and summit curvature distribution relative to sampling interval and that mean summit curvature is approximately equal to mean peak curvature. Furthermore, summit distributions are nearly Gaussian for all sampling intervals. Even though the distribution of non-dimensional slopes and curvatures does not depend strongly on sampling interval, their numerical (dimensional) value does depend strongly on it, as it was observed earlier—RMS slope and curvature are very sensitive to short wavelengths.

Several surfaces are indeed verified to be Gaussian, even if anisotropic (Persson, Albohr, *et al.*, 2005; Thomas, 1999). The major caveat in approximating surface height by a Gaussian distribution is the lack of representativeness in the tails of the distribution. There is a clear bound on summit height for real surfaces, which suggests a non-Gaussian distribution of heights near the tails—a truncated Gaussian would be adequate. Even though this seems a small approximation, contact mechanics applications concern mainly the tails of the distribution, where most of surface summits are present—the summits will be the first points in contact, so they require special care in this situation (Bhushan, 1998; Thomas, 1999). Some authors tried to fit other distribution to collections of heights measurement in order to find a better fit, yet Gaussian distribution prevails up to the date as reference (Thomas, 1999).

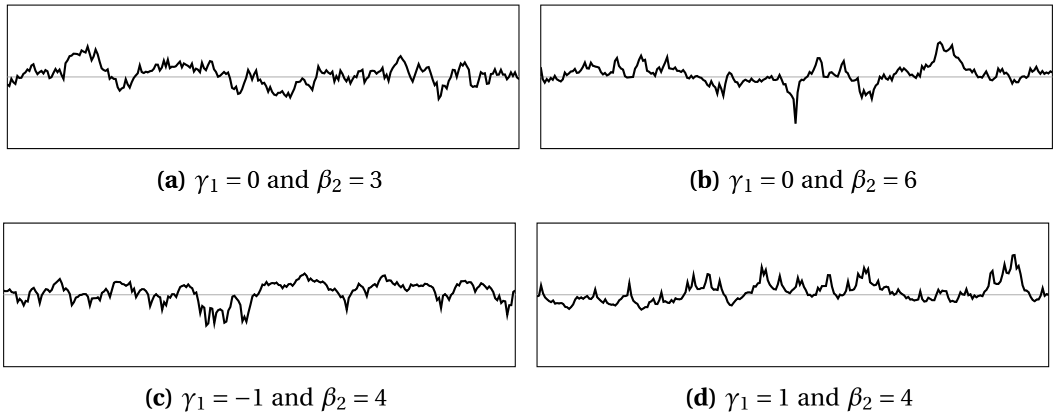
In contrast, most real surfaces are, in fact, non-Gaussian. A metal surface after grinding is certainly non-Gaussian, since this process removes mostly roughness summits, while leaving valleys intact—it would produce a negatively skewed surface. Figure 2.12 shows typical ranges for profile skewness and kurtosis, produced by a different conventional machining processes. Typical values of skewness lie between -2 and 2, while for kurtosis, value between 2 and 10 are reasonable. Nevertheless, value of skewness down to -6 and of kurtosis high as 100 have been measured in practice (Minet *et al.*, 2010).



(a) Effect of skewness

(b) Effect of kurtosis

Figure 2.10: Effect of skewness and kurtosis on the shape of PDF. The effect of skewness is portrayed in Figure 2.10a, and compared to a normal distribution. The three distributions have different mean values. A positively skewed distribution is *pushed* to positive values of z, while for negative skewness, the distribution is pushed for negative values of z. Figure 2.10b shows the effect of kurtosis, also compared with a normal distribution. This plot is presented with f_Z axis in log-scale in order to emphasize that the effect of kurtosis is mainly on the tail, which is associated with increasingly higher probability with increasing kurtosis.



(a) $\gamma_1 = 0$ and $\beta_2 = 3$

(b) $\gamma_1 = 0$ and $\beta_2 = 6$

(c) $\gamma_1 = -1$ and $\beta_2 = 4$

(d) $\gamma_1 = 1$ and $\beta_2 = 4$

Figure 2.11: Effect of skewness and kurtosis on the topography of a rough profile. All profiles have the same ACF. Figure 2.11a show a Gaussian profile, and Figures 2.11b to 2.11d non-Gaussian profiles. By changing only the value of kurtosis, Figure 2.11b, one gets a more *peaky* profile, both in valleys and peaks. Introducing a value for skewness, one biases the peakedness of the profile for valleys (Figure 2.11c) or peaks (Figure 2.11d).

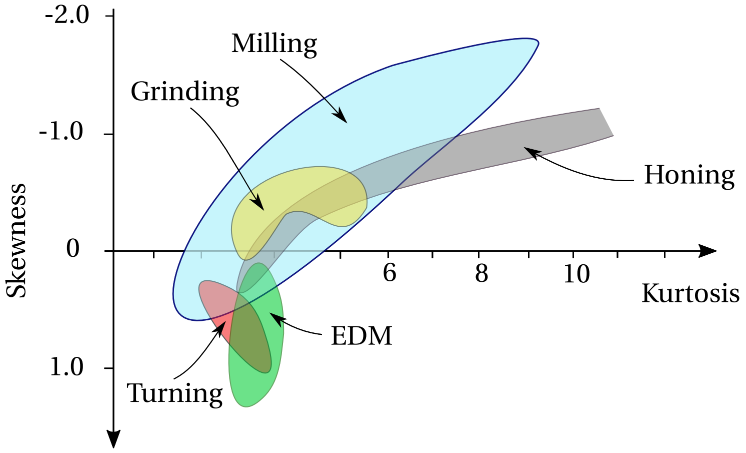


Figure 2.12: Typical values of skewness and kurtosis for surface produced by common machining processes. Representative values for skewness lie between -2 and 2, and between 1 and 10 for kurtosis. EDM stands for Electrical Discharge Machining. Adapted from Whitehouse (1994).

2.4 Self-affine rough surfaces and profiles. Fractal roughness

The mathematical tools examined earlier provide a complete framework for rough topography characterization. However, rough surface models, such as the exponential ACF, fail in providing a scale independent description this is, no consistent surface property has yet been discussed. Sayles and Thomas (1978) stated that rough surfaces are not stationary processes, based on the observation that profile power spectrum of several real surfaces increase with increasing cut-off length. In other words, roughness, and in particular, RMS roughness, increases with sampling length. If variance is dependent on sampling length, then rough topography violates the stationarity hypothesis. Furthermore, these authors verified that power spectrum of real surfaces followed a power law like

$$\Phi_\theta(k) = Bk^{-2}\,, \tag{2.50}$$

where B was termed the surface *topothesy* and has units of length. It is difficult to understand the physical interpretation of this quantity; sometimes, it is referred as the length of the cord over which a slope of 1 radian occurs (Russ, 1994). With this argument, B seemed a real property of the surface, because it was not dependent of sampling length, and the power-law in Equation (2.50) would be uniquely defined. In practice, Equation (2.50) was not verified for all surfaces, and it was reformulated such that it can accommodate other exponents for the power spectrum

$$\Phi_\theta(k) = Bk^{-c}\,. \tag{2.51}$$

In a logarithmic representation of Φ_θ as a function of k, the power law shows as a line, whose slope is equal to $-b$ and the intersect is $\log B$. Profiles whose spectrum verify a power law are self-affine profiles, which are a type of fractal curves (B. B. Mandelbrot, 1983; Peitgen and Saupe, 2012).

Fractal theory deals with geometric entities that are not encompassed by Euclidean geometry. For example, a fractal profile or surface is continuous, yet non-differentiable at all points. Fractal curves have infinite perimeter, yet finite enclosed area, which is other property that makes fractals unique geometrical constructions. Likewise, fractal surfaces have infinite surface area and finite volume. Fractals show an unfolding symmetry, technically termed self-similarity or self-affine behavior, i.e., they look similar at different magnifications. As an illustration, Figures 2.13 and 2.14 show two fractal curves which are self-similar and self-affine, respectively. The self-similar fractal in Figure 2.13 scales equally in both directions, meaning that when a magnification is applied, say in the horizontal direction, the incremental curve, scaled with the same magnitude, is appended to the previous magnification. Repeating this iterative process produces a self-similar geometry, which looks exactly equal at different magnifications. In contrast, self-affine fractals show different magnification for different directions. This can be seen in Figure 2.14, where the scale of roughness height becomes smaller with increasing horizontal magnification. Figures 2.13 and 2.14 also present other distinction between fractal geometries. In Figure 2.13, the curve is generated by a well defined rule, that states how to add new geometry features after each iteration, while for a rough profile in Figure 2.14, such rule do not exist. In fact, fractals does not need to have a regular structure, and quite complex shapes like the coastline of countries are fractal (B. Mandelbrot, 1967). The term *self-affine* for such cases imply a statistical resemblance, rather than an exact one.

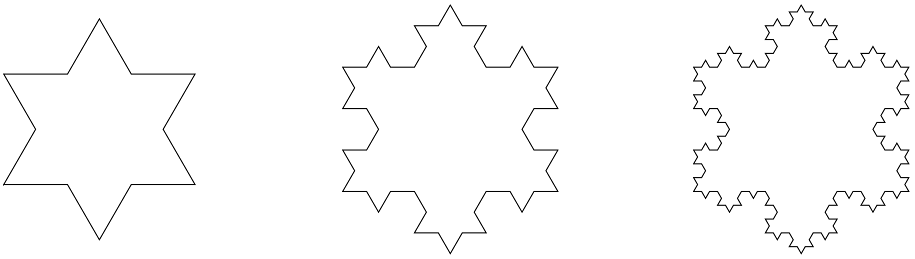


Figure 2.13: The Koch snowflake, an example of self-similar fractal. Increasing the level of magnification, i.e., the required detail, the curve is progressively more complex, and each new small geometry increment added at each magnification is a scaled version of the geometry increment at previous magnification.

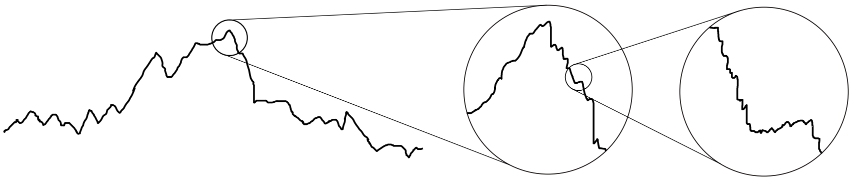


Figure 2.14: Example of self-affine rough profile. As the level of magnification increases, new roughness profiles are added the previous one. Yet, the roughness profile scale differently in horizontal and vertical direction, hence it is called self-affine, not self-similar.

Self-affine geometry is of special interest in surface analysis, as fractals provide mean to characterize roughness across several scales. More specifically, rough topography can be treated as a multiscale construct, on which nanoscale roughness exists on top of microscale roughness, and the same from the microscale to the next larger scales. The idea of roughness covered in roughness was proposed early by Archard, J. F. (1957), yet long before fractals where mathematically described.

Fractals are characterized by their fractal dimension, which is different from their topological dimension. The topological dimension of a self-affine profile is 1, since it is a curve. However, its fractal dimension D_p lies between 1 and 2, because it fills more area than a non-fractal curve but less than a surface. This argument similarly applies to self-affine surfaces, whose fractal dimensions D_s lies between 2 and 3. The fractal dimension of a profile taken from an isotropic surface, which results from the interception between that surface and vertical plane, is related with the fractal dimension of the surface by (Russ, 1994)

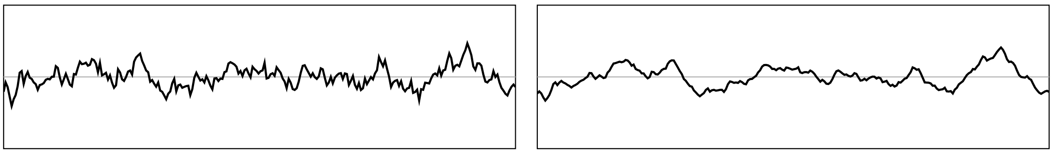
$$D_p = D_s - 1\,. \tag{2.52}$$

Commonly, fractal dimension is expressed by the Hurst roughness exponent, defined as the difference between the upper bound for a fractal dimension and the fractal dimension itself, this is

$$H = 2 - D_p\,; \tag{2.53a}$$

$$H = 3 - D_s\,. \tag{2.53b}$$

Fractal dimension and Hurst exponent measure the space-filling capacity of the topography. Higher fractal dimensions, conversely, lower Hurst exponents, indicate more space filling fractals. This can be observed in Figures 2.15 and 2.16, that show rough profiles and surfaces with varying Hurst exponent. This roughness measure takes values between 0.7 and 0.9, for several surfaces (B. B. Mandelbrot *et al.*, 1984).



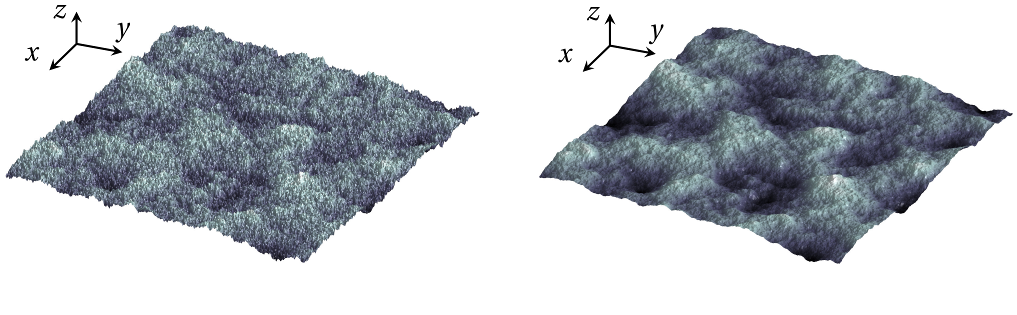
(a) $H = 0.3$

(b) $H = 0.8$

Figure 2.15: Effect of Hurst roughness exponent on profile topography. Both profiles have same RMS roughness and same phase field. Lower Hurst exponent increases the contribution of high frequency, creating a more space filling profile, when compared with a higher Hurst exponent.

One aspect of particular interest is that the PSD of self-affine isotropic rough surfaces can be related to their fractal properties. In particular, it can be expressed in a multitude of formulas, by using surface fractal dimension, fractal dimension of profiles contained in the surface or using Hurst exponent. This relation writes (Russ, 1994; J.-J. Wu, 2000a; Yastrebov, Anciaux, *et al.*, 2015)

$$\Phi(k_x, k_y) = \frac{g^{2D_p-2}}{\left(k_x^2 + k_y^2\right)^{3-D_p}} = \frac{g^{2D_s-4}}{\left(k_x^2 + k_y^2\right)^{4-D_s}} = \frac{g^{2(1-H)}}{\|\boldsymbol{k}\|^{2(H+1)}}\,. \tag{2.54}$$



(a) $H = 0.3$ **(b)** $H = 0.8$

Figure 2.16: Effect of Hurst roughness exponent on surface topography. Both profiles have same RMS roughness, phase field and both are isotropic. Lower Hurst exponent increases the contribution of high frequency, creating a more space filling surface, similar to what can be observed in Figure 2.15.

Regarding the profile PSD, it comes from the surface PSD by Equation (2.23), and reads[9]

$$\Phi_\theta(k) = \frac{G^{2D_p-2}}{k^{5-2D_p}} = \frac{G^{2(1-H)}}{k^{1+2H}} . \tag{2.55}$$

Thus, both self-affine surfaces and profiles have a power law PSD, whose slope in a log-log plot is related to the Hurst exponent. Constants g and G in Equations (2.54) and (2.55) are the surface and profile fractal scale constants, which express the absolute scale of roughness, while H gives the relative scale between frequency contributions.

Several engineering surfaces have been confirmed to be fractal, for example, machined and fracture surfaces (Russ, 1994). The fractal description of roughness does seem to be the answer for a scale independent characterization of surfaces. From experimental measurements, and power spectrum computation, the power law for PSD, thus self-affinity, has been observed for several cases. For example, Panda *et al.* (2016), which was cited earlier for his results showing that autocorrelation length is scale dependent, verified for several sampling lengths and sampling interval that the PSD matched for every case, with exception at high wavelengths, where instrument filtering shows a critical influence. Also in Persson (2014), the PSD computed from measurements at different scales with different instruments and resolutions, the invariance of the PSD is confirmed. Hence, PSD is mostly unbiased by sampling length and size. Experimental results for Hurst exponent can be repeated with good accuracy, however this is not always verified for fractal scale constants. These still seem to be instrument dependent in some cases, as for the experimental result from Majumdar and Bhushan (1991). Some discussion still exists about the topic

Real surfaces cannot be fractal at all scales. For wavelengths reaching atomic scale, fractal behavior ceases, alongside with the continuum hypothesis. Also, for large wavelengths, most surfaces do not evidence fractal properties. In practice, real surfaces have the PSD represented in Figure 2.17. The fractal behavior is observed between a high frequency

[9] For a complete derivation see J.-J. Wu (2000a).

k_s and a lower frequency, called *roll-off frequency* k_r, where the fractal behavior is lost. Between the roll-off frequency and a even lower frequency k_l a plateau of constant PSD is observed (Persson, Albohr, *et al.*, 2005; Dodds and Robson, 1973; Vallet *et al.*, 2009). For frequencies higher than k_s the fractal behavior is lost, and the PSD is often assumed to be null. The same happens for wavelengths larger than k_l, which is related to the sampling length. The plateau can be justified, e.g., for machined surfaces, because machining typically reduces high amplitude longer wavelengths, while keeping short wavelengths intact. Thus, machined surfaces are fractal at least at small scales. The determination of the cut-off frequencies k_l and k_s for different cases remains an open question. For these cases, the profile and power spectrum can be written with considerably more convenient expressions (C_0 and C_0' are just scale constants):

$$\Phi(k_x, k_y) = \begin{cases} C_0 & , \quad k_l \leq \|\boldsymbol{k}\| < k_r \\ C_0 \left(\dfrac{k_r}{\|\boldsymbol{k}\|} \right)^{2(H+1)} & , \quad k_r \leq \|\boldsymbol{k}\| \leq k_s \\ 0 & , \quad \text{elsewhere}\,; \end{cases} \tag{2.56}$$

$$\Phi_\theta(k) = \begin{cases} C_0' & , \quad k_l \leq k < k_r \\ C_0' \left(\dfrac{k_r}{k} \right)^{1+2H} & , \quad k_r \leq k \leq k_s \\ 0 & , \quad \text{elsewhere}\,. \end{cases} \tag{2.57}$$

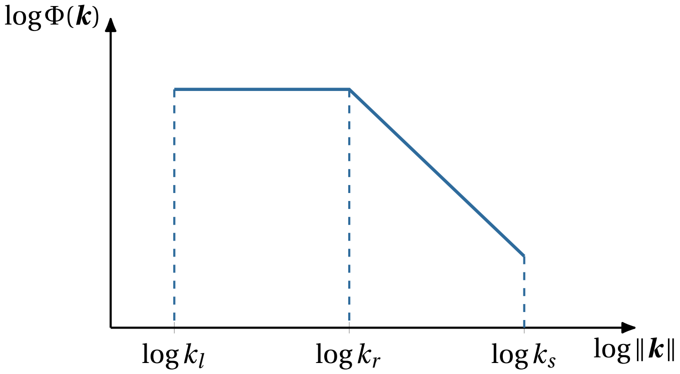


Figure 2.17: Typical PSD of an isotropic rough surface. Fractal behavior is observed between the frequency k_s associated with a short wavelength and a roll-off frequency k_r. For wavelengths larger than the roll-off wavelength there is a plateau of approximately constant Φ. For frequencies higher than k_s or lower than k_l the PSD is truncated, i.e., it is assumed null.

Other surfaces are multi-fractal. For these cases, different fractal behaviors dependent on the scale of observation are observed (Thomas, 1999). Using Figure 2.17 as reference, the PSD of a bifractal surface would have a Hurst exponent between k_r and k_s and other Hurst exponent between k_l and k_r, thus, different slopes at this two ranges. Surfaces produced by several processes, e.g., several machining processes, may verify a multi-fractal

PSD, with a fractal dimension relating to each process. A less practical example, yet certainly useful for understanding multi-fractal surfaces, concerns mountainous terrain covered with vegetation. Mountain topography, like roughness topography, is a fractal with a particular dimension at a large scale. Vegetation, which shows up at smaller scales, is also fractal with a different dimension. A broader justification for the restriction of fractal behavior for a specific scale is that no process acts across all scales.

It is interesting to confirm previous results on the scale dependence of RMS parameters, based on the freshly introduced fractal ideas. Considering a self-affine profile, whose power spectrum is given by Equation (2.55), and assuming that the spectrum is bounded at large scales by $k_l = k_r$ and at small scales by k_s, RMS roughness, slope and curvature comes from the profile spectral moments Equation (2.35)

$$z_{\mathrm{rms},x} = \int_{k_l}^{k_s} \frac{C_1}{k^{1+2H}}\,\mathrm{d}k = C_2\left(\frac{1}{k_s^{2H}} - \frac{1}{k_l^{2H}}\right) \approx C_3\lambda_l^{2H}\,; \tag{2.58a}$$

$$z'_{\mathrm{rms},x} = \int_{k_l}^{k_s} \frac{C_1 k^2}{k^{1+2H}}\,\mathrm{d}k = C'_2\left(\frac{1}{k_s^{-2+2H}} - \frac{1}{k_l^{-2+2H}}\right) \approx C'_3\lambda_s^{2-2H}\,; \tag{2.58b}$$

$$z''_{\mathrm{rms},x} = \int_{k_l}^{k_s} \frac{C_1 k^4}{k^{1+2H}}\,\mathrm{d}k = C''_2\left(\frac{1}{k_s^{-4+2H}} - \frac{1}{k_l^{-4+2H}}\right) \approx C''_3\lambda_s^{4-2H}\,. \tag{2.58c}$$

The constants C are included to express the idea of proportionality, although their exact value is not important for the ongoing discussion. In the last step of all previous equations, the wavenumber is substituted by the respective wavelength λ, and the following expressions are derived assuming $k_l << k_s$. The value of RMS roughness is dominated by the large wavelength, and increases with increasing longer wavelength, thus with sampling size, as already mentioned. Concerning RMS slope and curvature, since $2H < 2$ (cf. the definition of Hurst exponent) it is observed that both parameters are dominated by the shortest wavelength, or high frequency, and increase with increasing high frequency. Curvature increases faster the slope, because it involves a higher power of the short wavelength. From this analysis, one can conclude that there must be a high frequency filter in order to enable measurements on fractal surfaces, either in the measurement device itself or by a change in fractal behavior. This also justifies the unbounded increase on summit density predicted by Nayak (1971), since high frequencies create small summits around larger ones, progressively increasing summit density.

2.5 Anisotropy

It was referred in Section 2.3.1 that an exponential autocorrelation function in two dimensions could be used to describe anisotropic surfaces. Longuet-Higgins (1957b) discussed briefly anisotropy in his works, but only found expressions for the spatial density of summits. Nayak (1973) concluded that the problem of anisotropic statistical geometry could be reduced to the determination of seven invariants from five non-parallel rough profiles, which is quite unpractical. McCool (1978) characterized anisotropy by the ratio of maximum and minimum $m_{\theta 2}$ along several directions.

The concept of strong anisotropy was introduced by Bush, Gibson, and Keogh (1979), in contrast to that of weak anisotropy. Weakly anisotropic are stretched isotropic surfaces, while strong anisotropic surfaces show a distinct lay along which roughness exists. This lay can be interpreted, for example, as the direction of machining marks in a surface. Bush, Gibson, and Keogh (1979) stated that for a strongly anisotropic surface, the problem is reduced to the determination of 5 invariants, that can be computed from only 2 profiles, one perpendicular to the lay, and other parallel to it, providing very convenient strategy for practical applications.

The extension of the fractal theory to anisotropic surfaces is not straightforward. For an isotropic surface one can speak of a surface fractal dimension, that relates to the dimension of each profile, which is equal for all profiles. However, when dealing with anisotropic surfaces, profiles may have different fractal dimensions, and the concept of surface fractal dimension becomes blurry. Russ (1994) suggested that for a weakly anisotropic surface, the profile fractal dimension does not depend on the profile direction, but topothesy (or fractal scaling factor) does. Concerning strong anisotropy, Hall and Davies (1995) observed that the fractal dimension is constant nearly in every direction, expect along the lay, where it decreases rapidly. J.-J. Wu (2002) proposed analytical expressions for the PSD of both weakly and strongly anisotropic surface, based on the prepositions of Russ (1994) and Hall and Davies (1995), in order to perform numerical generation of random rough surfaces. For weakly anisotropic surface, Wu suggests the following PSD:

$$\Phi(k_X, k_y) = \frac{g^{2D_p-2}}{\left[\left(\frac{k_x}{b_x}\right)^2 + \left(\frac{k_y}{b_y}\right)^2\right]^{3-D_p}} = \frac{g^{2(1-H)}}{\left[\left(\frac{k_x}{b_x}\right)^2 + \left(\frac{k_y}{b_y}\right)^2\right]^{1+H}}, \tag{2.59}$$

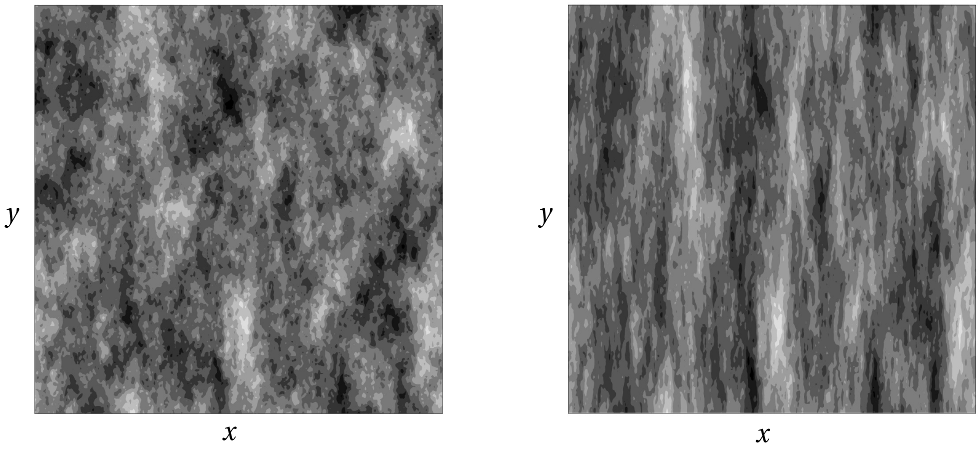
where b_x and b_y responsible for changing the profile scale factor along different directions and D_p is the profile fractal dimension of any profile in the surface. With respect to strong anisotropy, the same author proposes the following relation

$$\Phi(k_x, k_y) = \frac{J_x^{2D_p^x-2}\Delta(k_y)}{k_x^{5-2D_p^x}} + \frac{J_y^{2D_p^y-2}\Delta(k_x)}{k_y^{5-2D_p^y}} = \frac{J_x^{2(1-H_x)}\Delta(k_y)}{k_x^{1+2H_x}} + \frac{J_y^{2(1-H_y)}\Delta(k_x)}{k_y^{1+2H_y}}. \tag{2.60}$$

Here, $\Delta(\cdot)$ is a function which verifies $\Delta(0) = 1$, and is null elsewhere. Therefore, this PSD is null everywhere except along the lines $k_x = 0$ and $k_y = 0$. Opposed to what J.-J. Wu (2002) presented, the scale constants for strongly anisotropic rough surfaces are denoted here by J instead of G, to account for the difference between physical units of Φ and Φ_θ. Figures 2.18 and 2.19 show different textures, characterized by weak and strong anisotropy, respectively.

2.6 Application to discrete surfaces and profiles

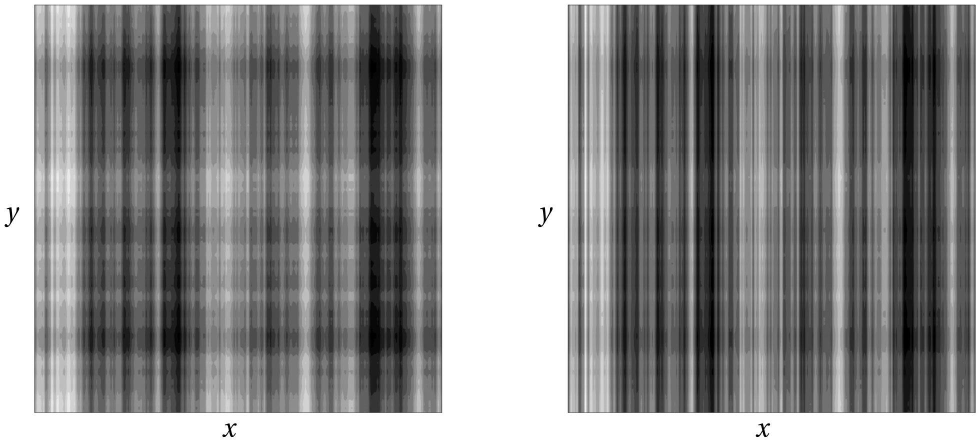
Most previous concepts, specifically those concerning random processes and spectral analysis of rough surfaces, were established for continuous topography. When working



(a) $H = 0.3$ and $b_x = 2b_y$

(b) $H = 0.3$ and $b_x = 4b_y$

Figure 2.18: Weakly anisotropic rough surfaces. Both surfaces have the same value of RMS roughness, and also the same field. By stretching Figure 2.18a on obtains Figure 2.18b.



(a) $H = 0.3$ and $J_y = 20J_x$

(b) $H = 0.1$ and $J_y = 20J_x$

Figure 2.19: Strongly anisotropic rough surfaces. The surface is very similar to an extrusion of a profile along x, in the y direction.

with data from real measurements, or with artificially synthesized surfaces, one needs to reformulate those concepts such that they can be applied for discrete surfaces and profiles. RMS parameters in Section 2.2 have already been derived assuming a discrete description of roughness. The extension of ACF, skewness and kurtosis for discrete data is straightforward, and follows the same idea. Firstly, consider the following: a discrete profile, sampled at N equally spaced points, identified by index n; a discrete surface sampled on a grid of N points in the x direction and M points in the y direction, denoted by the symbols n and m, respectively. The discrete estimate of ACF comes as

$$\hat{R}_q = \frac{1}{N-q}\sum_{n=0}^{N-1-q} z_n z_{n+q}, \qquad q = 0,1,2,....,N-1\,; \tag{2.61a}$$

$$\hat{R}_{p,q} = \frac{1}{(M-p)(N-q)}\sum_{m=0}^{M-1-p}\sum_{n=0}^{N-1-q} z_{m,n} z_{m+p,n+q}, \quad \begin{cases} p = 0,1,...,M-1 \\ q = 0,1,...,N-1 \end{cases} . \tag{2.61b}$$

One refers to $\hat{R}$ as an *estimate of the ACF*, because it is computed from a discrete set of data and, consequently, it will differ from R at the sampling points, in general. Also, note that with increasing shift, less points are available for ACF computation, and thus estimates become less reliable as the shifts become larger. Skewness and kurtosis can also be readily reformulated, assuming a zero mean data set. It is noteworthy to mention that skewness and kurtosis are stable with sampling size (Mainsah *et al.*, 2013). The discrete computation of these PDF moments follows

$$R_{sk} = \frac{1}{R_q^3}\frac{1}{N}\sum_{n=0}^{N-1} z_n^3\,; \tag{2.62a}$$

$$S_{sk} = \frac{1}{S_q^3}\frac{1}{MN}\sum_{m=0}^{M-1}\sum_{n=0}^{N-1} z_{m,n}^3\,; \tag{2.62b}$$

$$R_{ku} = \frac{1}{R_q^4}\frac{1}{N}\sum_{n=0}^{N-1} z_n^4\,; \tag{2.63a}$$

$$S_{ku} = \frac{1}{S_q^4}\frac{1}{MN}\sum_{m=0}^{M-1}\sum_{n=0}^{N-1} z_{m,n}^4\,. \tag{2.63b}$$

The discretization of Fourier transforms must be performed both in spatial and frequency. This procedure is accomplished by the Discrete Fourier Transform (abbreviated DFT, see A.4 for more details). The DFT of a rough topography comes

$$\mathrm{DFT}(z_n) = Z_q = \sum_{n=0}^{N-1} z_n e^{-\mathrm{i}2\pi qn/N}, \qquad q = 0,1,2,....,N-1\,; \tag{2.64a}$$

$$\mathrm{DFT}(z_{m,n}) = Z_{p,q} = \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} z_{m,n} e^{-\mathrm{i}2\pi(qm/M+pn/N)}, \quad \begin{cases} p = 0,1,...,M-1 \\ q = 0,1,...,N-1 \end{cases} ; \tag{2.64b}$$

and the inverse transform (IDFT) reads

$$\text{IDFT}(Z_q) = z_n = \frac{1}{N}\sum_{q=0}^{N-1} Z_q e^{\text{i}2\pi qn/N}, \qquad n = 0,1,2,....,N-1\,; \tag{2.65a}$$

$$\text{IDFT}(Z_{p,q}) = z_{m,n} = \frac{1}{MN}\sum_{p=0}^{M-1}\sum_{q=0}^{N-1} Z_{p,q} e^{\text{i}2\pi(qm/M+pn/N)}, \quad \begin{cases} m = 0,1,...,M-1 \\ n = 0,1,...,N-1 \end{cases}. \tag{2.65b}$$

The DFT writes a discrete rough profile as the superposition of N discrete sinusoidal waves with frequencies equally spaced between 0 and $\Omega_s/2$. The symbol Ω_s denotes the sampling frequency, which is related to the sampling interval l_s (defined as the spacing between sampled points) by

$$\Omega_s = \frac{2\pi}{l_s}\,. \tag{2.66}$$

The same holds for the two dimensional case, where the frequencies in each direction are sampled between 0 and the sampling frequency on that direction. When performing a DFT operation, one must have in mind that it is implicitly assumed that the profile or surface are periodic in each direction, with period equal to the number of sampled points in each direction. The discrete transform is also periodic in each direction, also with same period. Furthermore, the conjugate symmetry property still holds for the discrete scenario, writing

$$Z_q = Z^*_{-q} = Z^*_{N-q}\,; \tag{2.67a}$$

$$Z_{p,q} = Z^*_{-p,-q} = Z^*_{M-p,N-q}\,. \tag{2.67b}$$

The main difference relative to the continuous case, is that PSD is not the DFT of the discrete ACF $\hat{R}$ from Equation (2.61), but of the circular ACF, denoted by $\widetilde{R}$. The definition of circular ACF is similar to Equation (2.61), yet it assumes periodicity of the profile/surface, which allows the average to be performed over all domain. One writes the discrete circular ACF as

$$\widetilde{R}_q = \frac{1}{N}\sum_{n=0}^{N-1} z_n z_{n+q}, \qquad q = 0,1,2,....,N-1\,, \tag{2.68a}$$

$$\widetilde{R}_{p,q} = \frac{1}{MN}\sum_{m=0}^{M-1}\sum_{n=0}^{N-1} z_{m,n} z_{m+p,n+q}, \quad \begin{cases} m = 0,1,...,M-1 \\ n = 0,1,...,N-1 \end{cases}, \tag{2.68b}$$

where it is assumed that

$$z_n = z_{n+N}\,; \tag{2.69a}$$

$$z_{m,n} = z_{m+M,n} = z_{m,n+N} = z_{m+M,n+N}\,. \tag{2.69b}$$

Moreover, it is reasonable to name DFT of circular ACF as an estimate of the power spectral density, rather than PSD itself, since it is an estimation of a continuous PSD from discrete data. Henceforth, one can write the relation between the PSD estimate and circular autocorrelation

$$\text{DFT}\left(\widetilde{R}_n\right) = \hat{\Phi}^\theta_q = \sum_{n=0}^{N-1} \widetilde{R}_n e^{-\text{i}2\pi qn/N}, \qquad q = 0,1,2,....,N-1\,; \tag{2.70a}$$

$$\text{DFT}\left(\widetilde{R}_{m,n}\right) = \hat{\Phi}_{p,q} = \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} \widetilde{R}_{m,n} e^{-\text{i}2\pi(qm/M+pn/N)}, \quad \begin{cases} p = 0,1,...,M-1 \\ q = 0,1,...,N-1 \end{cases}. \tag{2.70b}$$

The inverse relation holds

$$\mathrm{IDFT}\left(\hat{\Phi}^{\theta}_{q}\right) = \widetilde{R}_n = \frac{1}{N}\sum_{q=0}^{N-1}\hat{\Phi}^{\theta}_{q} e^{\mathrm{i}2\pi qn/N}, \qquad n = 0,1,2,....,N-1\,; \tag{2.71a}$$

$$\mathrm{IDFT}\left(\hat{\Phi}_{p,q}\right) = \widetilde{R}_{m,n} = \frac{1}{MN}\sum_{m=0}^{M-1}\sum_{n=0}^{N-1}\hat{\Phi}_{p,q} e^{-\mathrm{i}2\pi(qm/M+pn/N)}, \quad \begin{cases} m = 0,1,...,M-1 \\ n = 0,1,...,N-1. \end{cases} \tag{2.71b}$$

From the autocorrelation theorem, the estimate of PSD can be related with the discrete spectrum of the sampled surface, which now comes

$$\hat{\Phi}^{\theta}_{q} = \frac{|Z_q|^2}{N}\,; \tag{2.72a}$$

$$\hat{\Phi}_{p,q} = \frac{|Z_{p,q}|^2}{MN}\,. \tag{2.72b}$$

Substituting the previous relations in the discrete surface's IDFT, one reaches an expression describing the discrete surface synthesis via the estimate of its PSD

$$z_n = \frac{1}{N}\sum_{q=0}^{N-1}\sqrt{N\hat{\Phi}^{\theta}_{q}}\; e^{\mathrm{i}(2\pi qn/N+\phi_q)}, \qquad n = 0,1,2,....,N-1\,; \tag{2.73a}$$

$$z_{m,n} = \frac{1}{MN}\sum_{p=0}^{M-1}\sum_{q=0}^{N-1}\sqrt{MN\hat{\Phi}_{p,q}}\; e^{\mathrm{i}(2\pi(qm/M+pn/N)+\phi_{p,q})}, \quad \begin{cases} m = 0,1,...,M-1 \\ n = 0,1,...,N-1 \end{cases} . \tag{2.73b}$$

Note that in Expressions (2.73) random phases ϕ_q and $\phi_{p,q}$ were introduced, since Expressions (2.72) only relate the magnitude of DFT and not to the phases.

A key point to mention here is the relation between the discrete estimate of the surface spectrum $\hat{\Phi}$ and the continuous Φ. While complex phenomena may happen in the transformation from the continuous into the discrete, such as frequency aliasing, there is no interest in exploring that topic here. Thus, one shall focus on a completely clean transformation of the continuous spectrum into the the discrete. Referring to Equation (A.22), on page 212, it can be seen that the discrete spectrum is obtained by sampling the continuous one, followed by a division with the sampling length l_s. Hence, for an aliasing free sampling, it writes

$$\hat{\Phi}^{\theta}[q] = \frac{1}{l_s}\Phi_{\theta}\left(\frac{q}{N}\Omega_s\right), \qquad q = 0,1,2,....,N/2\,, \tag{2.74}$$

$$\hat{\Phi}[p,q] = \frac{1}{l_{s_x} l_{s_y}}\Phi\left(\frac{q}{N}\Omega_{s_x}, \frac{p}{M}\Omega_{s_y}\right), \quad \begin{cases} p = 0,1,...,M/2 \\ q = 0,1,...,N/2 \end{cases} , \tag{2.75}$$

for profiles and surfaces, respectively.

At last, the computation of spectral moments from discrete data is addressed. Several strategies have been proposed in the literature. For example, Longuet-Higgins (1957b) proposed the computation of these quantities based on the density of extrema and zero crossings in a profile. Other method consists in computing RMS parameters, and the spectral moments are deduced from Equation (2.35) to (2.36). Still, another alternative is based on the numerical computation of the derivative of ACF at the origin. The problem of such method is that it relies either on the topography of a particular discrete surface and on numerical computation of derivatives. This is undesirable, since spectral moments would not be uniquely determined from PSD and would be dependent, for example, on the method adopted for derivative computation. Additionally, RMS slope and curvature are largely affected by noise at small scales, which would introduce errors in spectral moments. Yastrebov, Anciaux, *et al.* (2017) suggests to compute spectral moments directly from surface of profile spectrum, in order to avoid such errors. The spectral moment for a discrete profile writes

$$m_{\theta n} = \frac{1}{N} \sum_{j=-N/2+1}^{N/2} \left(\frac{2\pi j}{L}\right)^n \hat{\Phi}^\theta_j \,. \tag{2.76}$$

The zeroth moment is exactly equal to profile RMS roughness squared, but relation to RMS slope and curvature holds approximately, rather than exactly:

$$R_q^2 = m_{\theta 0}\,; \tag{2.77a}$$

$$R_{\Delta q}^2 \approx m_{\theta 2}\,; \tag{2.77b}$$

$$R_{\Delta^2 q}^2 \approx m_{\theta 4}\,. \tag{2.77c}$$

Similar expressions can be written for surfaces. Starting by defining the spectral moment for a discrete surface, it comes

$$m_{pq} = \frac{1}{MN} \sum_{i=-M/2+1}^{M/2} \sum_{j=-N/2+1}^{N/2} \left(\frac{2\pi j}{L_x}\right)^p \left(\frac{2\pi i}{L_y}\right)^q \hat{\Phi}_{i,j}\,. \tag{2.78}$$

The relation between surface spectral moments and discrete surface RMS parameters come analogously

$$S_q^2 = m_{00}\,; \tag{2.79a}$$

$$S_{\Delta q}^2 \approx m_{02} + m_{20}\,; \tag{2.79b}$$

$$S_{\Delta^2 q}^2 \approx \frac{m_{04} + 2m_{22} + m_{04}}{4}\,. \tag{2.79c}$$

While the previous strategy for computation of spectral moments does not depend on the discretization techniques for the derivative, it stills depends on the number of points used to sample a given PSD, and thus, will produce different results for different levels of discretization. To circumvent this issue, the best way to compute spectral moments is through analytical solution for each case, whenever possible.

Remark 2.6 on the definition of the Discrete Fourier transform.
It is important to emphasize the definition of the Fourier transform adopted in the current work, as remarked previously regarding the continuous transform. The DFT of a one dimensional discrete function is here defined as

$$\mathrm{DFT}(f_n) = F_q = \sum_{n=0}^{N-1} f_n e^{-\mathrm{i}2\pi q n/N}, \qquad q = 0, 1, 2, \ldots., N-1\,,$$

and the inverse transform (IDFT) writes

$$\mathrm{IDFT}(F_q) = f_n = \frac{1}{N}\sum_{q=0}^{N-1} F_q e^{\mathrm{i}2\pi qn/N}, \qquad n = 0,1,2,....,N-1\,.$$

Analogously to the continuous transform case, these definitions vary from the work of some authors to another, by moving the division by N from the inverse to the forward transform—this influences how several equations are written. The above definitions were adopted in the current work in order to match the ones used in the `Python` scientific computing library `SciPy`. This package was used for the computer implementation of the random topography generator, thus the connection between implementation and documentation is simpler if the same definition is used for both.

Page intentionally left blank

Chapter 3

Numerical generation of randomly rough topography

Numerical simulations are a powerful tool to predict mechanical interaction between rough surfaces. Techniques such as the Finite Element Method (FEM) rely on a discrete mesh to approximate rough topography, thus requiring a discretization of surface and profile geometry. Discrete profiles can be readily measured in real surfaces with stylus devices, in a relatively cheap and fast fashion. However, discrete surface measurements are very time consuming. Numerical generation of random rough topography comes as a very attractive tool, that provides artificially generated discrete heights having statistical and spectral properties identical to real world surfaces. Not only it allows the generation of a large number of rough topography realizations in an extremely short period of time—when compared to that that would be needed to acquire an equivalent amount of experimental data—but it also promotes the realization of parametric studies involving surface statistics. For instance, if the influence of surface kurtosis on real contact area was to be studied from experimental data, one would need to measure different surfaces until finding one which verified the required statistics. This seems a rather inefficient approach to the problem, when a random surface generator would be able to reproduce those statistics consistently and a lot faster.

Several rough surface generation strategies have been proposed since around 1970. In this chapter, a brief literature review on generation algorithms is presented. Two particular algorithms, one for Gaussian (J.-J. Wu, 2000b) and other concerning non-Gaussian topography (J.-J. Wu, 2004), are described in detail. Both algorithms were implemented using the `Python` programming language. Validation results and numerical tests are also documented, in order to access algorithm performance. This chapter closes with an assessment of these numerical methods against experimental data on profile and surface roughness.

3.1 Brief literature review

Following the ideas presented in Chapter 2, the challenge of numerical generation of profiles and surfaces is to create topographies with prescribed statistics, such as skewness and kurtosis, and spectral content—this is, verifying some required ACF or PSD. Most generation methods fall into the category of Autoregressive Moving Average techniques (ARMA), which are concepts rooted in the field of time series modeling. Generally, ARMA methods describe the height at position n as a function of the previous height values z_i and an input white noise η_i. In other words, it models the response of a linear closed loop system to white noise. For example, the general expression for ARMA methods when applied to rough profiles writes

$$z_n = \sum_{i=0}^{m} a_i \eta_{i+n} + \sum_{j=0}^{n-1} b_j z_j. \tag{3.1}$$

Equation (3.1) expresses height z_n as a function of previous heights (*autoregressive*) and a weighted average of the input white noise η_i. The average of η_i is computed over a set which changes with changing n—hence the designation *moving average*. These models can be reduced to simpler ones such as Autoregressive models (AR) or Moving Average models (MA), when coefficients a_i or b_i are set to zero, respectively. Complete ARMA methods are not as popular as their simplified versions, when it comes to random topography generation. The work of Gu and Huang (1990) is cited as an example of application of such methods.

AR methods were the first to get attention, by Staufert, G. (1979) and DeVries, W. R. (1979). Both authors generated Gaussian rough profiles from an input ACF, yet adopted different strategies for the computation of coefficients b_j. Whitehouse (1983) extended the application of AR models to Gaussian surface generation. The major drawback from AR models is that they can only consider few terms of the sum in Equation (3.1), hence only a small number of points near the origin of the ACF can be accounted.

Concerning MA models, the method proposed by Patir (1978) is one of most popular strategies for random surface generation, and serves as starting point for many other methods. Even though it can be considered as a moving average method, Patir's method does not follow the idea behind typical time series modeling. Instead, the moving average procedure is thought as a linear transformation of a random matrix. By imposing that all input points are uncorrelated (the autocorrelation function is zero everywhere except at the origin), and also that the output surface ACF must verify a set of prescribed values, a system of nonlinear equations on the transformation matrix coefficients can be written. In the original work, Patir proposed to solve this nonlinear system with Newton-Raphson method. Overall, this generation method revealed convergence issues for large autocorrelation lengths. Furthermore, once it needs to solve a system of nonlinear equations, memory requisites and computation times grow rapidly with increasing surface size. Patir (1978) also suggested a strategy to generate non-Gaussian surfaces. However, it did not become as notorious, and currently is rarely employed. Some authors proposed alternative methods based on Patir's, in order to reduce computational costs and increase speed. Bakolas (2003) reformulated Patir's method, transforming a root finding problem (solution of the nonlinear system of equations) into an optimization problem, which could be solved using the Nonlinear Conjugate Gradient Method (NCGM). Additionally, Bakolas used a FFT-based strategy to compute the ACF, to reduce computation time even further. Since the solution of a optimization problem is not necessarily equal to the corresponding root finding problem, due to the existence of local extrema, Bakolas' method may prove unsatisfactory in some cases. Liao *et al.* (2018) redefined the problem as a nonlinear least squares problem, to improve the method's efficiency and stability.

The cornerstone of non-Gaussian topography generation was laid by Watson and Spedding (1982), whose analysis started from ARMA models, and then focused on pure MA models. The authors presented formulas relating the skewness and kurtosis of the output of a pure moving average procedure z_n with coefficients a_i and the skewness and kurtosis of the input white noise η_i. By knowing *a priori* the MA coefficients, one can compute the skewness and kurtosis needed for η_i, such that they will result in the prescribed values at the output z_n. For this purpose, a non-Gaussian set random numbers η_i needs to be generated, which can be accomplished with Johnson translator system of frequency curves (N. L. Johnson, 1949; Elderton and N. L. Johnson, 1969). Johnson frequency curves are non-Gaussian probability density functions, which result from a transformation of the Gaussian distribution. This transformation involves 4 parameters and can be performed with 3 different types of frequency curves. Using the fitting algorithm proposed by I. D. Hill, R. Hill, *et al.* (1976), the curve type can be selected, and the parameters can be fitted such that the probability distribution verifies the prescribed values of skewness and kurtosis for η_i. Thus, requesting a particular combination of skewness and kurtosis for a rough topography, the required skewness and kurtosis for a random number set η_i is computed using the results from Watson and Spedding (1982), and such sequence η_i can be generated by transforming a set of normal numbers using Hill's algorithm. In fact, this is the mainstream scheme for generating artificial non-Gaussian surfaces, and is also integrated in the methods proposed by Bakolas (2003) and Liao *et al.* (2018).

While all the works on topography generation cited so far used as input an ACF, the possibility to prescribe a particular power spectrum is also of particular interest. Topography generation from MA models can be tackled with Fourier transforms, in particular FFT and IFFT. These tools provide a convenient framework to handle the power spectrum of artificial topography. Newland (1984) presented a method to generate rough surface either from PSD or ACF. A similar strategy was applied by Ganti and Bhushan (1995) for rough profiles, and J.-J. Wu (2000b) extended the previous method for surface generation. These methods consist in synthesizing rough topography as the superposition of waves, whose amplitudes depend on the input PSD, and the phases are randomly generated. This can be reduced to imposing a particular DFT to a rough topography, which is then generated by performing an IDFT operation on the explicitly built DFT. Newland's method differs from Wu's in the scheme adopted to compute the amplitude of each wave from PSD or ACF. In the pioneer work of Hu and Tonder (1992), topography generation was initially approached by transforming a MA model into the frequency space. Hu and Tonder's concept lies in the similarity between surface height generation from MA models and the application of a digital filter to an input signal. With this analogy, it is possible to generate rough surfaces from a given ACF. Moreover, generation from PSD is also straightforward. This work also incorporated the previously mentioned non-Gaussian surface generation framework, based on Watson and Spedding work, Johnson translator system and Hill's fitting algorithm. Reizer (2011) compared both Wu and Newland's method, and concluded that Wu's represents rough topography with higher accuracy, although differences are generally small. J.-J. Wu (2000b) carried out a comparison between his, Newland's and Hu and Tonder's method, regarding Gaussian topography. All methods showed good results, as long the autocorrelation length is kept small. Wu claims that, even though all method shows similar result, his method performs better, and points out mathematical mistakes in the formulation of Hu and Tonder's method. Despite such eventual mistakes, Hu and Tonder's method achieved great success amongst the scientific community studying rough contact, and it is currently used by many researchers (Yastrebov, Anciaux, *et al.*, 2015; Urzică *et al.*, 2012). However, generation of non-Gaussian surfaces with this method has revealed unsatisfactory results for skewness and kurtosis of artificially generated surfaces. This may owe directly to the method itself, or to the limitations of Johnson translator system (Ao *et al.*, 2002). An iterative process was proposed by J.-J. Wu (2004) in order to improve the performance of Hu and Tonder's method, concerning verification of spectral properties, skewness and kurtosis. Francisco and Brunetière (2016) presented an hybrid analytical/numerical method, in order to cope with the limitations of typical non-Gaussian generators, and to extend the range of skewness and kurtosis of generated surfaces.

Different techniques have also been applied to surface generation, yet with far less expression. Some examples are random field theory (Temizer, 2011), Monte Carlo simulations (Zou *et al.*, 2007), neural network schemes (Patrikar, 2004), random midpoint displacement method (Zahouani *et al.*, 1998) and fractal simulation with Weierstrass-Mandelbrot function (Majumdar and Tien, 1990). Surface generation methods have also been used to produce stratified surfaces, which result from the superposition of two independent surfaces with different height distributions (Pawlus, 2008).

Remark 3.1 on the selection of the generation methods.
Both methods proposed by Wu (J.-J. Wu, 2000b, 2004) were selected as rough topography generators for the present work. This selection was based, firstly, on the convenience of having at our disposal two generations methods, one for Gaussian topography and other for non-Gaussian (even though non-Gaussian generators can produce Gaussian surfaces as a particular case). With this division, the complexity and limitations associated with non-Gaussian methods are separated from Gaussian topography generation. Secondly, both methods are based on FFT, which means that the enforcement of the rough surface power spectra is simplified. The two methods are well established in literature, and are used frequently in rough contact numerical studies. Note that the work of Francisco and Brunetière (2016) seems a reliable alternative to explore in future works.

3.2 Gaussian topography

In the following section, a detailed description of the random rough Gaussian topography generation algorithm proposed in J.-J. Wu (2000b) is shown. It applies to both rough profiles and surfaces, yet only surface generation will be described, for simplicity.

The synthesis of a random rough surface can be performed as the superposition of discrete waves with different frequency, amplitude and phase. Recalling Expressions (2.73), the discrete surface $z_{m,n}$ is the result of an IDFT operation

$$z_{m,n} = \frac{1}{MN} \sum_{p=0}^{M-1} \sum_{q=0}^{N-1} \sqrt{MN\hat{\Phi}_{p,q}}\; e^{\mathrm{i}\left(2\pi(qm/M+pn/N)+\phi_{p,q}\right)}, \quad \begin{cases} m = 0,1,\ldots,M-1 \\ n = 0,1,\ldots,N-1 \end{cases} . \tag{3.2}$$

Recall that M and N are the number of sampled points in y and x direction, respectively, $\hat{\Phi}_{p,q}$ is the estimate of the real power spectral density (related to the circular autocorrelation function) and $\phi_{p,q}$ is an array of random phases. Note that the sampled points are uniformly spaced in the interval $[0,L[$, where L is the sample length on the considered direction. By specifying the values for $\hat{\Phi}_{p,q}$ and the phases $\phi_{p,q}$, a discrete surface can be generated by Equation (3.2). For the specification of $\hat{\Phi}$, the two cases of interest are generation from input PSD and ACF. Starting with surface synthesis from PSD, it is convenient to repeat that $\hat{\Phi}$ is a discrete estimate of the real power spectrum $\Phi(k_x, k_y)$. If the real PSD is bandwidth limited (i.e., that above a certain wavenumber the power spectrum is null) and further considering that the Nyquist frequency is higher than this limiting wavenumber, frequency aliasing, caused by the topography discretization, can be neglected. In such conditions, it is reasonable recall Equation (2.75), which states

$$\hat{\Phi}_{p,q} = \frac{1}{l_{s_x} l_{s_y}} \Phi\left(k_x = \frac{q}{N}\Omega_{s_x},\ k_y = \frac{p}{M}\Omega_{s_y}\right), \quad \begin{cases} p = -M/2+1,\ldots,M/2 \\ q = 0,1,\ldots,N/2 \end{cases} . \tag{3.3}$$

Note that, from the properties of discrete Fourier analysis, $\hat{\Phi}$ is periodic in both directions, with period equal to M and N on y and x directions, respectively. Hence, only M and N need to be defined, in order to compute $\hat{\Phi}_{p,q}$ from a required continuous PSD. Sampling frequencies in both direction are here denoted by Ω_s. It proves convenient to rewrite Equation (3.3) with both indexes starting at 0, which comes, in a compact notation, and attending to the discrete surface periodicity

$$\hat{\Phi}_{p,q} = \frac{1}{l_{s_x} l_{s_y}} \Phi\left(k_x = \frac{q}{N}\Omega_{s_x},\ k_y = \left[(1-\chi)\frac{p}{M} + \chi\left(1-\frac{p}{M}\right)\right]\Omega_{s_y}\right), \quad \begin{cases} p = 0,\ldots,M-1 \\ q = 0,\ldots,N/2 \end{cases} , \tag{3.4}$$

with

$$\chi = \begin{cases} 0, & p \in \left[0, \dfrac{M}{2}\right] \\ 1, & p \in \left]\dfrac{M}{2}, M-1\right] \end{cases} . \tag{3.5}$$

In Equation (3.3), index p spans all M values, while q index spans only half of required points. This is due to the conjugate symmetry property, expressed by Equations (2.67), which states that only half of frequency space needs to be regarded, since the remaining points are obtained by

$$\hat{\Phi}_{p,q} = \hat{\Phi}_{M-p,N-q}, \quad \begin{cases} p = 0,\ldots,M-1 \\ q = 0,\ldots,N/2 \end{cases} . \tag{3.6}$$

In the previous discussion it was assumed that index p spans all indexes, while q covers only half the needed indexes. The inverse would also be true, in which case p would only span half the indexes and q would cover available points in that direction. The first option was adopted for no particular reason, and shall be kept throughout the text.

Moving to the generation from a prescribed linear autocorrelation function R, recall that it is the circular ACF $\widetilde{R}$ that relates to the estimate of surface power spectrum $\hat{\Phi}_{p,q}$ by Equations (C.13)

$$\mathrm{FFT}\left(\widetilde{R}_{m,n}\right) = \hat{\Phi}_{p,q}\,. \tag{3.7}$$

In Equation (3.7), the notation for Fast Fourier Transform (FFT) replaces DFT, so that implementation details are explicit in the algorithm. Circular ACF is symmetric relative to the origin, and periodic in both directions—also with period equal to M and N in y and x directions. Combining these two properties, it results in symmetry relative to lines $m = M/2$ and $n = N/2$. Nonetheless, one needs the relation between linear and circular ACF needed, such that the coefficients $\hat{\Phi}_{p,q}$ contain information of the required linear ACF. Instead, the circular ACF can be related to the estimate of linear ACF $\hat{R}$ by, namely (Newland, 1984)

$$\widetilde{R}_{m,n} = \frac{M-m}{M}\frac{N-n}{N}\hat{R}_{m,n} + \frac{m}{M}\frac{N-n}{N}\hat{R}_{M-m,n} + \frac{M-m}{M}\frac{n}{N}\hat{R}_{m,N-n} + \frac{m}{M}\frac{n}{N}\hat{R}_{M-m,N-n}. \tag{3.8}$$

Notice that the estimate of linear ACF is computed over the discrete surface, which is not generated yet. Hence it is difficult to explicitly apply Equation (3.8). Newland's method consists in replacing $\hat{R}_{m,n}$, in Equation (3.8), with the prescribed linear correlation $R_{m,n}$. In contrast, Wu suggests to build the circular ACF directly with $R_{m,n}$, avoiding Equation (3.8). This writes

$$\widetilde{R}_{m,n} = \widetilde{R}_{M-m,n} = \widetilde{R}_{m,N-n} = \widetilde{R}_{M-m,N-n} = R_{m,n}, \quad \begin{cases} p = 0,\ldots,M/2 \\ q = 0,\ldots,N/2 \end{cases}. \tag{3.9}$$

From this circular ACF, $\hat{\Phi}_{p,q}$ can be computed from Equation (3.7), and it verifies the conjugate symmetry property, necessarily. Having computed the amplitudes, the array of random phases remains the only unknown. While each wave verifies a deterministic amplitude, which only depends on the input PSD, or input ACF, the random nature of rough topography will be ensured in the phases $\phi_{p,q}$. These should be generated by a random number generator following a uniform distribution between 0 and 2π, in order to avoid some phases showing with higher probability than others. This would synchronize the respective frequencies and deteriorate the required random behavior. Following the conjugate symmetry property, the phases must ensure

$$\phi_{m,n} = -\phi_{M-m,N-n}, \quad \begin{cases} p = 0,\ldots,M-1 \\ q = 0,\ldots,N/2 \end{cases}. \tag{3.10}$$

Combining the conjugate symmetry property and the DFT periodicity, the conjugate symmetry property applies to points which lie on other period of the DFT, hence, it must be condense in a relation between points in the same period. Then, the follow relationships must also be verified

$$\phi_{0,0} = \phi_{0,N/2} = \phi_{M/2,0} = \phi_{M/2,N/2} = 0\,; \tag{3.11a}$$

$$\phi_{m,0} = -\phi_{M-m,0}, \quad m = 0,\ldots,M/2\,; \tag{3.11b}$$

$$\phi_{0,n} = -\phi_{0,N-n}, \quad n = 0,\ldots,N/2\,. \tag{3.11c}$$

By using the amplitudes computed form the prescribed PSD or ACF, and generating a uniformly distributed grid of phases $\phi_{p,q}$ verifying the conjugate symmetry property, alongside with the periodic behavior DFT, a discrete rough surface can be generated from Equation (3.2). To enhance computational efficiency of the generation algorithm, Equation (3.2) shall not be implemented explicitly, but using a Inverse Fast Fourier Transform algorithm as

$$z_{m,n} = \mathrm{IFFT}\left(\sqrt{MN\hat{\Phi}_{p,q}}\, e^{\mathrm{i}\phi_{p,q}}\right). \tag{3.12}$$

The Gaussian distribution of heights is guaranteed from the superposition, of independent random variables—the height of each wave—at every point. From the central limit theorem, it comes that the height distribution is Gaussian. The flowchart of this algorithm is presented in Figures 3.1 and 3.2 for surfaces and profiles, respectively.

Remark 3.2 on the distinction between generated length and period.
There are two different lengths that shall be recognized in the profile/surface generation methods. The first one is the period L over which the points are sampled. This is, L is the period of the hypothetically continuous topography. Focusing on the profile case, but with conclusions valid for surfaces as well, when N points are sampled over the period, the points lie between $x = 0$ and $x = L(N-1)/N$, inclusively. The generation algorithm only takes these points into account, since the points at $x = L$ is equal to $x = 0$ and does not required explicit generation. Hence, one can distinguish between the generated profile length $L(N-1)/N$ and the periodic profile length L.

Usually, it is more convenient to work with the periodic profile length, since by determining the wavelength associated with each discrete frequency,

$$\frac{q}{N}\Omega_s = \frac{q}{N}\frac{2\pi}{l_s} = \frac{q}{N}\frac{2\pi N}{L} = \frac{2\pi}{L/q},$$

each discrete frequency can be associated with the wavelength L/q. This goes up to the Nyquist frequency, where the wavelength is $2L/N = 2l_s$, which is the maximum frequency that the discretization can represent. Furthermore, note operations such as discrete PSD computation and ACF estimate are computed over the generated length.

Remark 3.3 on the input of the power spectrum.
It was emphasized earlier that the discrete spectrum is a sampled and scaled version of the continuous spectrum, see Equation (3.3). *However, the roughness models discussed in Chapter 2 were formulated for continuous topography. Thus, one needs to scale the models for different levels of discretization of the topography. Adding this inconvenience to the fact that the generated profile is often normalized in a post processing step, one shall adopt the inverse strategy. The models in Equation* (3.3) *shall be used for the discrete spectrum, and whenever the continuous spectrum is required, one multiplies the model by the respective sampling lengths. For instance, any reference to parameters involved in fractal modes, such as the profile fractal scale factor G, refer directly to the discrete spectrum.*

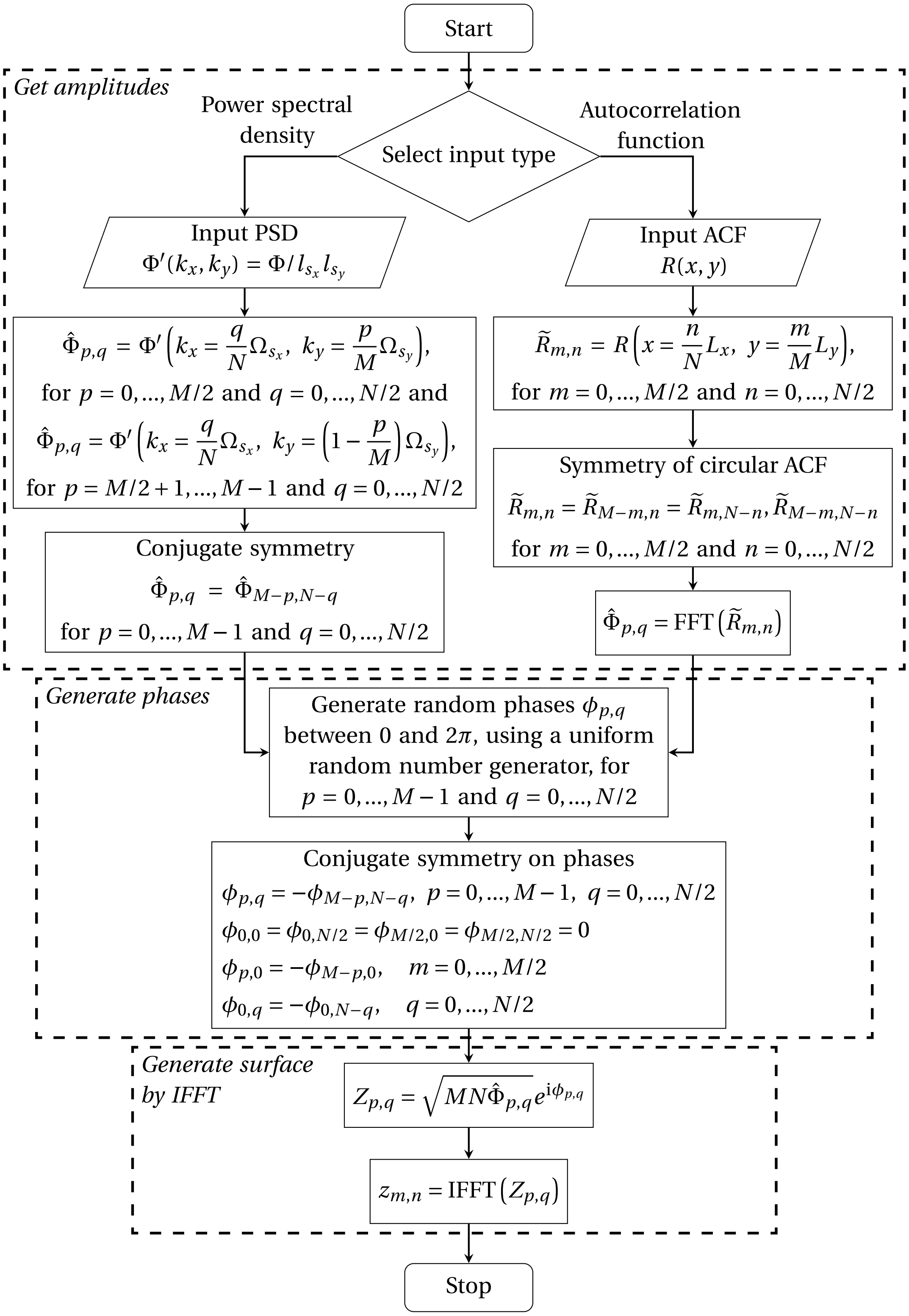

Start
Get amplitudes
Power spectral density
Select input type
Autocorrelation function
Input PSD
$\Phi'(k_x, k_y) = \Phi / l_{s_x} l_{s_y}$
Input ACF
$R(x, y)$
$\hat{\Phi}_{p,q} = \Phi'\left(k_x = \frac{q}{N}\Omega_{s_x},\ k_y = \frac{p}{M}\Omega_{s_y}\right),$
for $p = 0, ..., M/2$ and $q = 0, ..., N/2$ and
$\hat{\Phi}_{p,q} = \Phi'\left(k_x = \frac{q}{N}\Omega_{s_x},\ k_y = \left(1 - \frac{p}{M}\right)\Omega_{s_y}\right),$
for $p = M/2+1, ..., M-1$ and $q = 0, ..., N/2$
$\widetilde{R}_{m,n} = R\left(x = \frac{n}{N}L_x,\ y = \frac{m}{M}L_y\right),$
for $m = 0, ..., M/2$ and $n = 0, ..., N/2$
Symmetry of circular ACF
$\widetilde{R}_{m,n} = \widetilde{R}_{M-m,n} = \widetilde{R}_{m,N-n}, \widetilde{R}_{M-m,N-n}$
for $m = 0, ..., M/2$ and $n = 0, ..., N/2$
Conjugate symmetry
$\hat{\Phi}_{p,q} = \hat{\Phi}_{M-p,N-q}$
for $p = 0, ..., M-1$ and $q = 0, ..., N/2$
$\hat{\Phi}_{p,q} = \mathrm{FFT}\left(\widetilde{R}_{m,n}\right)$
Generate phases
Generate random phases $\phi_{p,q}$ between 0 and 2π, using a uniform random number generator, for $p = 0, ..., M-1$ and $q = 0, ..., N/2$
Conjugate symmetry on phases
$\phi_{p,q} = -\phi_{M-p,N-q},\ p = 0, ..., M-1,\ q = 0, ..., N/2$
$\phi_{0,0} = \phi_{0,N/2} = \phi_{M/2,0} = \phi_{M/2,N/2} = 0$
$\phi_{p,0} = -\phi_{M-p,0}, \quad m = 0, ..., M/2$
$\phi_{0,q} = -\phi_{0,N-q}, \quad q = 0, ..., N/2$
Generate surface by IFFT
$Z_{p,q} = \sqrt{MN\hat{\Phi}_{p,q}}e^{\mathrm{i}\phi_{p,q}}$
$z_{m,n} = \mathrm{IFFT}\left(Z_{p,q}\right)$
Stop

Figure 3.1: Flowchart of Gaussian random rough surface generation algorithm.

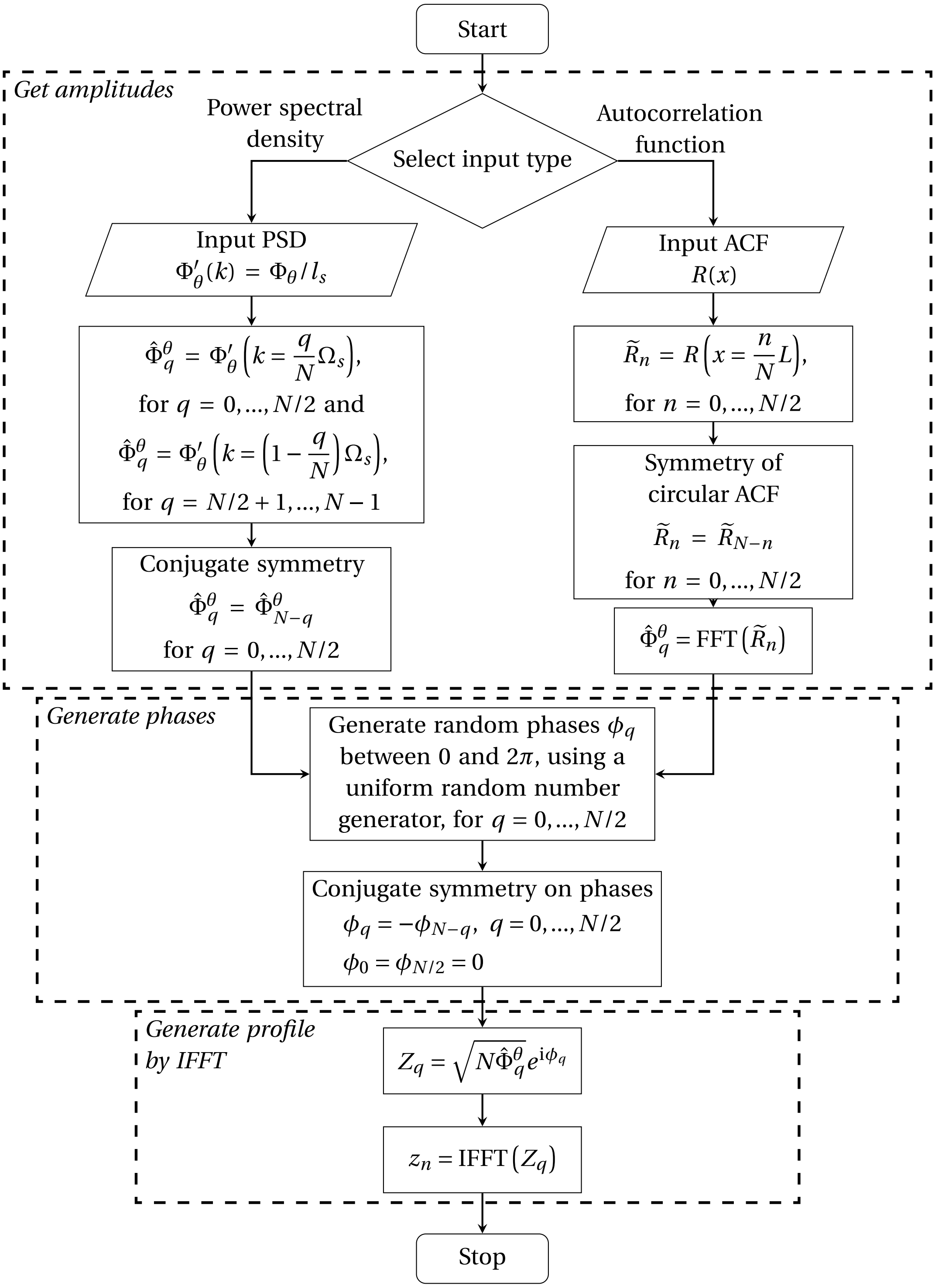
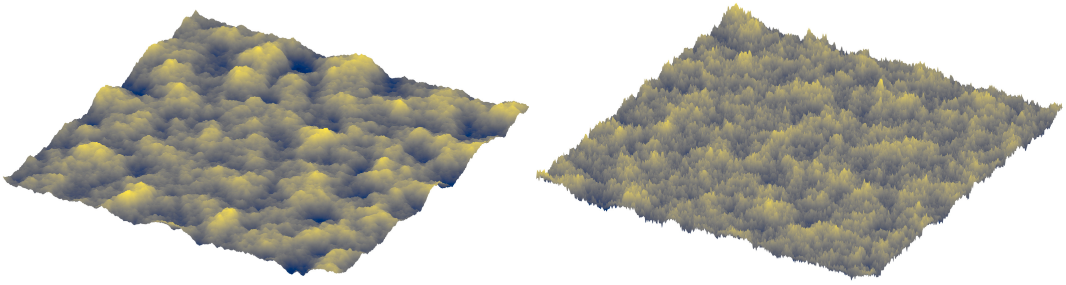

Start
Get amplitudes
Power spectral density
Select input type
Autocorrelation function
Input PSD
$\Phi'_\theta(k) = \Phi_\theta / l_s$
Input ACF
$R(x)$
$\hat{\Phi}^\theta_q = \Phi'_\theta\left(k = \frac{q}{N}\Omega_s\right)$,
for $q = 0, \ldots, N/2$ and
$\hat{\Phi}^\theta_q = \Phi'_\theta\left(k = \left(1 - \frac{q}{N}\right)\Omega_s\right)$,
for $q = N/2 + 1, \ldots, N - 1$
$\widetilde{R}_n = R\left(x = \frac{n}{N}L\right)$,
for $n = 0, \ldots, N/2$
Symmetry of circular ACF
$\widetilde{R}_n = \widetilde{R}_{N-n}$
for $n = 0, \ldots, N/2$
Conjugate symmetry
$\hat{\Phi}^\theta_q = \hat{\Phi}^\theta_{N-q}$
for $q = 0, \ldots, N/2$
$\hat{\Phi}^\theta_q = \mathrm{FFT}\left(\widetilde{R}_n\right)$
Generate phases
Generate random phases ϕ_q between 0 and 2π, using a uniform random number generator, for $q = 0, \ldots, N/2$
Conjugate symmetry on phases
$\phi_q = -\phi_{N-q}, \ q = 0, \ldots, N/2$
$\phi_0 = \phi_{N/2} = 0$
Generate profile by IFFT
$Z_q = \sqrt{N\hat{\Phi}^\theta_q}\, e^{\mathrm{i}\phi_q}$
$z_n = \mathrm{IFFT}\left(Z_q\right)$
Stop

Figure 3.2: Flowchart of Gaussian random rough profile generation algorithm.

3.2.1 Numerical tests

In order to evaluate the quality of implementation of the Gaussian topography generation algorithm, and outline its performance under several circumstances, different numerical experiments were carried out with the method. In particular, both profiles and surfaces were synthesized by prescribing ACF and PSD, and these functions were recomputed from the generated topography and compared with the input ones. Profile and surface generation follow a very similar structure, hence only results concerning one of the topography types will be shown for each case, which shall be emphasized at the time—conclusions are identical for both cases. Although physical units are not relevant for numerical tests, they will be displayed in graphical representation of results, such that some aspects can be clarified. As a matter of fact, almost every profile and surface used so far, for illustration of theoretical aspects, were generated by the implemented generation algorithms. Figure 3.3 shows two more examples of artificially generated Gaussian isotropic surfaces, verifying a fractal PSD.



(a) $H = 0.8$ **(b)** $H = 0.2$

Figure 3.3: Examples of artificially generated isotropic Gaussian surfaces. Both surfaces are generated from 1024 points in each direction with $L_x = L_y = 1\,\text{mm}$, and their heights are normalized, such that $S_q = 0.01\,\text{mm}$. The high frequency cut-off is set to $\lambda_s = L/256$ and the low cut-off to $\lambda = L/4$. There is no roll-off in both surfaces, i.e., $\lambda_r = \lambda_l$.

Starting with profile generation from input exponential ACF, Figure 3.4 shows the ACF computed over 5 artificially generated profiles, with $R_q = 1\,\text{mm}$, $L = 1\,\text{mm}$ and 1024 points, alongside with the theoretical ACF. For every case, the computed ACF recovers its value at the origin with practically zero error. For longer autocorrelation lengths however, the ACF computed from the artificial profile shows increasingly higher deviations relative to the exact value at each points, and these deviations also increase, on average, with the distance to the origin. For autocorrelation lengths shorter than $L/50$, the input ACF is recovered almost exactly. This results show good agreement with the ones presented by J.-J. Wu (2000b). A disadvantage of topography generation from exponential ACF, which was already referred earlier in this text, is that artificial profiles and surfaces have a non-zero mean height, which contradicts roughness definition. It is convenient to investigate how the output ACF will behave, if the mean value of height is removed from the profile. Figure 3.5 illustrates the influence of the mean value on ACF, for two different autocorrelation lengths. Both profiles are generated from 1024 points, from input data $R_q = 1\,\text{mm}$ and $L = 1\,\text{mm}$. It can be seen that the mean value has little influence for low autocorrelation lengths, while for longer ones it distorts the output ACF.

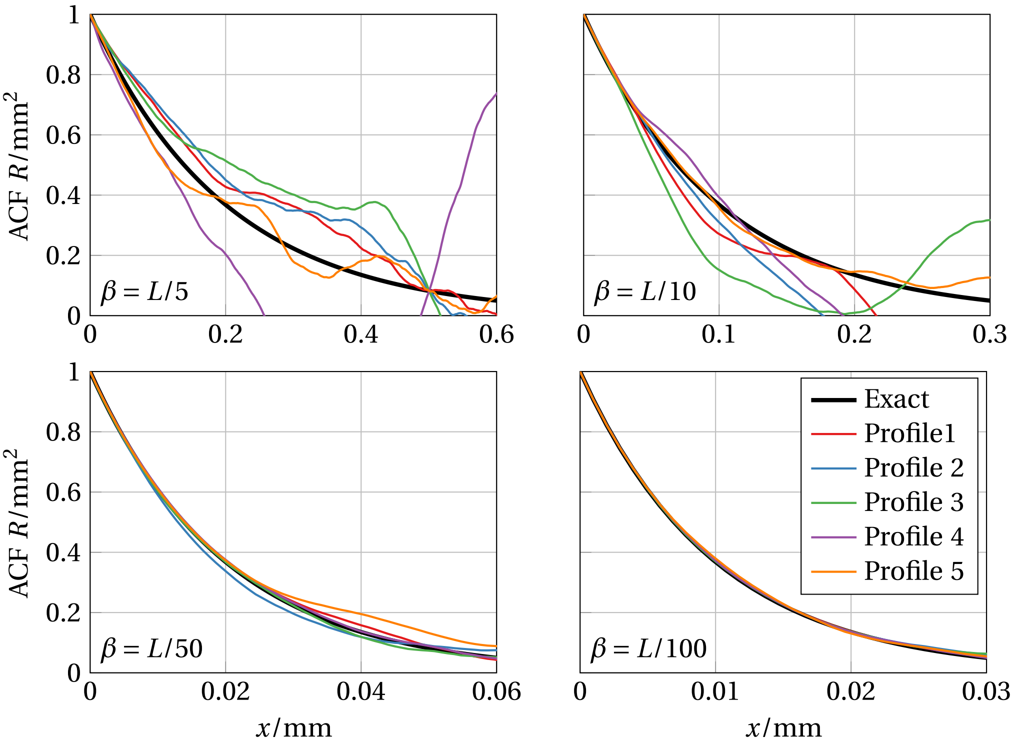


Figure 3.4: Effect of correlation length on the accuracy of the Gaussian profile generator. ACF computed from 5 profiles generated with $L = 1\,\text{mm}$, $R_q = 1$ and 1024 points are plotted. With increasing autocorrelation length, the generated profile show increasingly higher deviations relative to the input ACF, and the mean deviation increases with increasing distance to the origin. For $\beta < L/50$ the algorithm can recover the input ACF with high accuracy, at every point.

A feature of generation algorithms based on FFT, is that they imply periodicity of the generated topography. The discrete Fourier transform assumes that the discrete profile z_n or surface $z_{m,n}$ are periodic in every direction over which the transform is computed. This is emphasized in Figure 3.6, regarding a rough profile. From this figure, one sees that points $x = 0$ and $x = L$ are correlated, once it is very likely that profile height is similar at both points. Actually, some authors, who used FFT-based algorithms for surface generation, actually mention that they considered periodic surfaces or profiles, in their analysis.

Coming to topography generation from PSD, since the amplitude of DFT is explicitly prescribed in the algorithm (cf. Figures 3.1 and 3.2), one expects that the input PSD is recovered exactly. This result is confirmed in Figure 3.7, where the PSD of two fractal profiles, generated artificially, are plotted. The artificial profiles in question have 1024 points,

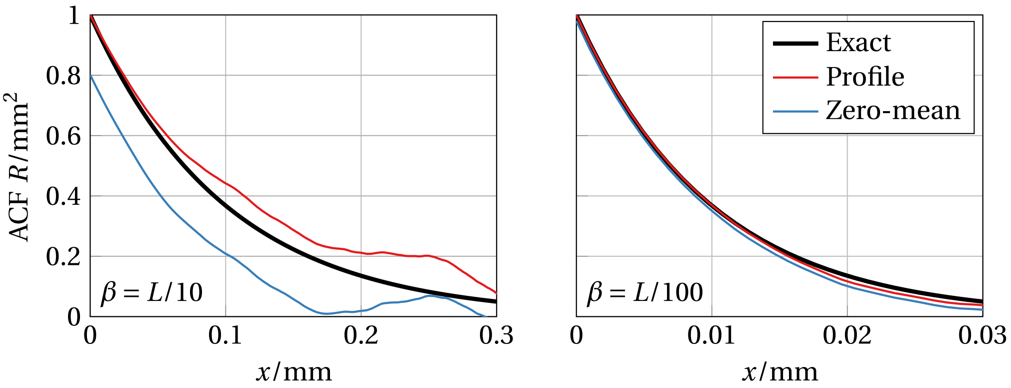


Figure 3.5: Influence of the profile mean value in ACF. Two profiles are generated with $L = 1$, $R_q = 1$ and 1024 points. Profiles artificially generated from an exponential ACF do not verify the zero mean height condition implied in roughness definition. If the mean value is removed from the profile and the ACF of this new profile is computed, the ACF moves away from the theoretical curve. The error due to mean value removal increases with increasing autocorrelation length.

and the algorithm inputs are $L = 1\,\text{mm}$, $G = 1$, $H = 0.5$, $\lambda_l = L$, and $\lambda_s = L/256$, where λ denotes wavelength.[1] One of the profiles is generated without roll-off ($\lambda_l = \lambda_r$) and other with roll-off $\lambda_r = L/10$. Following the previous comment on artificial topography periodicity, an approach to remove correlation between extreme points in a profile consists in trimming the profile, i.e., considering just a part of it. By doing so, only one extreme of the profile is kept, and the resemblance with the other extreme is discarded. However, it should be expected that the input quantities, such as the PSD, will no longer be verified with same accuracy. Together with the validation of generation from input PSD,

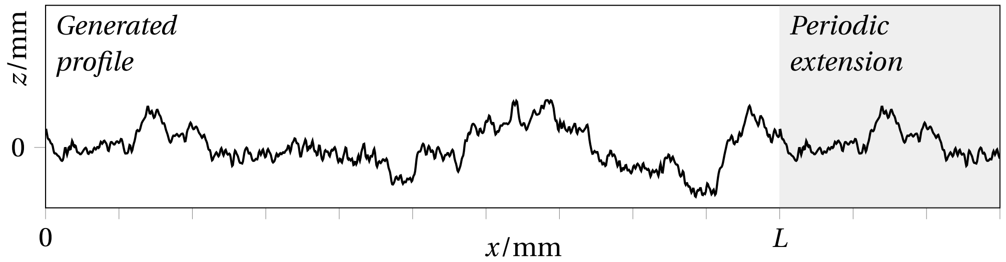


Figure 3.6: Illustrating periodicity of artificially generated profiles. IDFT assumes periodicity of the profile, with period equal to the sampled length. Thus, profiles generated from FFT-based algorithms are always periodic, which suggests that points $x = 0$ and $x = L$ are correlated.

[1] G is not dimensionless, yet dimensional compatibility between this parameter, k and Φ_θ, results in strange powers of length, which may lead to difficult physical interpretation. It was opted to ignore the dimensions of such quantity, for simplicity. The same reasoning applies to physical dimensions of surface fractal scale factor g and profile-surface parameter J.

Figure 3.7 shows the PSD of the profile with roll-off after trimming. The trimming procedure reduces the sampling length, but leaves sampling frequency unchanged. Hence, the frequency range that is covered by PSD is the same, yet the number of discrete frequencies frequency is smaller—frequency resolution is reduced. The power spectrum relative to the trimmed profile is different from the untrimmed, but its points are positioned around the prescribed function.

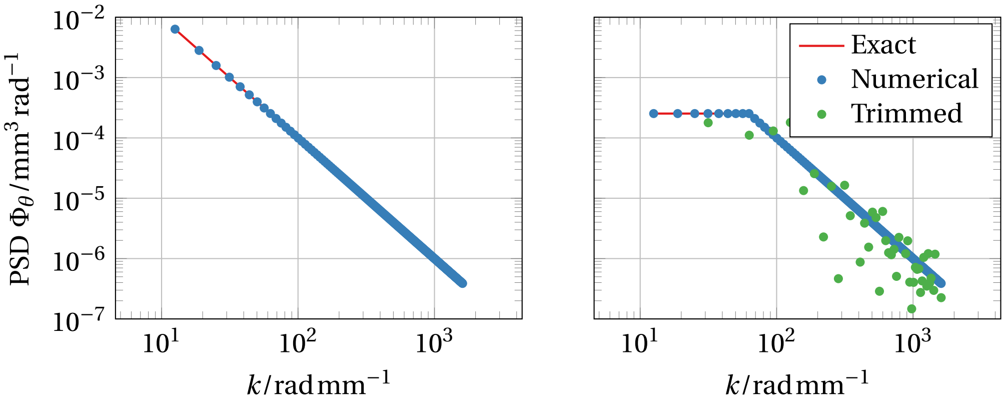


Figure 3.7: Power spectral density of artificially generated profiles. Profiles are generated from a fractal PSD with $L = L_x = L_y = 1\,\text{mm}$, $G = 1$, $H = 0.5$, $\lambda_l = L$, $\lambda_s = L/256$ and 1024 points. One profile is generated without roll-off, i.e., $\lambda_l = \lambda_r$, and other with roll-off $\lambda_r = L/10$. Since the method lies in explicitly setting the profile DFT, the input PSD is recovered exactly. If the profile is trimmed, in order to remove the correlation between extreme points, due to periodicity, the output PSD is distorted relative to the exact values, yet it follows a similar trend.

A last aspect which must be raised is the verification of whether the Gaussian topography generator does, indeed, synthesize Gaussian sets. This point was addressed by Yastrebov, Anciaux, *et al.* (2015), regarding Hu and Tonder's method, and the authors observed that for longer low wavelength cut-offs λ_l, artificial surfaces were non-Gaussian, even though the mean distribution was Gaussian. This distinction is paramount, since, citing the previous authors, '*the averaged mechanical response of non-Gaussian surfaces is not equivalent to the response of the averaged surface, whose distribution is Gaussian*'. Based on these observations, similar tests to those carried by Yastrebov, Anciaux, *et al.* (2015) for the method from Hu and Tonder, were explored. The test consisted in generating 100 rough surfaces with 1024×1024 points, from an input fractal PSD with $g = 1$ and $H = 0.8$. For each realization, the height distribution is computed, dividing the the z domain in 1000 bins (subintervals) between the maximum and minimum height value generated. From all probability distributions relative to each surface, the averaged height distribution is then computed, together with the standard deviation of the probability density in each bin. This process was repeated for several values of the low frequency cut-off k_l and the high frequency cut-off k_s. A reference Gaussian curve is plotted, with mean and standard deviation computed over all generated heights. The results for each combination of low and high cut-offs are presented in Figure 3.8.

Looking at the results, they are similar to the ones published by Yastrebov, Anciaux, *et al.* (2015). The main conclusion is that the effect of high cut-off frequency on height distribution is practically unnoticeable. In contrast, long low cut-off wavelengths distort the height distribution, which is a consequence of the discrete synthesis of rough surfaces. For $\lambda_l = L$ the averaged distribution is very near the Gaussian reference, yet the standard deviation is high, and individual realizations are considerably non-Gaussian. By reducing the low cut-off wavelength, the standard deviation of height distribution decreases, and individual realizations height distribution are closer to the Gaussian curve. For $\lambda_l = L/16$, the averaged distribution, the reference Gaussian and each individual distribution are almost coincident, and standard deviation vanishes, in each bin. Figure 3.9 repeats Figure 3.8 using logarithmic scale in the probability density axis, in order to highlight the tails of the distribution. These curves support previous observations that the height distribution is essentially Gaussian, even in the extremes. It should be mentioned that the previous results on topography height are similar for profiles, and even for topography generated from ACF. In this case, longer autocorrelation lengths have a similar effect to longer low cut-off wavelengths, i.e., longer ACL leads to artificial topography whose height distribution differs from a Gaussian curve.

As a final comment on Gaussian topography generation, one shall refer to computation time. In a machine equipped with a quad-core Intel® Core™ i7-7700HQ CPU at 2.8 GHz, the algorithms takes in average less than 1 second to compute a 1024×1024 surface, from an input PSD, without any parallelization strategy. For this case, which is already a rather extreme one, computation time is extremely small, hence for simpler cases cases, i.e., surfaces and profiles with smaller number of points, computation time is even smaller.

3.3 Non-Gaussian topography

In this section, the non-Gaussian rough topography generation method proposed by J.-J. Wu (2004) is described . Actually, the implemented algorithm is slightly different from the one presented by Wu, which will be clarified later. Following a similar structure to the presentation of the Gaussian generator, the algorithm will only be described for surface generation, yet algorithm flowcharts will be given for both scenarios. Despite the fact that the non-Gaussian algorithm departs from different grounds, when compared to the Gaussian generator, they share several aspects, such as the generation of phases verifying conjugate symmetry, and amplitude computation from input PSD or ACF. These routines are highlighted in Figures 3.1 and 3.2, such that the following algorithms can be simplified, by referring to the previous ones.

Non-Gaussian topography generation is based on the analogy between surface generation and digital filtering. Rough surfaces $z_{m,n}$ can be thought as the output of a digital filtering operation applied to an input white noise $\eta_{m,n}$.[2] Denoting filter coefficients by $h_{p,q}$, the resulting rough surface is written as the convolution of the filter coefficients

[2] For a more detailed discussion on digital filters, and in particular, digital filters in two dimension, the interested reader is referred to Lu (1992).

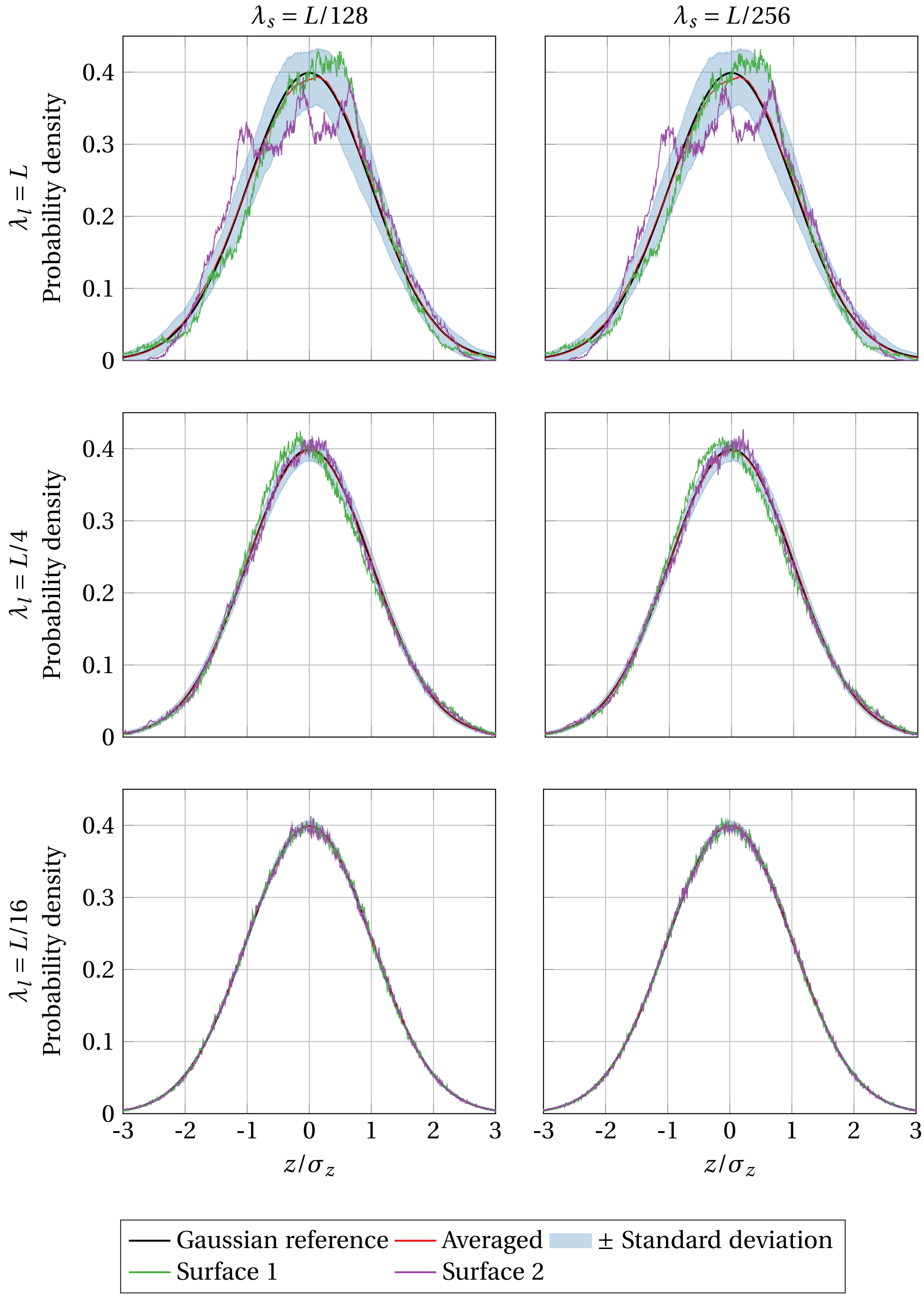


Figure 3.8: Probability density of artificial surfaces for several low and high frequency cut-offs. 100 surfaces with 1024×1024 points were generated from a fractal PSD, with $H = 0.8$, $g = 1$. For each realization, the height distribution was computed, using 1000 subintervals. Each plot contains the averaged height distribution, the standard deviation of probability density at each bin and the distribution of two particular realizations.

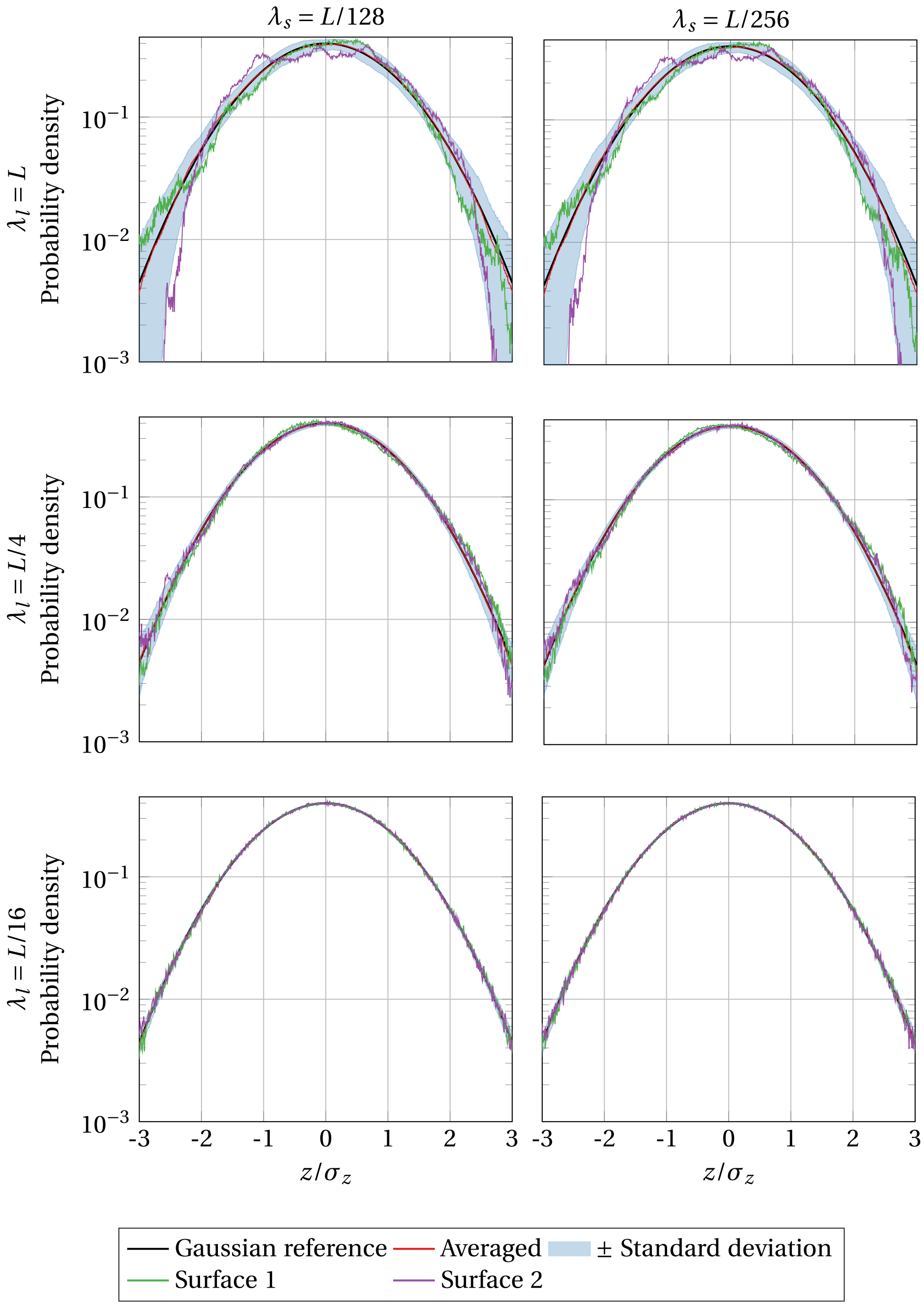


Figure 3.9: Probability density of artificial surfaces for several low and high frequency cut-offs: semi-logarithmic scale. It is a repetition of Figure 3.8 using a semi-logarithmic scale on the probability density axis, to highlight differences in the distribution’s tails.

with the *periodic* input white noise $\eta_{m,n}$, i.e.

$$z_{m,n} = \sum_{p=0}^{M-1}\sum_{q=0}^{N-1} h_{p,q}\eta_{m-p,n-q}, \quad \begin{cases} m = 0,\dots,M-1 \\ n = 0,\dots,N-1 \end{cases} . \tag{3.13}$$

Again, note that both the input signal $\eta_{m,n}$ and the output rough surface $z_{m,n}$ are periodic, with period M and N in y and x directions, respectively. From the convolution theorem of discrete Fourier transforms, the rough surface DFT can be written as the frequency-wise product of the DFT of $h_{m,n}$ and $\eta_{m,n}$,

$$Z_{p,q} = H_{p,q}A_{p,q}, \quad \begin{cases} p = 0,\dots,M-1 \\ q = 0,\dots,N-1 \end{cases} , \tag{3.14}$$

where $H_{p,q}$ denotes the DFT of the filter coefficients, also called the transfer function of the system, and $A_{p,q}$ the DFT of $\eta_{m,n}$. Multiplying each side of Equation (3.14) by the complex conjugate of $Z_{p,q}$, and dividing by the total number of points, Equation (3.14) can be rewritten in terms of the power spectrum of both signals

$$\hat{\Phi}^{(z)}_{p,q} = |H_{p,q}|^2\hat{\Phi}^{(\eta)}_{p,q}, \quad \begin{cases} p = 0,\dots,M-1 \\ q = 0,\dots,N-1 \end{cases} . \tag{3.15}$$

If the input signal $\eta_{m,n}$ is pure white noise, i.e., if its PSD is one for every frequency, the transfer function follows directly

$$H_{p,q} = \sqrt{\hat{\Phi}^{(z)}_{p,q}} . \tag{3.16}$$

Filter coefficients can be computed using an IFFT algorithm, and setting the phases of the transfer function arbitrarily to zero

$$h_{m,n} = \mathrm{IFFT}\left(H_{p,q}\right). \tag{3.17}$$

Observe that the hypothesis of $H_{p,q}$ being equal to the square root of the discrete rough surface power spectrum is coherent with the assumption that the phases of the filter transfer function are zero. A random signal $\eta_{m,n}$ which verifies a unit PSD for all frequencies can easily be generated recalling Equation (3.2)

$$\eta_{m,n} = \frac{1}{MN}\sum_{p=0}^{M-1}\sum_{q=0}^{N-1}\sqrt{MN}\, e^{\mathrm{i}\left(2\pi(qm/M+pn/N)+\phi_{p,q}\right)}, \quad \begin{cases} m = 0,1,\dots,M-1 \\ n = 0,1,\dots,N-1 \end{cases} . \tag{3.18}$$

Phases $\phi_{p,q}$ must verify the conjugate symmetry relations, expressed by Equations (3.10) and (3.11). For implementation purposes, it is more efficient to write $\eta_{m,n}$ as

$$\eta_{m,n} = \mathrm{IFFT}\left(\sqrt{MN}e^{\mathrm{i}\phi_{p,q}}\right). \tag{3.19}$$

Furthermore, note that once the phases of $H_{p,q}$ are zero, the phases of $Z_{p,q}$ will necessarily be equal to the phases of $A_{p,q}$, i.e., they are equal to $\phi_{p,q}$—this result comes from

Equation (3.14). The rough surface can then be synthesized from Equation (3.12), which rewrites here as

$$z_{m,n} = \mathrm{IFFT}\left(\sqrt{MN\hat{\Phi}^{(z)}_{p,q}}\, e^{\mathrm{i}\phi_{p,q}}\right). \tag{3.20}$$

In sum, a rough surface can be generated by IFFT, as the result of digital filtering of an input white noise (signal with unit PSD). The filter coefficients are related with $\hat{\Phi}^{(z)}_{p,q}$, which in turn are computed from the input PSD or ACF, following exactly the same expressions presented in Section 3.2. This procedure is reduced to the Gaussian generator algorithm, viewed from the perspective of digital filters.

The purpose of reformulating the Gaussian algorithm, as starting from a filtering operation given by Equation (3.13), relates to the work of Watson and Spedding (1982). These authors derived expressions relating the skewness and kurtosis of the input white noise $\eta_{m,n}$ of a pure MA process (digital filter), with the same statistics of the output signal $z_{m,n}$. This relation is expressed as a function of the filter coefficients. For two dimensional signals, it writes

$$S^{(z)}_{sk} = \frac{\sum_{r=0}^{MN-1} h_r^3}{\left(\sum_{r=0}^{MN-1} h_r^2\right)^{3/2}} S^{(\eta)}_{sk}, \tag{3.21a}$$

$$S^{(z)}_{ku} = \frac{S^{(\eta)}_{ku} \sum_{r=0}^{MN-1} h_r^4 + 6\sum_{r=0}^{MN-2}\sum_{p=r+1}^{MN-1} h_p^2 h_r^2}{\left(\sum_{r=0}^{MN-1} h_r^2\right)^2}, \tag{3.21b}$$

with $h_r = h_{m,n}$ and $r = mN + n$ for $m = 0, ..., M-1$ and $n = 0, ..., N-1$. This is, h_r is the one dimensional array which results from stacking the rows of $h_{m,n}$ in ascending order. Hence, from Equations (3.21) one could compute the particular values of skewness and kurtosis for $\eta_{m,n}$, which would produce the prescribed values of surface skewness and kurtosis in $z_{m,n}$. If one could generate a random set of numbers, verifying unit PSD for all frequencies, and whose skewness and kurtosis were $S^{(\eta)}_{sk}$ and $S^{(\eta)}_{ku}$, then, a rough surface with prescribed spectral and statistical properties can be synthesized from Equation (3.12) or Equation (3.13). This idea relies on the generation of a random set of non-Gaussian numbers, which can be accomplished with Johnson system of frequency curves. When $\eta_{m,n}$ is a Gaussian signal, from Equations (3.21) it follows that the output surface is also Gaussian. This case reduces the algorithm to a Gaussian generator, and supports the Gaussian algorithm described in Section 3.2. In fact, it is not necessary to resort to Equations (3.21) to reach such conclusion, since the result that the output of a linear system is Gaussian as long as the input is also Gaussian (Lu, 1992).

3.3.1 Johnson frequency curves

Johnson frequency curves are probability density functions of non-Gaussian variables, which result from transformations of Gaussian variables (N. L. Johnson, 1949; Elderton and N. L. Johnson, 1969). They are defined from three different types of curves, which transform a standardized normal variable ζ (zero mean and unit variance) into a Johnson variable η. The three types of curves are the lognormal system S_L, the unbounded system S_U and the bounded system S_B. For each system, the transformation of the Gaussian variable follows

$$S_L: \quad \eta = \gamma + \delta \ln(\zeta - \xi), \quad \xi < \zeta; \tag{3.22a}$$

$$S_U: \quad \eta = \gamma + \delta \sinh\left(\frac{\zeta - \xi}{\lambda}\right); \tag{3.22b}$$

$$S_B: \quad \eta = \gamma + \delta \ln\left(\frac{\zeta - \xi}{\xi + \lambda - \zeta}\right), \quad \xi < \zeta < \xi + \lambda. \tag{3.22c}$$

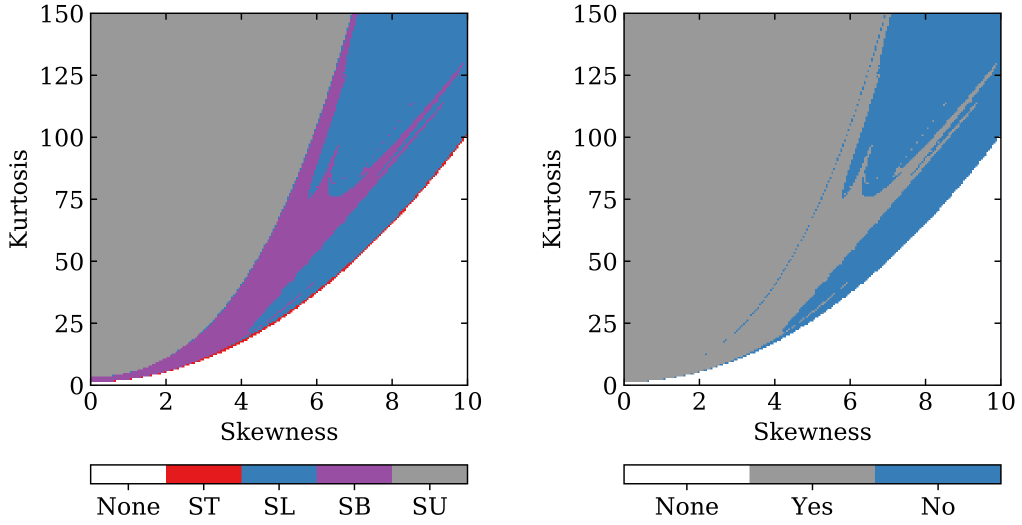
One can generate a non-Gaussian number from a Gaussian set of random numbers by using the transformations defined in Equations (3.22). However, the generation of non-Gaussian random numbers verifying a specific value of mean, standard deviation, skewness and kurtosis needs a judicious analysis, since for each combination of prescribed statistics, a particular transformation must be selected, and parameters γ, δ, ξ and λ need to be fitted carefully. I. D. Hill, R. Hill, *et al.* (1976) developed an algorithm which performs this fitting procedure—it selects the curve type, and computes the parameters γ, δ, ξ and λ, taking as input the four first moments of the PDF, which are required for the output. An implementation of the Gaussian-Johnson transformation, which corresponds to inverse relations of Equations (3.22) was also proposed by I. D. Hill (1976). Note that a remark was issued by I. D. Hill and Wheeler, E. (1981), concerning the definition of probability curves which was assumed in both algorithms, and providing corrections for alternative conventions. These algorithms were originally written in `FORTRAN 77`, and were rewritten in `Python` for the current work.

Johnson translator system, and in particular, Hill's algorithm, cannot span all skewness-kurtosis plane. In particular, Johnson curves must verify Equation (2.49), hence combinations of skewness and kurtosis which violate that inequality cannot be fitted. Furthermore, Hill's algorithm does not always converge to a good fit, and alternative solutions are suggested by the algorithm itself, even though the user is warned to check if that solution produces good results. This is illustrated in Figure 3.10, which shows the distribution type and convergence of the algorithm for several combinations of skewness and kurtosis. The fitting algorithm shows convergence issues near the boundary specified by Equation (2.49). The S_T distribution is a particular case of S_B curves, when points are very near the admissible boundary. Negative values of skewness are not presented, since symmetry exists relative to the kurtosis axis.

Even though Johnson system provides a framework for transforming Gaussian to non-Gaussian variables with prescribed statistics, it should be noted that these statistics are not guaranteed for all generated sets. In fact, it may be needed to generate several non-Gaussian signals, in order to get one with small error in the prescribed statistics.

3.3.2 Iterative procedure

It has been concluded in previous sections that, if one generates a set of non-Gaussian numbers $\eta_{m,n}$, verifying unit PSD and particular values of skewness and kurtosis given by Equation (3.21), then a rough surface can be synthesized from Equation (3.20). However, it is impossible, or at least, very difficult, to guarantee both spectral and statistical properties simultaneously with currently employed methods. Preceding any non-Gaussian



(a) Distribution type

(b) Algorithm convergence

Figure 3.10: Distribution type and convergence of Hill's fitting algorithm, as function of the input skewness and kurtosis. The algorithm shows convergence problems near the skewness-kurtosis boundary. Negative skewness are not presented, due to symmetry.

transformation, one needs to first get a Gaussian set, and Wu suggests that this initial Gaussian set shall verify unit PSD. The generation of this set follows Equation (3.19), with phases begin generated from a uniform distribution between 0 and 2π, just like in the algorithm regarding Gaussian topography synthesis. Then, one computes the values of skewness and kurtosis needed for $\eta_{m,n}$, in order to get the prescribed values of these moments in the rough profile

$$S_{sk}^{(\eta)} = \frac{\left(\sum_{r=0}^{MN-1} h_r^2\right)^{3/2}}{\sum_{r=0}^{MN-1} h_r^3} S_{sk}^{(z)} \, ; \tag{3.23a}$$

$$S_{ku}^{(\eta)} = \frac{S_{ku}^{(z)} \left(\sum_{r=0}^{MN-1} h_r^2\right)^2 - 6\sum_{r=0}^{MN-2} \sum_{p=r+1}^{MN-1} h_p^2 h_r^2}{\sum_{r=0}^{MN-1} h_r^4} \, ; \tag{3.23b}$$

with $h_r = h_{m,n}$ and $r = mN + n$ for $m = 0, \ldots, M-1$ and $n = 0, \ldots, N-1$. From the resulting moments $S_{sk}^{(\eta)}$ and $S_{ku}^{(\eta)}$, the Gaussian set $\eta_{m,n}$ is transformed into a non-Gaussian one, denoted by $\eta'_{m,n}$, with Hill's algorithm and Johnson system. The input of Hill's algorithm should be the mean and standard deviation of $\eta_{m,n}$, and PDF moments $S_{sk}^{(\eta)}$ and $S_{ku}^{(\eta)}$. The non-Gaussian set $\eta'_{m,n}$ will have statistics near the prescribed values $S_{sk}^{(\eta)}$ and $S_{ku}^{(\eta)}$, but it will no longer verify the condition of unit PSD. Since skewness and kurtosis are mostly phase sensitive parameters, a new array of phases $\phi'_{p,q}$ can be extracted from $\eta'_{m,n}$ as

$$\phi'_{p,q} = \angle \mathrm{FFT}\left(\eta'_{m,n}\right) . \tag{3.24}$$

A new set of non-Gaussian numbers with unit PSD can be synthesized, again from Equation (3.19), but using phases $\phi'_{p,q}$, instead:

$$\eta''_{m,n} = \mathrm{IFFT}\left(\sqrt{MN}e^{\mathrm{i}\phi'_{p,q}}\right) . \tag{3.25}$$

With this transformation, the condition of unit PSD is recovered, at dispense of accuracy on surface statistics. This is, skewness and kurtosis will change relative to $\eta'_{m,n}$, and, in general, the error relative to the required values will increase, when compared with $\eta'_{m,n}$ statistics. Sequence $\eta''_{m,n}$ does not even need to be created, since the result from convolution expressed in Equation (3.13) is know *a priori*. Thus, the rough surface can be readily generated by

$$z_{m,n} = \mathrm{IFFT}\left(\sqrt{MN\hat{\Phi}^{(z)}_{p,q}}\,e^{\mathrm{i}\phi'_{p,q}}\right) . \tag{3.26}$$

This procedure permits the generation of a surface with exact PSD, however output skewness and kurtosis will deviate considerably from the input values, most of the time. With the purpose of improving the quality of the output surface skewness and kurtosis, Wu proposes the realization of an iterative procedure on surface statistics. It consists in evaluating the error of S_{sk} and S_{ku} of the output $z_{m,n}$ at the end of the first iteration (Equation (3.26)), and if the result is not satisfactory, one adjusts the required statistics at the output $S^{(z)}_{sk}$ and $S^{(z)}_{ku}$ to new values $S^{(z)_2}_{sk}$ and $S^{(z)_2}_{ku}$, and repeats the procedure from Equations (3.23). This process is based on the idea that, if by prescribing the values of required skewness $S^{(z)}_{sk}$ and $S^{(z)}_{ku}$ in Equations (3.23), the output surface does not match these values, then, by adjusting these parameters, it is possible move towards the correct values. This condenses in an optimization problem, which tests new values of $S^{(z)_i}_{sk}$ and $S^{(z)_i}_{ku}$, then goes from Equations (3.23) to (3.26), and evaluates the quality of the output skewness and kurtosis. The strategy for defining the new values of $S^{(z)_i}_{sk}$ and $S^{(z)_i}_{ku}$ was not specified by Wu, and he just referred to the use of some optimization technique, such as bisection method. Note that the random Gaussian set $\eta_{m,n}$ is generated only once in the algorithm, and the iterative process operates over that particular set, which allows the optimization to be carried only with variables $S^{(z)_i}_{sk}$ and $S^{(z)_i}_{ku}$.

From the numerical experience gained in this work with this algorithm, where Powell's method was used for the optimization procedure, it has been observed that convergence was extremely dependent on the initial Gaussian random set $\eta_{m,n}$—on the initial guess for the optimization problem. This can be justified as: first, based on the fact that output from Gaussian to Johnson transformation will not necessarily verify the required skewness and kurtosis values, which will distort the theoretical output for $z_{m,n}$ in Equation (3.21); last, Equations (3.21) are exact only for completely uncorrelated input signals, whose ACF in non-zero only at the origin, hence they hold approximately true for real signals (Francisco and Brunetière, 2016). These two factors introduce errors in the method, which are dependent on the initial Gaussian set $\eta_{m,n}$—actually, on the combination of the initial random set and prescribed parameters. Furthermore, it was also verified that small variations in $S^{(z)_i}_{sk}$ and $S^{(z)_i}_{ku}$ lead to large changes in $S^{(\eta)_i}_{sk}$ and $S^{(\eta)_i}_{ku}$. Then, if the optimization algorithm is carried with $S^{(\eta)_i}_{sk}$ and $S^{(\eta)_i}_{ku}$ instead, one would have a finer control over the output surface statistics, since smaller steps on $S_{sk}^{(z)_i}$ and $S_{ku}^{(z)_i}$ could be obtained.

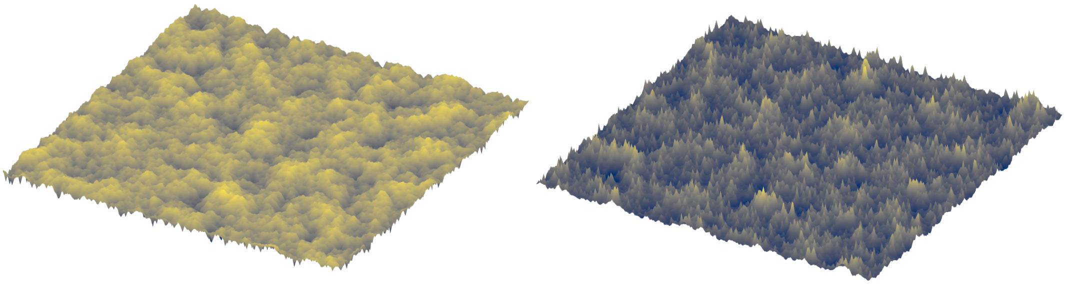
To cope with such limitations, Wu's algorithm was slightly modified, with a rather *brute force* strategy. Before applying the optimization procedure, the previously named first iteration is repeated several times, generating different random Gaussian sets $\eta_{m,n}^{(i)}$, by using $S_{sk}^{(z)}$ and $S_{ku}^{(z)}$ in Equations (3.23). This locks the inputs and, necessarily, the outputs of Hill's algorithm. Then, the set $\eta_{m,n}^{(i)}$ from which the surface with the lowest error on statistics was generated is used as the initial guess, in the optimization procedure. For the optimization, Powell's method is used, and the dependent variables are chosen to be $S_{sk}^{(\eta)_i}$ and $S_{ku}^{(\eta)_i}$. If a large number of surfaces is generated in the trial-and-error process, it is very likely that a solution with statistics very similar to the required ones is generated, hence only fine tuning may be necessary in order to converge to accurate values—which justifies the optimization to be carried on $S_{sk}^{(\eta)_i}$ and $S_{ku}^{(\eta)_i}$.

Remark 3.4 on the computer implementation of Equations (3.23).
The non-Gaussian topography generator requires the computer implementation of Equations (3.23). These expressions, involve a double sum over the coefficient filter vector, whose length is MN. For surfaces, this length can grow quite rapidly, since it depends, roughly speaking, on the square of number of points in each direction. For a 256×256 *surface, this vector would hold 65536 elements. The computation of the double sum is then crucial for computation time. For an efficient computing of this quantity, note that for two successive values of r, the inner sum for both r differs only in one term. Hence, the outer sum shall start on the largest index $r = MN-2$, for which the inner sum reduces to only one term, which is saved into a variable* `s`*. The double sum value* `ds`*, initialized at zero, is incremented by* `s` *for each r. For decreasing r, the term h_{r+1}^2 is added to* `s` *and* `ds` *is incremented by h_r^2·*`s`*. This strategy reduces the number of operations, and greatly decreases computation time for large data sets.*

The flowchart of the non-Gaussian surface generator algorithm is presented on Figure 3.11. Note that some operations already presented for the Gaussian algorithm (cf. Figure 3.11, on page 58) are condensed in Figure 3.11, in order to simplify the diagram and focus on the additional aspects introduced in the present method. Also for this purpose, other two sub-processes, namely, the computation of required statistics for white noise, and the sequence for generating a non-Gaussian surface from Hill's algorithm, are presented in Figures 3.15 and 3.16. Observe that the first step of the algorithm is the computation of amplitudes from the input ACF or PSD, followed by the initial necessary skewness $S_{sk}^{(\eta)}$ and kurtosis $S_{ku}^{(\eta)}$ for the white noise $\eta_{m,n}$. Then, a finite number of Gaussian sets $\eta_{m,n}^{(i)}$ are randomly generated, and transformed with Hill's algorithms to verify the aforementioned statistics $S_{sk}^{(\eta)}$ and $S_{ku}^{(\eta)}$. Next, a non-Gaussian surface $z_{m,n}^{(i)}$ is synthesized, following the sequence in Figure 3.16. The best surface $z_{m,n}^{(i)}$ is selected as the initial guess $z_{m,n}^{[1]}$ for Powell's method. It is important to remark the change of notation, where the superscript (i) indicates a surface generated in the trial-and-error process for selecting the best random set, and $[j]$ relates to the result of each iteration of the optimization process. By picking the best surface $z_{m,n}^{(i)}$, one is actually saving the random phases $\phi_{p,q}^{(i)}$, which result in the smallest statistics error, amongst all randomly generated sets. Finally, in the optimization process, the values of skewness and kurtosis input in Hill's algorithm, regarding the transformation of $\eta_{m,n}$, are iterated. This process uses $S_{sk}^{(\eta)}$ and $S_{ku}^{(\eta)}$ as starting values, and updates these values to $S_{sk}^{(\eta^{[j]})}$ and $S_{ku}^{(\eta^{[j]})}$ on each iteration $[j]$, until acceptable errors are obtained. In sum, the first block performs an optimization by changing the phase field with a trial-and-error scheme, and the second block improves this result by iterating the skewness and kurtosis specified for the transformation of $\eta_{m,n}$, on Hill's algorithm. Figure 3.14 repeats the flowchart for the non-Gaussian profile generation, and Figures 3.15 and 3.16 present the auxiliary sub-processes. Note that in Figure 3.15 there is no need to transform h_n into a vector, since it is already one, hence the notation comes simplified.

3.3.3 Numerical tests

Numerical performance and accuracy of the non-Gaussian generation algorithm has also been evaluated with simples tests. Since amplitude prescription is analogous in both Gaussian and non-Gaussian algorithms, their performance in recovering the input ACF or PSD is identical. Therefore, tests comparing input and output spectral content will be disregarded—see Section 3.2.1 for these results. Recall that the input PSD is recovered exactly, and output ACF shows increasingly smaller deviations with decreasing autocorrelation length. Since the quality of output PSD and ACF is already verified, focus is given to the algorithm capacity to recover input skewness and kurtosis. Figure 3.17 shows two examples of artificial surfaces generated by the implemented non-Gaussian generator.



(a) $S_{sk} \approx -1$ and $S_{ku} \approx 5$

(b) $S_{sk} \approx 1$ and $S_{ku} \approx 5$

Figure 3.17: Examples of artificially generated isotropic non-Gaussian surfaces. Both surfaces are generated from 256 points in each direction with $L = L_x = L_y = 1\,\text{mm}$, and heights are normalized to verify $S_q = 0.01\,\text{mm}$. The high frequency cut-off is set to $\lambda_s = L/128$, the low to $\lambda = L/4$, there is no roll-off, and $H = 0.2$.

In order to evaluate the algorithm accuracy on surface statistics, examples with different combinations of skewness and kurtosis were performed. For each combination, the number of attempts in the trial-and-error process were varied, without further optimization. For the largest number of attempts analyzed, an additional test was performed, where optimization was introduced. The results of these tests are gathered in Table 3.1. If

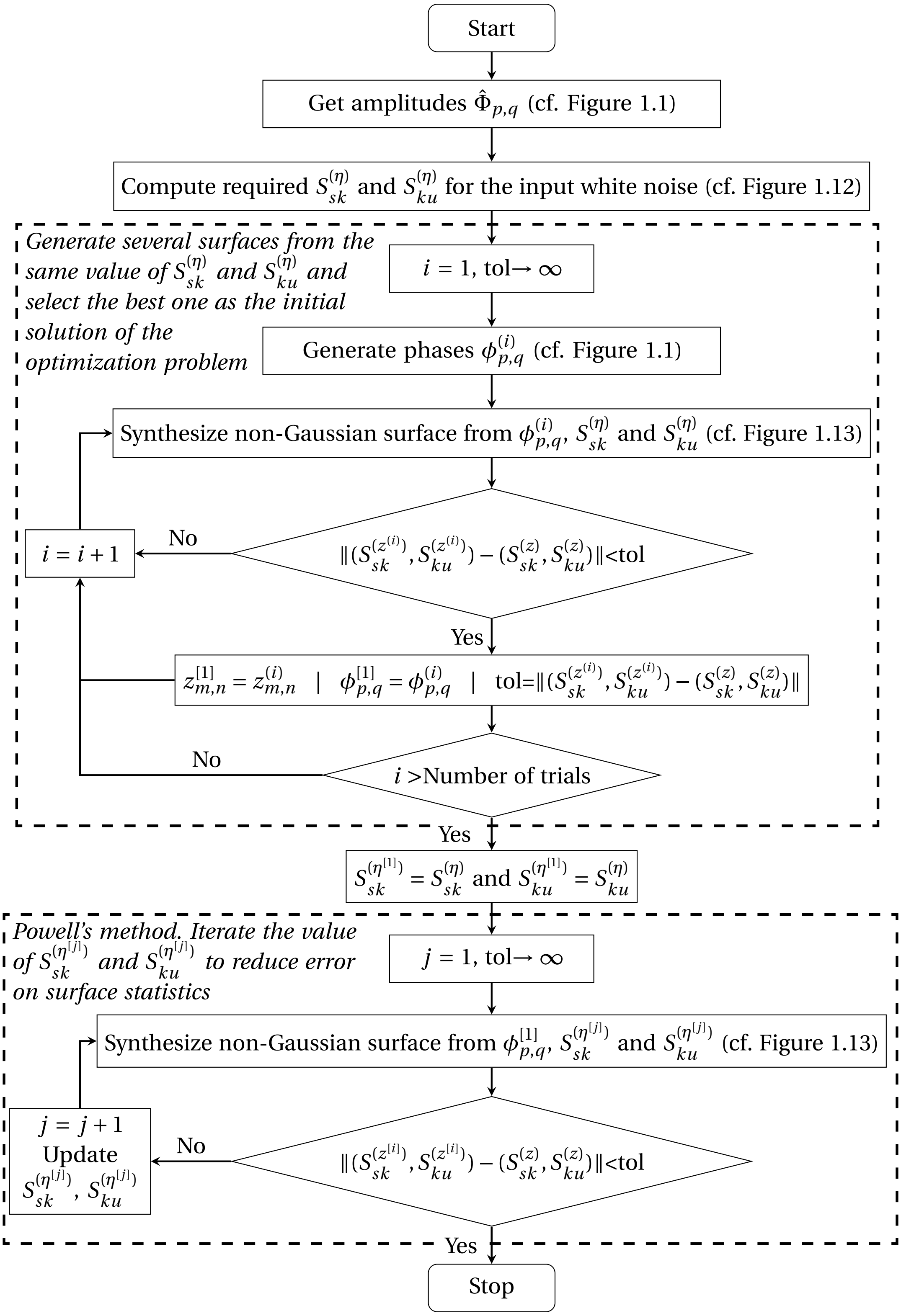

Start
Get amplitudes $\hat{\Phi}_{p,q}$ (cf. Figure 1.1)
Compute required $S_{sk}^{(\eta)}$ and $S_{ku}^{(\eta)}$ for the input white noise (cf. Figure 1.12)
Generate several surfaces from the same value of $S_{sk}^{(\eta)}$ and $S_{ku}^{(\eta)}$ and select the best one as the initial solution of the optimization problem
i = 1, tol→ ∞
Generate phases $\phi_{p,q}^{(i)}$ (cf. Figure 1.1)
Synthesize non-Gaussian surface from $\phi_{p,q}^{(i)}$, $S_{sk}^{(\eta)}$ and $S_{ku}^{(\eta)}$ (cf. Figure 1.13)
$\|(S_{sk}^{(z^{(i)})}, S_{ku}^{(z^{(i)})}) - (S_{sk}^{(z)}, S_{ku}^{(z)})\|$<tol
No
$i = i + 1$
Yes
$z_{m,n}^{[1]} = z_{m,n}^{(i)}$ | $\phi_{p,q}^{[1]} = \phi_{p,q}^{(i)}$ | tol=$\|(S_{sk}^{(z^{(i)})}, S_{ku}^{(z^{(i)})}) - (S_{sk}^{(z)}, S_{ku}^{(z)})\|$
i >Number of trials
No
Yes
$S_{sk}^{(\eta^{[1]})} = S_{sk}^{(\eta)}$ and $S_{ku}^{(\eta^{[1]})} = S_{ku}^{(\eta)}$
Powell's method. Iterate the value of $S_{sk}^{(\eta^{[j]})}$ and $S_{ku}^{(\eta^{[j]})}$ to reduce error on surface statistics
j = 1, tol→ ∞
Synthesize non-Gaussian surface from $\phi_{p,q}^{[1]}$, $S_{sk}^{(\eta^{[j]})}$ and $S_{ku}^{(\eta^{[j]})}$ (cf. Figure 1.13)
$\|(S_{sk}^{(z^{[i]})}, S_{ku}^{(z^{[i]})}) - (S_{sk}^{(z)}, S_{ku}^{(z)})\|$<tol
No
$j = j + 1$
Update
$S_{sk}^{(\eta^{[j]})}$, $S_{ku}^{(\eta^{[j]})}$
Yes
Stop

Figure 3.11: Flowchart of non-Gaussian random rough surface generation algorithm.

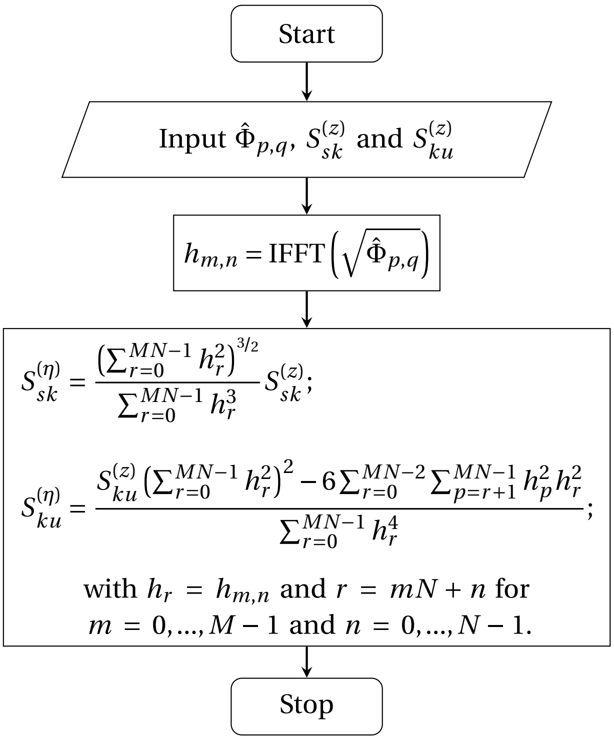

Start
Input $\hat{\Phi}_{p,q}$, $S_{sk}^{(z)}$ and $S_{ku}^{(z)}$
$h_{m,n} = \text{IFFT}\left(\sqrt{\hat{\Phi}_{p,q}}\right)$
$S_{sk}^{(\eta)} = \frac{\left(\sum_{r=0}^{MN-1} h_r^2\right)^{3/2}}{\sum_{r=0}^{MN-1} h_r^3} S_{sk}^{(z)};$
$S_{ku}^{(\eta)} = \frac{S_{ku}^{(z)}\left(\sum_{r=0}^{MN-1} h_r^2\right)^2 - 6\sum_{r=0}^{MN-2}\sum_{p=r+1}^{MN-1} h_p^2 h_r^2}{\sum_{r=0}^{MN-1} h_r^4};$
with $h_r = h_{m,n}$ and $r = mN + n$ for $m = 0, \dots, M-1$ and $n = 0, \dots, N-1$.
Stop

Figure 3.12: Required skewness and kurtosis for surface white noise: block flowchart.

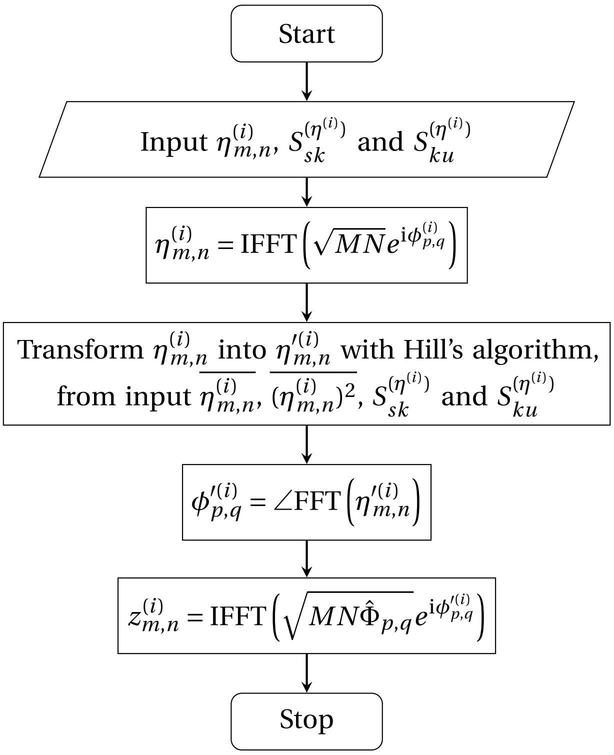

Start
Input $\eta_{m,n}^{(i)}$, $S_{sk}^{(\eta^{(i)})}$ and $S_{ku}^{(\eta^{(i)})}$
$\eta_{m,n}^{(i)} = \text{IFFT}\left(\sqrt{MN} e^{\mathrm{i}\phi_{p,q}^{(i)}}\right)$
Transform $\eta_{m,n}^{(i)}$ into $\eta_{m,n}^{\prime(i)}$ with Hill's algorithm, from input $\overline{\eta_{m,n}^{(i)}}$, $\overline{(\eta_{m,n}^{(i)})^2}$, $S_{sk}^{(\eta^{(i)})}$ and $S_{ku}^{(\eta^{(i)})}$
$\phi_{p,q}^{\prime(i)} = \angle\text{FFT}\left(\eta_{m,n}^{\prime(i)}\right)$
$z_{m,n}^{(i)} = \text{IFFT}\left(\sqrt{MN\hat{\Phi}_{p,q}} e^{\mathrm{i}\phi_{p,q}^{\prime(i)}}\right)$
Stop

Figure 3.13: Non-gaussian surface generation sequence: block flowchart.

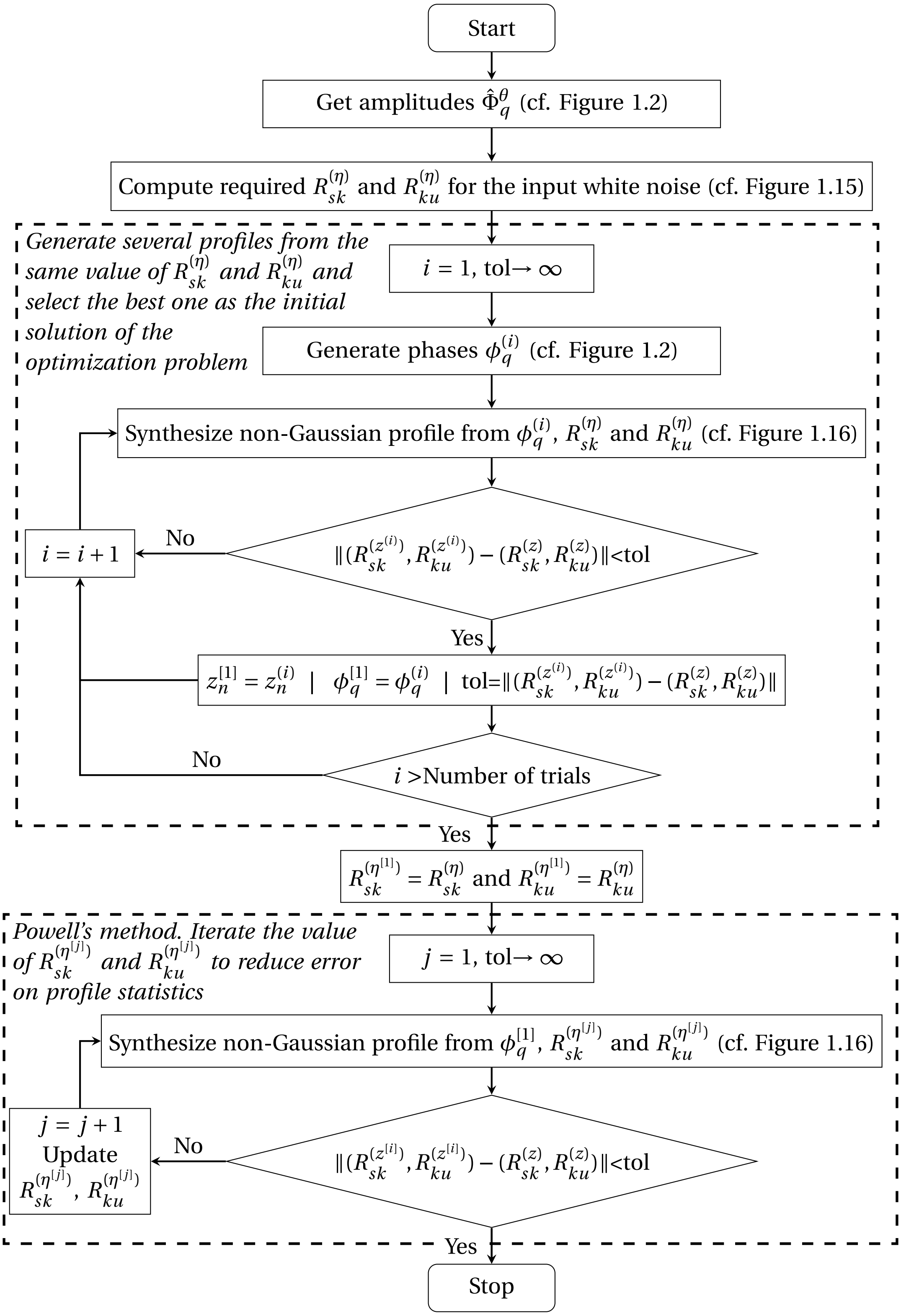

Start
Get amplitudes $\hat{\Phi}_q^{\theta}$ (cf. Figure 1.2)
Compute required $R_{sk}^{(\eta)}$ and $R_{ku}^{(\eta)}$ for the input white noise (cf. Figure 1.15)
Generate several profiles from the same value of $R_{sk}^{(\eta)}$ and $R_{ku}^{(\eta)}$ and select the best one as the initial solution of the optimization problem
$i = 1$, tol→ ∞
Generate phases $\phi_q^{(i)}$ (cf. Figure 1.2)
Synthesize non-Gaussian profile from $\phi_q^{(i)}$, $R_{sk}^{(\eta)}$ and $R_{ku}^{(\eta)}$ (cf. Figure 1.16)
$\|(R_{sk}^{(z^{(i)})}, R_{ku}^{(z^{(i)})}) - (R_{sk}^{(z)}, R_{ku}^{(z)})\|$<tol
No
$i = i + 1$
Yes
$z_n^{[1]} = z_n^{(i)}$ | $\phi_q^{[1]} = \phi_q^{(i)}$ | tol=$\|(R_{sk}^{(z^{(i)})}, R_{ku}^{(z^{(i)})}) - (R_{sk}^{(z)}, R_{ku}^{(z)})\|$
i >Number of trials
No
Yes
$R_{sk}^{(\eta^{[1]})} = R_{sk}^{(\eta)}$ and $R_{ku}^{(\eta^{[1]})} = R_{ku}^{(\eta)}$
Powell's method. Iterate the value of $R_{sk}^{(\eta^{[j]})}$ and $R_{ku}^{(\eta^{[j]})}$ to reduce error on profile statistics
$j = 1$, tol→ ∞
Synthesize non-Gaussian profile from $\phi_q^{[1]}$, $R_{sk}^{(\eta^{[j]})}$ and $R_{ku}^{(\eta^{[j]})}$ (cf. Figure 1.16)
$\|(R_{sk}^{(z^{[i]})}, R_{ku}^{(z^{[i]})}) - (R_{sk}^{(z)}, R_{ku}^{(z)})\|$<tol
No
$j = j + 1$
Update
$R_{sk}^{(\eta^{[j]})}$, $R_{ku}^{(\eta^{[j]})}$
Yes
Stop

Figure 3.14: Flowchart of non-Gaussian random rough profile generation algorithm.

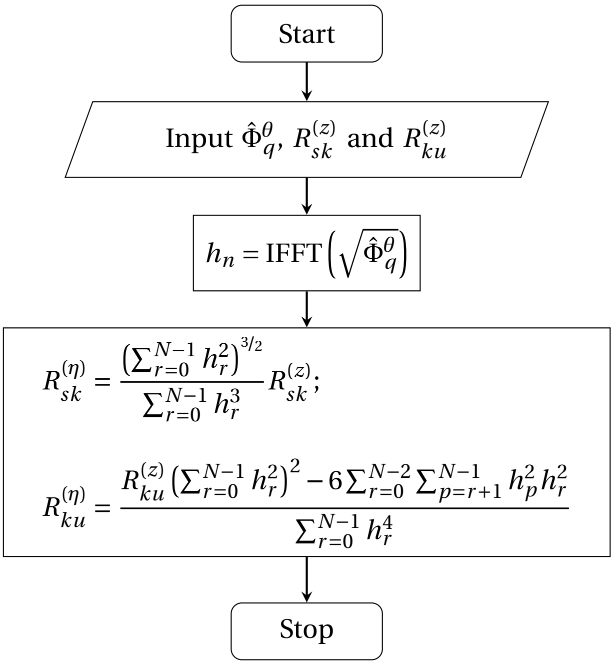

Start
Input $\hat{\Phi}_q^\theta$, $R_{sk}^{(z)}$ and $R_{ku}^{(z)}$
$h_n = \mathrm{IFFT}\left(\sqrt{\hat{\Phi}_q^\theta}\right)$
$R_{sk}^{(\eta)} = \frac{\left(\sum_{r=0}^{N-1} h_r^2\right)^{3/2}}{\sum_{r=0}^{N-1} h_r^3} R_{sk}^{(z)};$
$R_{ku}^{(\eta)} = \frac{R_{ku}^{(z)}\left(\sum_{r=0}^{N-1} h_r^2\right)^2 - 6\sum_{r=0}^{N-2}\sum_{p=r+1}^{N-1} h_p^2 h_r^2}{\sum_{r=0}^{N-1} h_r^4}$
Stop

Figure 3.15: Required skewness and kurtosis for profile white noise: block flowchart.

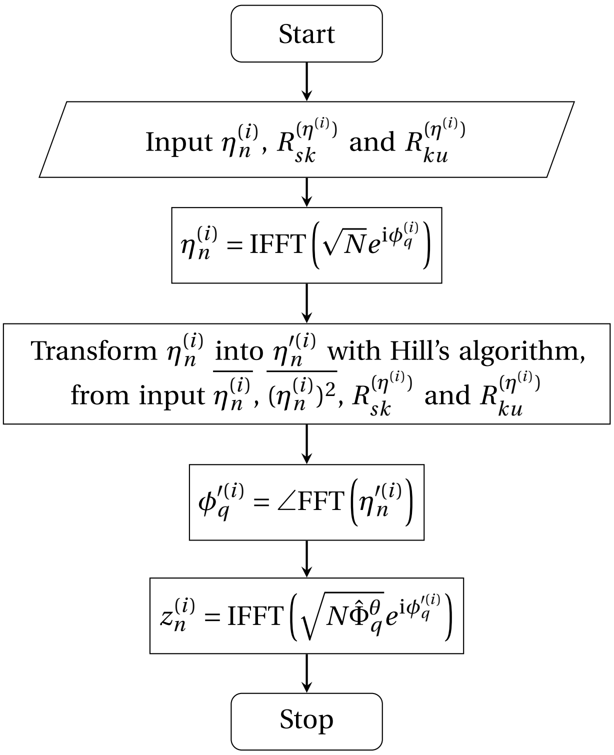

Start
Input $\eta_n^{(i)}$, $R_{sk}^{(\eta^{(i)})}$ and $R_{ku}^{(\eta^{(i)})}$
$\eta_n^{(i)} = \mathrm{IFFT}\left(\sqrt{N} e^{\mathrm{i}\phi_q^{(i)}}\right)$
Transform $\eta_n^{(i)}$ into $\eta_n'^{(i)}$ with Hill's algorithm,
from input $\overline{\eta_n^{(i)}}$, $\overline{(\eta_n^{(i)})^2}$, $R_{sk}^{(\eta^{(i)})}$ and $R_{ku}^{(\eta^{(i)})}$
$\phi_q'^{(i)} = \angle\mathrm{FFT}\left(\eta_n'^{(i)}\right)$
$z_n^{(i)} = \mathrm{IFFT}\left(\sqrt{N\hat{\Phi}_q^\theta} e^{\mathrm{i}\phi_q'^{(i)}}\right)$
Stop

Figure 3.16: Non-gaussian profile generation sequence: block flowchart.

a single surface is generated in the trial-and-error process, and optimization is ignored, the output surface cannot reproduce surface statistics accurately, in most cases. Yet, by increasing the number of attempts, surface skewness and kurtosis get closer to prescribed values. Exact values for these statistics are practically recovered for non-skewed surfaces ($S_{sk} = 0$). When non-zero skewness is specified, the algorithm convergence slows down, and results show larger deviations, for the same number of attempts, in the selection of the initial guess. When optimization is performed, results improve considerably, at the cost of increasing computation time. Thus, in general, the algorithm can reproduce input statistics if a sufficiently good initial guess for the optimization problem can be found—if a reasonable number of attempts are generated, to find a good initial solution for the optimization algorithm. In general, 100 trials are enough to ensure good precision on the results, yet this depends on several factors, from which the number of points is stressed. Extending the number of attempts further leads to an inconvenient increase of computation time, and it may not necessarily reduce statistics error. The fact that when non-zero skewness are prescribed, algorithm convergence is impacted is also a paramount result. Also, note that while computation time for generating a 1024 × 1024 rough surface using the Gaussian generation was about 1 second, the generation of a 256 × 256 non-Gaussian surface with relatively accurate statistics can take around 20 seconds (in a personal computer, with a quad-core Intel® Core™ i7-7700HQ CPU at 2.8 GHz). Thus, the generation of a non-Gaussian surface with 8 times less points can take up to 20 times longer, which is also a key point to mention, regarding the performance analysis of this algorithm.

Recalling Figure 3.8, on page 65, similar figures can be produced, in order to check if height distribution of artificial topography is non-Gaussian, by visual inspection of the probability density function. Following the same test structure adopted for Gaussian surface testing, 100 random non-Gaussian surfaces with 1024 × 1024 were generated from a fractal PSD, requiring $S_{sk} = -0.5$ and $S_{ku} = 4$. For each realization, the height distribution is computed, using 1000 bins between the maximum and minimum generated height. The average PDF is plotted in Figure 3.18, alongside with the Gaussian reference—a Gaussian PDF with mean and standard deviation computed across all generated surfaces. As expected, the non-Gaussian averaged PDF shows a higher peak relative to the Gaussian curve, since the specified kurtosis is higher than 3, and its peak is moved to the right of the mean value—zero, in this case—which is also normal for negatively skewed surfaces (see Section 2.3.4).

As it was seen from Table 3.1, convergence for accurate statistics may prove difficult, specially when the prescribed skewness is high. Other very inconvenient scenario happens when convergence is not met at all, and unrealistic topography is generated, with skewness and kurtosis very different from the prescribed. In Figure 3.19, a profile which was produced by the non-Gaussian algorithm in such conditions is plotted. This is explained by the fact that skewness and kurtosis ($R_{sk}^{(\eta)}$, $R_{ku}^{(\eta)}$), resulting from Equations (3.23), do not necessarily verify the inequality expressed by Equation (2.49). This is, the required skewness and kurtosis for the input white noise may violate the limit established for Johnson frequency curves, hence Hill's algorithm cannot fit any distribution, and sequences with those statistics are impossible to generate with Johnson system (cf. Figure 3.10). Since the required statistics for $\eta_{m,n}$ cannot be computed, all the algorithm is compro-

Table 3.1: Output surface skewness and kurtosis from non-Gaussian surface generator. Non-Gaussian surfaces with 256 × 256 points and $L = L_x = L_y = 1\,\mathrm{mm}$ are generated from a fractal PSD verifying $\lambda_l = L/4$, $\lambda_r = \lambda_l$, $\lambda_s = L/127$, $H = 0.2$ and $g = 1$. Surfaces are generated for several input skewness and kurtosis, with different number of attempts in the trial-and-error process, with and without optimization.

		Number of trials			
	Required	1	10	100	100 + opt.
S_{sk}	0.00000	−0.00149	−0.00149	0.01119	0.00000
S_{ku}	3.00000	2.98724	2.98724	3.00124	3.00000
Time/s	-	0.14610	0.31897	2.04780	10.59405
S_{sk}	0.00000	−0.03523	0.03142	0.03142	0.00000
S_{ku}	4.00000	3.22111	3.58994	3.58994	4.00000
Time/s		0.14749	0.32911	2.12166	11.33867
S_{sk}	−0.75000	−0.48945	−0.52122	−0.62639	−0.73081
S_{ku}	4.00000	3.51683	3.70461	3.80637	4.00424
Time/s		0.15130	0.40695	2.99757	19.45790
S_{sk}	0.00000	−0.05110	0.06620	−0.15507	0.00000
S_{ku}	5.00000	3.39848	4.29412	4.65645	5.00000
Time/s		0.14610	0.32887	2.15530	18.00355
S_{sk}	−1.00000	−0.69050	−0.70674	−0.77952	−0.86319
S_{ku}	5.00000	4.14864	4.41829	4.54488	5.01758
Time/s		0.15427	0.43070	3.28910	21.38146
S_{sk}	0.50000	0.10348	0.38700	0.38700	0.50000
S_{ku}	8.00000	3.70467	6.54360	6.54360	8.00000
Time/s		0.14469	0.33831	2.20978	16.51218

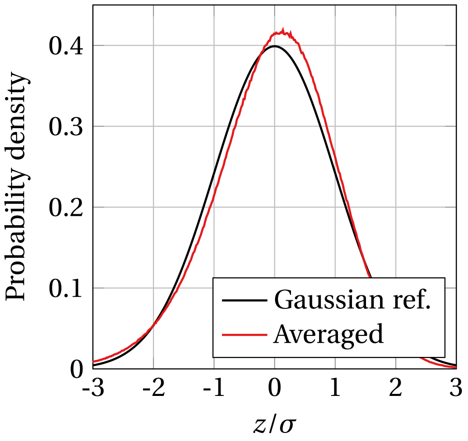


Figure 3.18: Probability density of artificial non-Gaussian surfaces. 100 non-Gaussian surfaces were generated with 1024 × 1024 from a fractal PSD, with $S_{sk} = -0.5$ and $S_{ku} = 4$. A Gaussian curve with same mean and standard deviation of all generated heights is plotted along with the averaged surface PDF. The averaged probability density shows a higher peak, which is moved into positive values of z. This is expected behavior for kurtosis higher than 3 and negatively skewed surfaces.

mised, and the result will diverge from the prescribed values. The divergence of the non-Gaussian generation algorithm is unpredictable, once it depends on the required statistics for $z_{m,n}$ and on the filter coefficients of Equations (3.23). In turn, these coefficients are related to the input ACF or PSD and the number of points. The synthesis of non-Gaussian rough topography must be preceded by a simple check, with the purpose of verifying whether the algorithm can converge for that case or not. From numerical experience, high Hurst exponents H and larger low wavelength cut-off, i.e., high amplitude low frequencies, tend to narrow the range of applicability of the algorithm. On the other hand, low H allow the simulation of a wider range of skewness and kurtosis on the output.

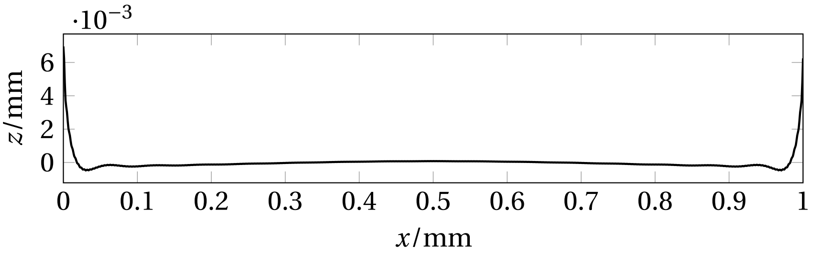


Figure 3.19: Example of artificial profile generated by the non-Gaussian algorithm, when convergence is not met. This profile is generated from a fractal PSD with $R_{sk} = -1$ and $R_{ku} = 4$, and the outputs verifies $R_{sk} = 8.1607$ and $R_{ku} = 82.17906$.

3.4 Application to real topography

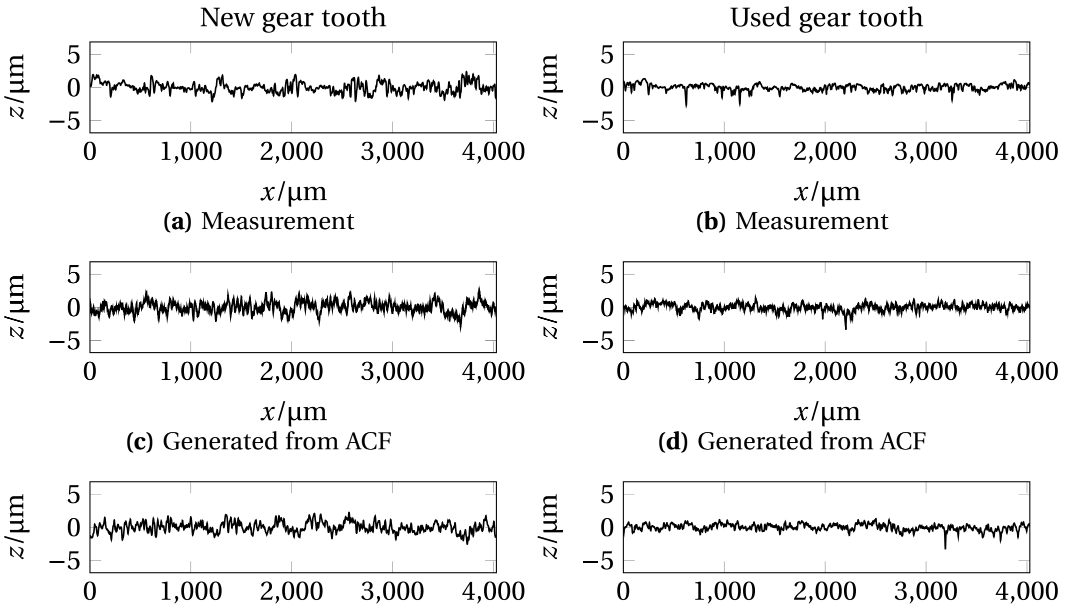
Following the numerical validation of rough topography generation algorithms, this topic is closed with their application to real roughness measurements. Experimental data on profile roughness, obtained with a stylus device, and areal measurements, acquired by an optical instrument based on interferometry, were kindly provided by CETRIB, a tribology laboratory in Faculty of Engineering of University of Porto (FEUP). Profile measurements were performed on the flank of a gear tooth with the contact profilometer Hommelwerke LV-50, equipped with a TKL 300 stylus probe, with tip radius 0.5 µm. The flank roughness was in two different stages: the first measurement concerned the surface after machining, termed the new tooth from here on; the second measurement was carried after the gear was subjected to power loss tests, and the measured tooth will be termed the used tooth. Profiles were measured along a sampling length equal to 4.8 mm (six times the cut-off wavelength 0.8 mm), and with a sampling interval of 1 µm. A Gaussian filter is then applied to profile data, which is finally trimmed, so that it results in a 4 mm long scan. Regarding surface measurements, the raceways of the housing washer from a cylindrical roller thrust bearing SKF® 81107 TN was used. Measurements were also performed before and after power loss tests, involving the roller bearing, using the optical measurement device Burker NPFLEX™. In order to perform areal measurements over a 5 mm×5 mm area, since the equipment cannot focus regions that large at once, a *stitching* procedure is adopted to link all smaller measured regions. The mean value, and a polynomial background is removed from surface data, and outliers, which may indicate badly measured points, are discarded. The application of numerical generation algorithms to real topography starts with the computation of real statistical and spectral measures—ACF, PSD, skewness and kurtosis. Then a random artificial topography is generated from these quantities, and it is checked whether those input measures are verified in the output topography. As an ultimate resource, one validates topography reconstruction by visual inspection. Note that all surfaces and profiles presented verify the zero mean conditions, regarding roughness definition, and that PSD gives directly the standard deviation of PDF, i.e., either R_q or S_q—recall Section 2.6. For this reason, standard deviation and mean shall not be referred when comparing heights distribution.

Starting with profile generation, Figure 3.20 shows the real topography of the profiles measured on the new and used tooth, side by side with artificial topography generated from both real ACF and PSD. Profile skewness and kurtosis, for real and artificial profiles, is presented in Table 3.2. Note that the new tooth is nearly Gaussian, hence, artificial profile generation may follow the Gaussian algorithm, in this case. On the other hand, the used tooth is clearly non-Gaussian, then artificial profiles regarding this case are synthesized with the non-Gaussian algorithm, from 1000 attempts in the trial-and-error process, followed by the optimization procedure. From Table 3.2, one sees that for the used tooth (non-Gaussian), these parameters are recovered with zero error, while for the new tooth (Gaussian), since no restrictions are imposed on these statistics, the recomputed values are not exact, yet they are acceptably close to the reference. As a side note, it can be seen that the power loss test led to a reduction of R_q (see ACF value at the origin on Figure 3.21), and transformed the height distribution, from an initially Gaussian, into a

Table 3.2: Skewness and kurtosis of experimentally measured profiles and artificial profiles generated from ACF and PSD.

	New tooth			Used tooth		
	Real	ACF	PSD	Real	ACF	PSD
R_{sk}	0.242411	-0.086841	0.041112	-0.640556	-0.640556	-0.640556
R_{ku}	2.925664	3.167411	2.952987	5.177983	5.177983	5.177983

negatively skewed and leptokurtic—the power loss test removed roughness peaks, but as little influence on the valleys. Figure 3.21 plots the ACF of the experimentally measured profile and of the artificial one, generated from the real ACF. Autocorrelation function of the output profiles match closely the real ACF. An exact fit was not expected, based on the conclusions obtained in Section 3.2.1. Power spectrum of real profiles is plotted in Figure 3.22, and it matches the PSD of any artificial profile generated from it, following earlier observations. It is interesting to point out the similitude between the real profiles PSD and a theoretical fractal PSD (cf. Figure 2.17). For very high wavenumbers, there is a small tail on the PSD, deviating from the fractal-like curve. This is mainly due to measurement noise, and similar results have been reported by Panda *et al.* (2016).



(a) Measurement

(b) Measurement

(c) Generated from ACF

(d) Generated from ACF

(e) Generated from PSD

(f) Generated from PSD

Figure 3.20: Profiles measured on gear tooth flanks, after machining and power loss test, in comparison with artificially generated profiles from input ACF and PSD.

Finally, by visual inspection of Figure 3.20, one sees that profiles generated from input ACF seem more noisy than the original. In general, generation from input PSD produces profiles which are visually similar to the reference, even though statistically, both genera-

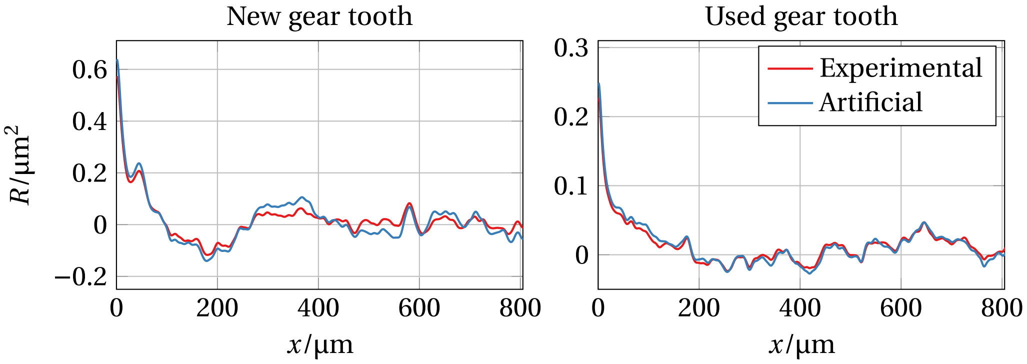


Figure 3.21: Autocorrelation function of experimental profiles and recovered from artificial profiles, generated from input ACF.

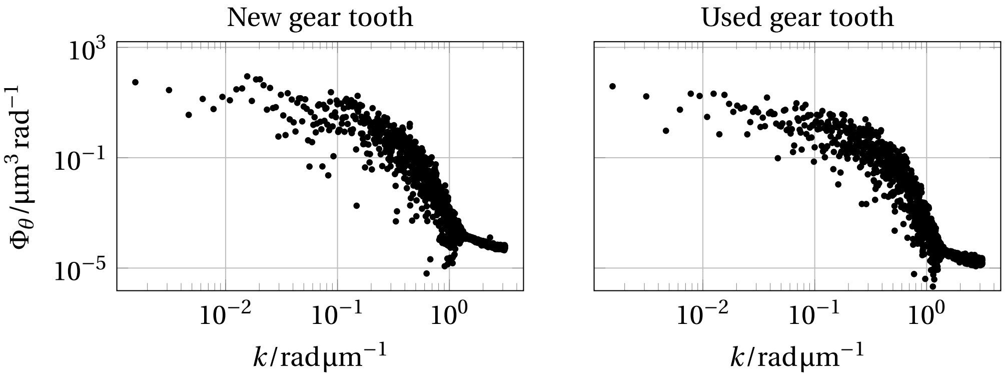


Figure 3.22: PSD of the experimental and artificial profiles.

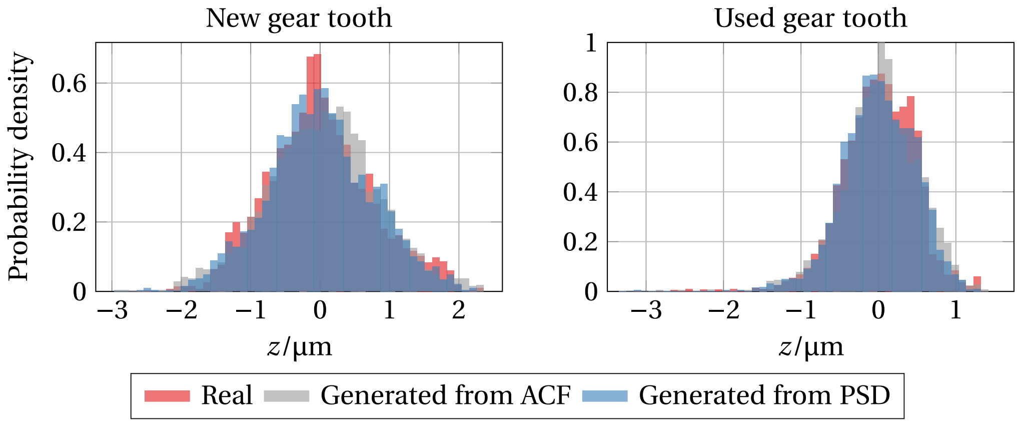


Figure 3.23: Heights distribution of the experimental profile and artificial profiles generated from input ACF and PSD.

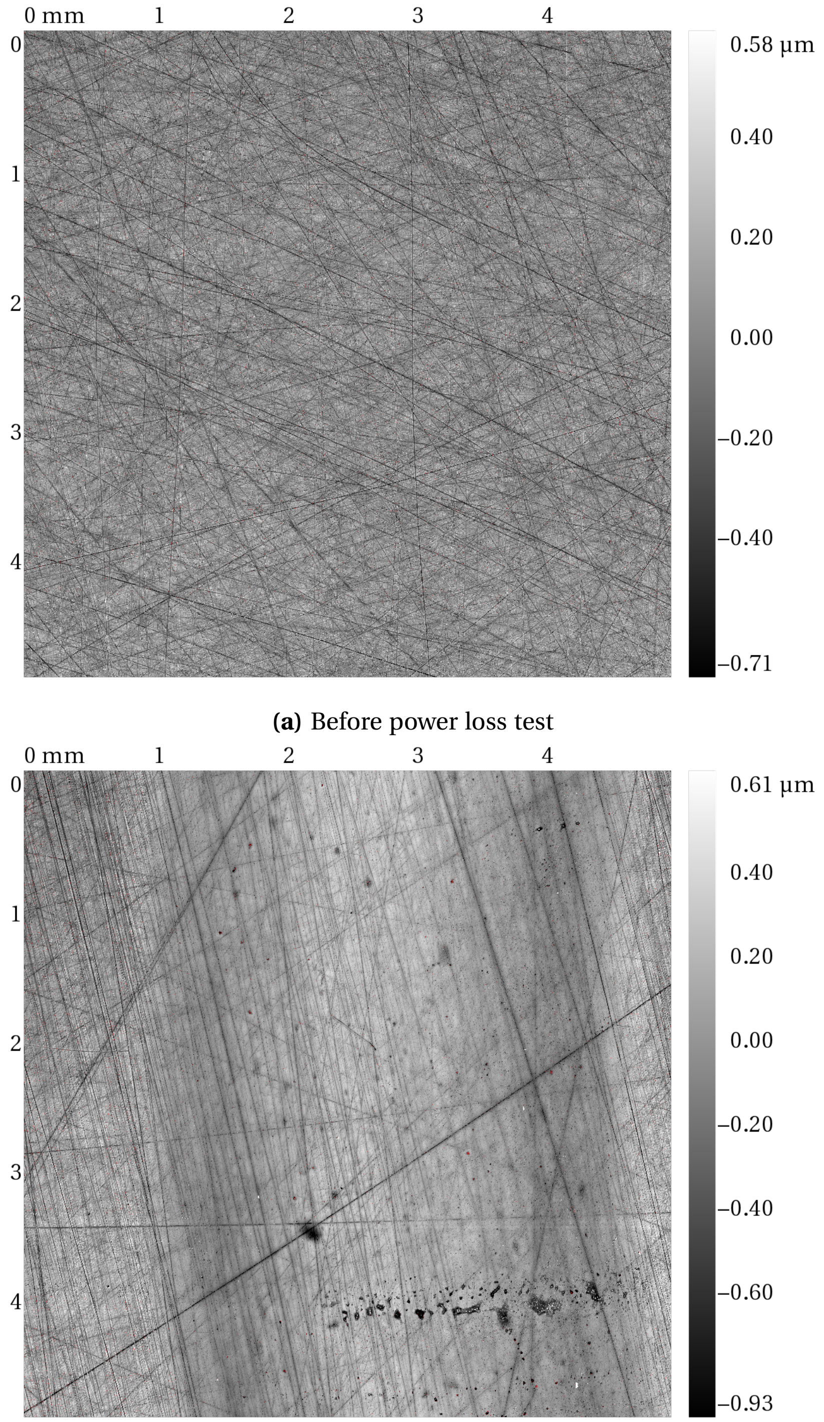
tion alternatives give accurate results. Peaks on generated profiles from PSD look sharper, yet all statistical and spectral properties are identical between real and artificial topographies. Height distribution for all cases are plotted in Figure 3.23. Naturally, the height distribution differs slightly between real and artificial profile. It is suggested, based on the height distribution referring to the real profile of the used tooth, that skewness and kurtosis cannot completely describe the heights PDF. In fact, this distribution appears too be bimodal, from the existence of two separate peaks of probability density. This is supported by a previous statement on the height distributions of stratified surface. For such surfaces, heights distribution results from the superposition of two independent distributions, produced by different processes—machining and wear, in this case. This effect cannot be reproduced by Johnson system, which raises a new limitation for the applicability of the presented algorithms to such profiles, where skewness and kurtosis are insufficient to characterize heights distribution.

With respect to the generation of artificial surfaces, Figure 3.24 shows the 5 mm by 5 mm image resulting from the measurement performed on the washer raceway, after leveling, preceding and proceeding the power loss test—the positive direction moves out of the page. About 5000 points were recorded in each direction, which represents a dataset which is oversized for a practical numerical application. Thus, in order to apply any generation algorithm, one shall restrict to a small patch of the surface with dimensions 1 mm×1 mm, that is discretized in 1022 points in each direction. Figure 3.25 shows the region of interest for the artificial surface generation. In these two cases, scratches are visible on the surface, which may result either from machining or wear. Apart of these scars, one can say that before the power loss test, the surface is isotropic, with randomly oriented scratches, while after the power loss test it is markedly anisotropic, and scratches are preferably oriented in one direction.

Artificial surfaces were generated only from PSD, based on previous observations on profile generation. Table 3.3 shows the skewness and kurtosis of the real and artificially generated surfaces. Both experimental surfaces are non-Gaussian, hence the non-Gaussian generator was applied. For each case, the non-Gaussian algorithm was employed with 1000 trials for the selection of the initial solution for the optimization process, followed by the optimization. Skewness and kurtosis are recovered almost exactly, regarding the surface previous to the power loss test. For the surface proceeding the test, both kurtosis and skewness increase, in magnitude, and the algorithm converges to a solution with relatively small error on kurtosis, and a larger deviation on skewness—this is mainly due to the prescription of a larger value, in magnitude, of skewness.

Table 3.3: Skewness and kurtosis of experimentally measured surfaces and artificial surfaces generated from PSD.

	Before test		After test	
	Real	PSD	Real	PSD
S_{sk}	-0.334944	-0.334944	-1.027649	-0.674759
S_{ku}	4.506406	4.506406	6.542165	6.569589



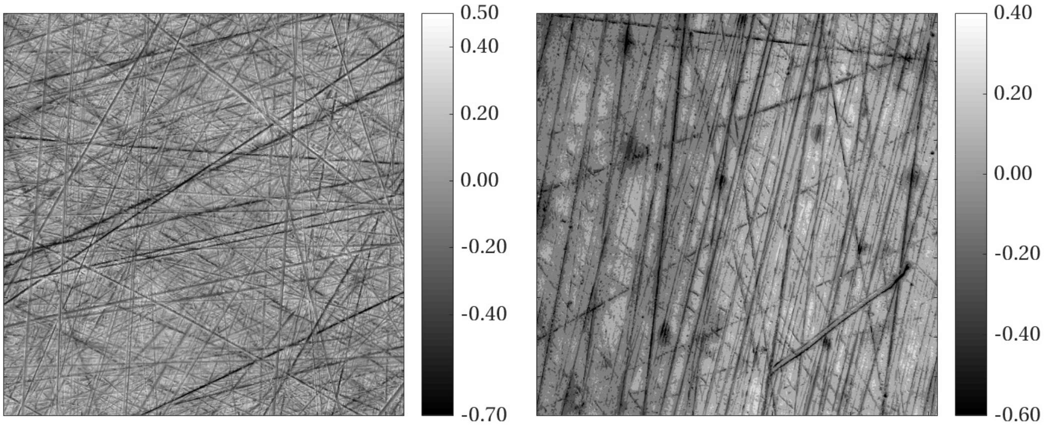
(a) Before power loss test

(b) After power loss test

Figure 3.24: Roughness areal measurement on the housing washer of the roller bearing in analysis, obtained with an optical instrument. The positive direction moves out of the page.

The artificial surfaces generated from the measured surfaces' PSD are plotted in Figure 3.26. Analyzing this figure, one can conclude that the overall trend of surface roughness is captured by the algorithm, even though individual marks are not exactly generated. In particular, the algorithm reproduces the mean isotropic and anisotropic characteristic of surface roughness correctly, at the cost of neglecting local topography features. As a matter of fact, the individual scratches on each surface are likely to result from the synchronization of specific harmonics in the surface, which is completely out of the scope of the implemented generation algorithm.

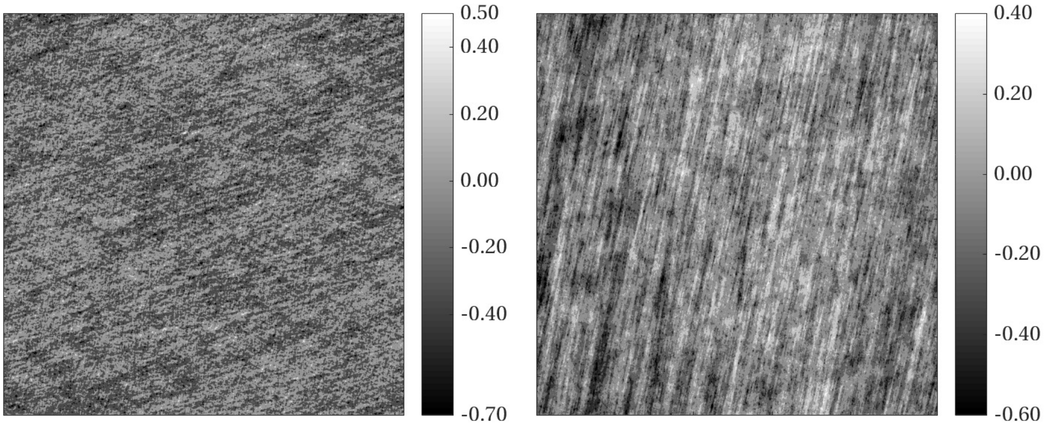
To complete the comparison between the two topographies, Figure 3.27 plots the heights distribution for real and artificial surface, in both scenarios in analysis. Differences between heights PDF are small, even though skewness and kurtosis are very similar. This results does also suggest that these parameters do not define the probability distribution in a strict sense, and small deviations in the heights distribution may still occur for the same PDF moments.



(a) Before the test

(b) After the test

Figure 3.25: Square region, with side 1 mm and 1022 points in each direction, used for surface generation. Wear and machining marks are noticeable in the surface in both cases. however, before the power loss test, the surface is generally isotropic, and marks are randomly oriented, while after the test, it is greatly anisotropic, with marks preferably oriented in one direction.



(a) Before test

(b) After test

Figure 3.26: Artificial surfaces generated by the PSD computed from surfaces plotted in Figure 3.25.The general trend of surface geometry is captured, namely isotropy and anisotropy. However, individual surface marks cannot be generated.

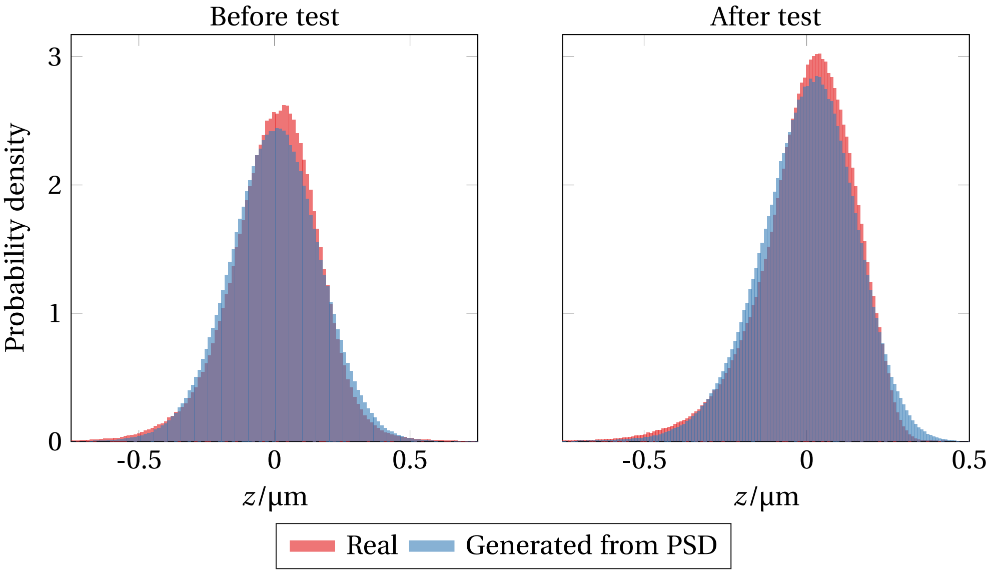


Figure 3.27: Heights distribution of real surface measured in an optical instrument, compared with the same distribution regarding a synthetic surface, generated from the real PSD.

Chapter 4

Micromechanical elastic contact: analytical models

When two nominally flat surfaces come into contact, they only touch locally at regions where the roughness features of both interfere. As a consequence, the *real* contact area is expected to be smaller than the *apparent*, or nominal, contact area. This has already been demonstrated by several experiments, e.g., by Dieterich and Kilgore (1994) (see Figure 4.1). In fact, owing to the multiscale nature of roughness, the real contact area is resolution dependent, and can be also characterized by fractals (Borri-Brunetto *et al.*, 1999). Although the surfaces of solids are typically assumed smooth in classical contact mechanics problems, the ability to evaluate the real contact area is paramount in several applications, such as contact conductance, sealing, wear and friction. The range of values that it can take depends on each specific situation. In general, one can consider 20% as a reference for the upper bound of real contact area fraction, for typical applications.

The link between friction and real contact area reports back to Leonardo da Vinci (1452-1519), who formulated that the frictional force is proportional to weight, and does not depend on the apparent contact area. However, it was through Guillaume Amontons (1663-1705) that the laws of friction were first recognized by the scientific community. Among Amontons' postulates, the proportionality between friction force and applied normal pressure and its independence on sliding velocity are the most notorious. Following Amontons' work, Charles-Augustin de Coulomb (1736-1806) carried an extensive experimental research, in order to investigate the influence of several physical parameters on the coefficient of friction, such as contact time and size.[1]

Early experimental data, and also recent numerical results, suggest that the frictional force is proportional to real contact area. Following this observation and Amonton's laws, it can be stated that the contact area is proportional to the applied normal load. This was initially explained by Bowden and Tabor, in 1950, based on the hypothesis that roughness summits would undergo plastic deformation very rapidly after contact—the friction force

[1] An interesting review of the work of Amontons and Coulomb on friction was authored by Popova and Popov (2015). There is no consensus about the individual to whom the laws of friction shall be attributed. Some authors suggest that laws of friction should be named after Leonardo da Vinci, instead (Israelachvili, 2015). In numerous textbooks the designation *Coulomb's laws of friction* is preferred.

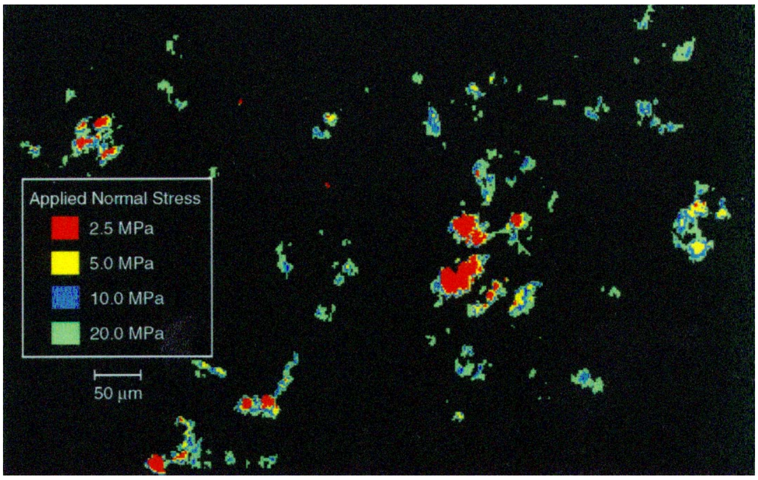


Figure 4.1: Photomicrograph from Dieterich and Kilgore (1994) highlighting the real contact area between a surface of acrylic plastic and other of soda-lime glass, for different load stages. With increasing load, the size of each contact spot increases, together with the number of contact spots.

would come as the material's shear strength and the real contact area (Bowden *et al.*, 2001). However, their theory proved unrealistic for several practical applications.

Micromechanical contact theories generally describe a constitutive law for the physics at the contact interface between rough surfaces, establishing relations between the applied load, separation and contact area. When the contact between two elastic surfaces is considered to be frictionless and nonadhesive, the problem can be reduced to the contact of an equivalent elastic flat surface with a rigid rough indenter (K. L. Johnson, 1987). Rough contact models pursue the linear relation between real contact area and load, based on the purely elastic deformation—in opposition to the model of Bowden and Tabor. It should be noted, though, that there are no experimental evidences supporting linearity between friction and real contact area for all ranges of applied load and contact area fraction (Paggi and Ciavarella, 2010). For example, experimental results suggest that linearity is lost for very large normal loads and very small slidings. Rabinowicz (1965) compiles a large collection of experimental data, showing the influence of several parameters on the coefficient of friction, which also supports the nonexistence of linearity for very wide ranges of nominal pressure. Curiously enough, proportionality between area and load holds even if adhesion is present (Persson, Sivebaek, *et al.*, 2008).

The contact between rough surfaces is inherently a three dimensional problem. Thus the vast majority of analytical models available are formulated in 3D, indeed. However, numerical simulations of such problems, e.g. by using the Finite Element Method, are computationally expensive. As a workaround, 2D simulations provide a cheaper and faster alternative to model rough contact, at the cost of some physical significance of the results. This is, the contact of rough profiles is less representative of practical application, yet it serves as a framework from which qualitative information can be obtained quickly.

This chapter reports analytical models for the elastic, frictionless, nonadhesive, rough contact of surfaces and profiles. Several models available in literature are reviewed, and their most important aspects are documented. Multiasperity models and Persson's diffusive contact theory, the two most popular rough contact theories, are discussed in more detail, and analytical expressions relating real contact area with load are presented. Despite the large number of models provided in the literature, a physically precise explanation of Amontons' laws of friction is yet to be reported (Carbone and Bottiglione, 2008; Persson, Sivebaek, *et al.*, 2008).

4.1 Bibliographic survey

The linearity between real contact area and load was initially explained by Bowden and Tabor, based on the hypothesis of plastic deformation of roughness summits. Nonetheless, it proves unrealistic to consider that a real surface, which completes several contact cycles during its life, will withstand the load with plastic deformation in every cycle, without suffering excessive damage. Archard, J. F. (1957) proved that proportionality could be obtained from elastic deformation, by using, what would be called now, a multiscale or fractal theory—long before fractals were introduced in mathematics. In Archard's model, roughness was idealized as protuberances (spheres), covered with micro-protuberances, which in turn were covered with micro-micro-protuberances, and so on (cf. Figure 4.2). Archard did not describe how the quantities required by his model could be obtained from profilometers traces. This compromised the success of his theory, when compared with others that relied on experimental data, effectively.

4.1.1 Review of multiasperity models

The first widely accepted micromechanical contact model was proposed by Greenwood J. A. and Williamson J. B. P. (1966). Their seminal work originated, throughout the years, the class of multiasperity contact theories. Greenwood and Williamson model (denoted GW from here on) limits the contact to a set of geometrical entities, called asperities (summits), under the hypothesis that all asperities are spherical and share the same radius of curvature. By doing so, and modeling the contact between each asperity and a flat plane with Hertz (1882) theory, the real contact area and load can be predicted as functions of

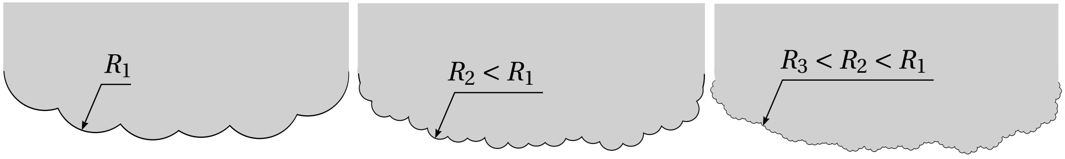


Figure 4.2: Archard's multi-scale roughness model: in each iteration, spheres with progressively smaller radius of curvature are padded to the previous ones. This a fractal concept, since with increasing magnification, more and more detail on the surface is revealed, even though fractals were only introduced many years after Archard published this model. Adapted from Archard, J. F. (1957)

the separation between the surface and the reference plane. From this, area and load can be related indirectly by the separation. GW model became popular amongst tribologists, since it relies exclusively on quantities that can be computed from profilometer traces—the most reliable roughness data available at that time. By predicting that the area of each contact spot increases with the load, but so does the number of contact spots, such that the average contact spot area remains constant, this model provided a relatively simple justification for linearity between contact area and load. In fact, if the number of contacts is constant, the area will increase with load as $F^{2/3}$, while for increasing number of contact, linearity can be achieved (K. L. Johnson, 1987; Greenwood J. A. and Williamson J. B. P., 1966). GW theory was improved by McCool (1986), who referred to the random theory of Nayak (1971) and the results from Bush, Gibson, and Keogh (1976), both on isotropic Gaussian surfaces, in order to compute more accurate values for the asperity radius, standard deviation of summit heights and density of asperities (this improved GW model shall be referred as GW-McCool).

The most complete multiasperity theory was developed by Bush, Gibson, and Thomas (1975) (abbreviated BGT), which accounted for the variation of asperity principal curvatures in each direction with height, by recalling the work of Nayak (1971) and Longuet-Higgins (1957b,a), as well. Summits are modeled as elliptic paraboloids, and the probability of having a certain combination of principal curvatures changes with asperity height. For this purpose, BGT model requires the joint probability distribution of summit heights and principal curvatures, or radius or curvature. J. A. Greenwood (2006) proposed a simplification of BGT theory, where asperities were treated as having spherical caps, with curvature equal to the square root of the product of principal curvatures (referred as GW-SE in the present document, denoting Greenwood-Williamson Simplified Elliptic model). An approach similar to GW-SE considers the arithmetic mean of principal curvatures, instead.[2] The last two models reduce the mathematical complexity inherent to BGT, by eliminating one degree of freedom of the problem, while trying to conserve its accuracy.

Asymptotic limit for small nominal pressure

All multiasperity contact models rely on Hertz contact theory, hence they must verify its fundamental hypothesis. In particular, the consideration that the contact area is small compared with the radius of curvature of contacting asperities (K. L. Johnson, 1987). These models do not include interaction between contacts, i.e., the deflection of neighboring asperities due to the compression of a certain summit—this effect is illustrated in Figure 4.3. Additionally, coalescence of contact areas is not modeled, neither. For these reasons, multiasperity theories are expected to verify only in infinitesimal contact, at very low loads and contact areas. The asymptotic limit of multiasperity models, for vanishingly small loads, reads[3]

$$A_c = \varkappa \frac{F}{E^* \sqrt{\overline{\|\nabla z(x,y)\|^2}}} \,, \tag{4.1}$$

[2] The origins of this last model are unclear (Carbone and Bottiglione, 2008), reporting to the work of Thomas (1999), and is commonly called the Nayak-Thomas model.

[3] This relation, derived from analytical grounds can also be predicted from dimensional analysis, as in Prodanov *et al.* (2014).

which can also be written as

$$\frac{A_c}{A} = \Bbbk \frac{p_0}{E^* \sqrt{\overline{\|\nabla z(x,y)\|^2}}} . \tag{4.2}$$

The symbols used in Equations (4.1) and (4.2) denote

A	Nominal contact area
A_c	Real contact area
F	Applied load
p	Nominal external pressure
E^*	Effective Young modulus, given by $\frac{1}{E^*} = \frac{1-\nu_1^2}{E_1} + \frac{1-\nu_2^2}{E_2}$
$\sqrt{\overline{\|\nabla z\|^2}}$	Surface RMS slope
$\Bbbk$	Linearity coefficient

For multiasperity theories accounting for the variation of asperity curvature with height, it can be proved that the coefficient $\Bbbk$ is $\sqrt{2\pi}$ (Bush, Gibson, and Thomas, 1975; Carbone and Bottiglione, 2008). Thus, every multiasperity theory cited so far verifies this asymptotic relation, with the exception of GW and GW-McCool. However, the linear limit of multiasperity theories is reached for unrealistic values of contact area. Divergence occurs for fractions of real contact area smaller than 0.0001, and separations about six times the RMS roughness, where the existence of contact is questionable. Furthermore, convergence to the linear relationship is very sensitive to the spectrum breadth α. In particular, for large values of spectrum breadth α, which are representative of several real surfaces, the linear asymptote is verified for decreasingly smaller and unrealistic values of contact area (Carbone and Bottiglione, 2008; Paggi and Ciavarella, 2010).

Remark 4.1 on the non-dimensionalization of load.
Some authors opt to use a different form for the normalized external pressure, in Equation (4.2). *In those cases, the parameter* $\sqrt{m_2/\pi}$ *substitutes the RMS slope, and the non-dimensionalization uses the parameter* $\Omega = \sqrt{m_2/\pi}E^*$ *to write the dimensionless pressure as* $F/(\Omega A)$. *Naturally, this changes the coefficient of linearity between the real contact area fraction and dimensionless load. Here, the version with the RMS slope will be preferred, whenever possible.*

Inclusion of asperity interaction and coalescence

Many different strategies were pursued in order to incorporate, in multiasperity models, the interaction between the displacement field produced by each contact. Ciavarella, J. A. Greenwood, *et al.* (2008) introduced a zeroth-order interaction on the GW model, by applying a uniform deformation of the subtract, due to a uniformly distributed contact pressure. This is equivalent to an increase in the effective separation between surfaces. Paggi and Ciavarella (2010) followed the same strategy for the BGT model. Ciavarella, Delfine, *et al.* (2006) included first-order interaction on a discrete version of the GW model, and Paggi and Ciavarella (2010) simplified the former formulation in order to include only zeroth-order interaction effects. By introducing interaction effects on contact models, their accuracy increases considerably, approaching state-of-the-art results

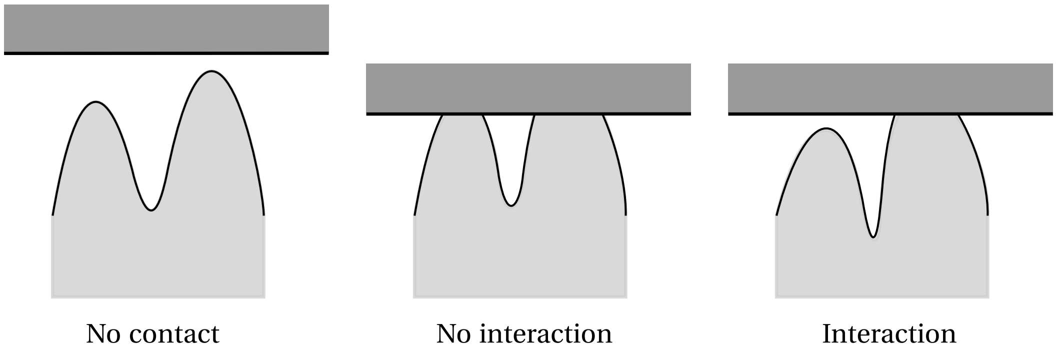


Figure 4.3: Effect of asperity contact interaction on real contact area. When one asperity is compressed by the flat plane, all points in the subtract are displaced, thus moving the neighbor asperity away from the plane. If this effect is included, it leads to a decrease in contact area relative to a model without interaction, for the same applied load.

(Paggi and Ciavarella, 2010). The problem of coalescing contact areas was addressed by Afferrante *et al.* (2012), by replacing two coalescing asperities with a single one with equivalent properties. By considering both interaction and coalescence, the former method extended the linear regime for real contact area fractions up to 20%.

Two dimensional contact model

A two dimensional GW model (denoted here by GW-2D) was proposed by J. A. Greenwood *et al.* (2011). It restricts the original 3D model to the contact of rough profiles, whose peaks are modeled as circles having constant radius of curvature. A 2D contact model raises a problem related with the line contact of a cylinder with an elastic half-space, where the penetration is dependent on the *thickness* of the half-space. Actually, the relative displacement between the center of the cylinder and a point inside the half-space depends on how deep the point is located inside the half-space. This results from the fact that penetration, in line contact, cannot be determined uniquely from the local contact stresses, requiring the prescription of the stress distribution on the bulk (K. L. Johnson, 1987). Owing to this theoretical detail, the formulation of GW-2D lacks the cleanliness of its 3D relative, but may provide a simple tool for comparison with numerical results, despite the inconvenient dependence on half-space thickness.

Remarks and models comparison

Among all multiasperity models, GW is arguably the least accurate, yet it also is the most flexible for particular applications, once it requires only simple statistical measures of the surface topography. In contrary, BGT and GW-SE rely heavily on the assumption of Gaussian isotropic surfaces, for which a random theory is fully developed (Nayak, 1971; Longuet-Higgins, 1957b,a). However, the contact area predicted by GW-McCool and BGT do not differ considerably, which suggests that GW-McCool is the most effective theory, even though it does not verify the same asymptotic limit (Carbone and Bottiglione, 2008).

Multiasperity models still predict asymptotic linearity between area and load, even for asperity heights distribution different from those associated with Gaussian surfaces. In particular, Ciavarella (2016) tested GW-McCool theory with several Weibull distributions, for different parameters, finding a linear relation between area and load for small contact areas. The resulting linearity coefficient was near to that predicted by Gaussian surfaces. Paggi and Ciavarella (2010) also verified that small deviations from the Gaussian distribution of heights do not correlate with errors associated with the models.

With the rise of fractal characterization of roughness, multiasperity models started being questioned, since the definition of asperity seemed no longer obvious. For a long time, asperities were identified in profile and surface measurements by the three and five points rules (Greenwood J. A. and Williamson J. B. P., 1966). In a profile trace, a point which is higher than its two neighbors was defined as a peak. The same was followed for summits in area measurements, and the five points rule. From the moment that one considers that an asperity is, in fact, covered with several smaller asperities, this definition turns unclear. J. Greenwood and J. Wu (2001) published *an apology* concerning this issue, stating that *asperities* can no longer be defined as peaks or summits, but as *what makes contact* (cf. Figure 4.4). In spite of the criticism, multiasperity models prevail in the scientific literature due to their overall simplicity, and to the fact that they capture the qualitative behavior of micromechanical contact— proportionality between real contact area and load and approximately negative exponential dependence of pressure on separation (Persson, 2007; Ciavarella, 2016).

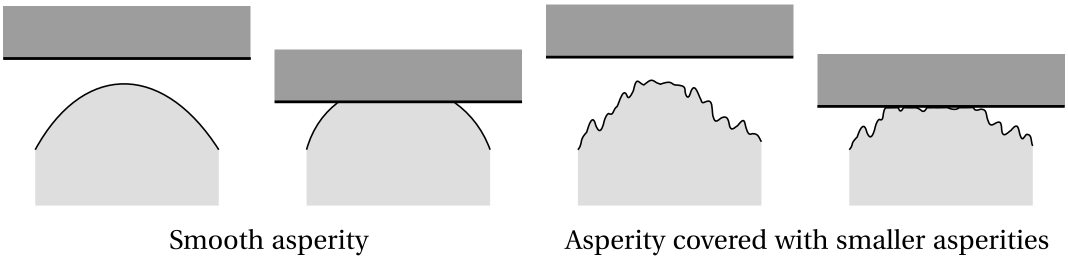


Figure 4.4: Comparison between a smooth asperity, as idealized by GW model, and a fractal asperity. If surface summits were smooth, there would be no conflict in the definition of asperity. However, due to the self-affine nature of rough surfaces, summits will contain smaller microscopical summits. Comparing both cases, the smaller asperities will only be relevant for contact stiffness at high separations. Thus, the definition of asperity is not straightforward once fractal roughness is accepted. Adapted from J. Greenwood and J. Wu (2001).

4.1.2 Fractal models

Several fractal models have been proposed in the literature, aiming at modeling the contribution of the multiscale roughness features. Many of fractal models make use of the Weierstrass-Mandelbrot function to simulate fractal roughness (Majumdar and Tien, 1990). A list of fractal and multiscale contact theories is provided in Jackson and Green (2011), and a comparison between multiasperity and fractal approaches is presented by

Kogut and Jackson (2005). The global popularity of fractal contact models, compared with other famous developments, is quite low, and they shall not be addressed in this work.

4.1.3 Persson's model

Persson (2001a,b) proposed a novel micromechanical contact theory, which intrinsically models the multiscale roughness characteristics and does not rely on Hertz contact theory. Instead of considering topographical features, Persson worked with the contact stress probability distribution, and how it changed as new frequency components are added to the surface. In other words, and following the nomenclature adopted in the original work, Persson evaluates the probability distribution of contact pressure at different magnifications.

For the case of two nominally flat surfaces in contact, when no frequencies are present, the only value that contact stress can take is the nominal pressure p_0. By successively adding shorter wavelengths to the topography, contact stress can take values in an increasingly wider interval, since points in the surface will be compressed by different amounts. Thus, while in the initial magnification there was only one value for the contact stress, and the probability distribution would reduce to a Dirac delta function centered on p_0, by increasing the magnification of observation, contact stress are allowed to take different values and the probability distribution broadens. This is, the probability distribution of contact stress *diffuses* with magnification (see Figure 4.5, in page 97, for a graphical representation). In fact, this is precisely the result obtained by Persson: a diffusive partial differential equation. It is derived assuming full contact between an elastic flat surface and a rigid rough one, which implies that the power spectrum of both the deformed and rigid surfaces are equal. Partial contact is imposed by a boundary condition, specifying that traction cannot exist—nonadhesive contact. From Persson's model, the result for the contact area simply comes

$$\frac{A_c}{A} = \operatorname{erf}\left(\frac{\sqrt{2} p_0}{E^* \sqrt{\overline{\|\nabla z(x,y)\|^2}}} \right), \tag{4.3}$$

where $\operatorname{erf}(\cdot)$ is the error function, defined as

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-a^2} \, \mathrm{d}a \,. \tag{4.4}$$

For small values of the nominal pressure, the asymptotic limit of Equation (4.3) is

$$A_c = \sqrt{\frac{8}{\pi}} \frac{F}{E^* \sqrt{\overline{\|\nabla z(x,y)\|^2}}} \,. \tag{4.5}$$

Thus, Persson's theory also predicts linearity between load and area, with linearity coefficient $\hbar = \sqrt{8/\pi}$—smaller than the value predicted by multiasperity models.

Comparison with multiasperity models

Persson's theory brings several improvements over asperity-based models. The asymptotic linearity holds approximately true up to realistic values of the real contact area,

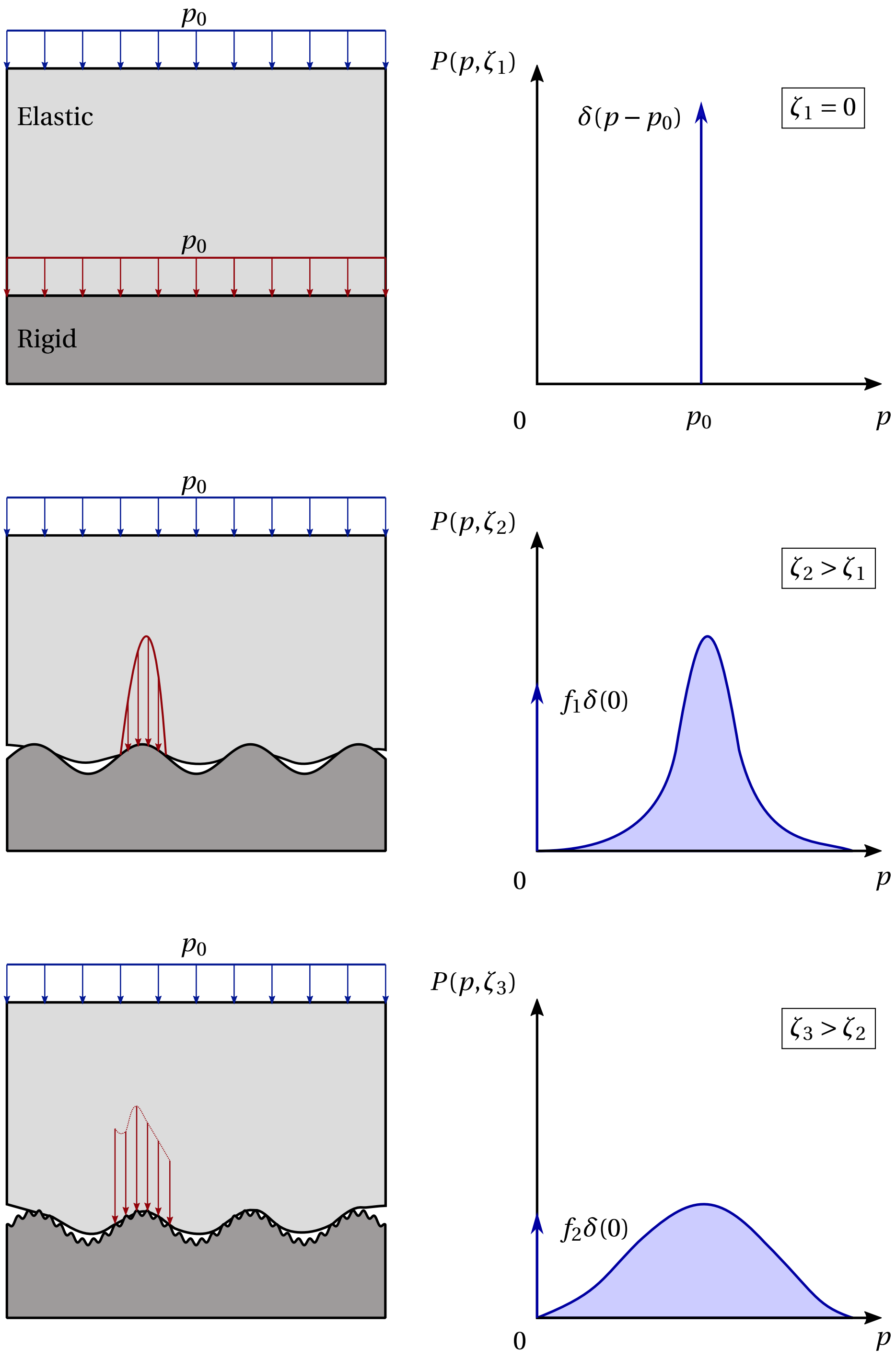


Figure 4.5: Schematics of Persson's contact model. In the initial magnification $\zeta_1 = 0$, only the zero frequency is included, thus contact pressure p is equal to the nominal pressure p_0 at every point, and the probability distribution of contact stresses $P(p,\zeta)$ reduces to a Dirac delta function, centered on nominal pressure. When magnification increases, more frequencies of the spectrum are included in the surface, and the probability distribution of contact stresses spreads over p. Due to partial contact, infinite points have null contact pressure, hence, one makes P go to zero at $p = 0$, and a new Dirac delta is added at this point, multiplied by a factor f_1 (or f_2), such that the integral of P along p is unitary.

around 15-20%. Furthermore, since it does not rely on Hertz theory, a full contact solution is provided. Actually, it starts from the exact solution at full contact, and departs from it, in order to model partial contact. Along with the full contact solution, it also predicts that A_c goes to A with increasing nominal pressure, which was not predicted by multiasperity models. Interaction between different contact zones is naturally handled, and, even though the 3D theory produces better results, the extension for two dimensional contact is straightforward (Carbone, Lorenz, *et al.*, 2009; Carbone, Scaraggi, *et al.*, 2009).

Multiasperity and Persson's model have been extensively compared in literature, c.f. the work by Carbone and Bottiglione (2008), Paggi and Ciavarella (2010), Persson (2006), and Zavarise and Paggi (2007). In general, these works highlight the inherent advantage of Persson's theory over multiasperity models for large values of contact area. In contrast to multiasperity theories, Persson's model does not depend on the spectrum breadth α, but recent numerical results suggest such dependence may exist (Yastrebov, Anciaux, *et al.*, 2017)

Theoretical limitations and criticism

Despite accounting for multiscale roughness and interaction effects, this theory is not exact. In fact, numerical results suggest, repeatedly, that Persson's theory underestimates contact area (Pei *et al.*, 2005; Hyun, Pei, *et al.*, 2004; Yang and Persson, 2008). The theoretical grounds of Persson's model were assessed by Manners and J. A. Greenwood (2006) and Dapp, Prodanov, *et al.* (2014).

Manners and J. A. Greenwood (2006) emphasized the contradiction inherent to the boundary conditions specifying partial contact, within a derivation of differential equations relying on full contact. Thus, these assumptions are contradictory, and it is not easy to understand how do they influence the accuracy and even the validity of Persson's model. Other problem pointed by these authors is the independence of the contact pressure p and the increment of contact pressure $\mathrm{d}p$, assumed in the derivation of the differential equation. While this is correct in full contact, under partial contact conditions these two quantities are not independent, since traction cannot exist. By specifying that p cannot be negative, a relation between $\mathrm{d}p$ and p is created, which renders p and $\mathrm{d}p$ dependent.

Dapp, Prodanov, *et al.* (2014) explored several explicit and implicit assumptions in this theory, and suggested that the accuracy of Persson's model might result form a fruitful cancellation of errors associated with several hypothesis—which might not happen in all situations. Some rely on more mathematical than physical grounds, and there is no interest here in covering all of them. However, it is curious to refer the re-entry effect, neglected by Persson's model. Figure 4.6 shows the contact at different magnifications, as idealized by this theory. One can see that as magnification increases, the global deflection is not much affected, but the local topography changes considerably. By adding more and more length scales, some points might fall out of contact at a certain magnification, and re-enter contact in a higher magnification. This effect is not predicted by Persson's model, and contributes for the underestimation of contact area.

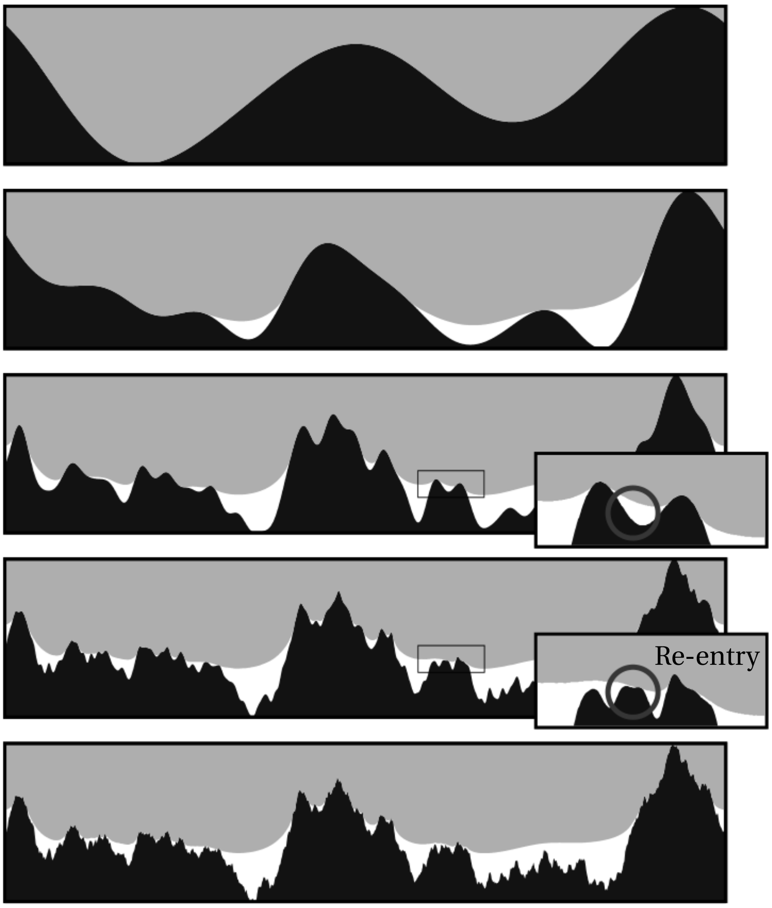


Figure 4.6: Contact between a rigid rough subtract (blue) and an elastic, initially flat surface (orange), at different magnification, illustrating the re-entry effect. The gap between the two surfaces is colored in white. When the magnification is increased, some points which have fallen out of contact in a previous magnification may re-entry contact, which is not modeled by Persson's contact theory. Adapted from Dapp, Prodanov, *et al.* (2014).

4.2 Multiasperity contact models

The following section covers a brief mathematical treatment of multiasperity models. First, GW and GW-McCool are formulated. They illustrate, concisely, the underlying physics of asperity-based theories, and do not require complex tools to express the results. Next, the BGT model is presented, followed by GW-SE. A complete derivation of the equations will only be provided for GW and GW-McCool models, because the analytical and algebraic procedure required for more complex theories becomes cumbersome, and strays from the main purpose of this work.

4.2.1 Greenwood-Williamson model

Consider an arbitrary surface $z(x, y)$, whose summits heights z_s satisfy a probability density function denoted by $\varphi_{\text{sum}}(z_s)$. Additionally, assume that all summits are spherical, with radius of curvature $R = -1/\kappa$.[4] When a summit at height z_s is compressed by δ, it originates a circular contact area with radius $a = \sqrt{\delta R}$ (K. L. Johnson, 1987). By the definition of PDF, the probability of summit having a particular value of height z_s is

$$\Pr\left(\mathcal{Z}_s \in [z_s,\ z_s + \mathrm{d}z_s]\right) = \varphi_{\text{sum}}(z_s)\mathrm{d}z_s\,, \tag{4.6}$$

Then, if there are N_{sum} summits in a surface with nominal area A, the number of summits with height z_s is $N_{\text{sum}}\varphi_{\text{sum}}(z_s)\mathrm{d}z_s$. The contact area associated with summits of height z_s equals the area of a single contact times the number of contacts

$$\mathrm{d}A_c = \pi a^2 N_{\text{sum}}\varphi_{\text{sum}}(z_s)\mathrm{d}z_s = \pi\delta R N_{\text{sum}}\varphi_{\text{sum}}(z_s)\mathrm{d}z_s\,. \tag{4.7}$$

Assuming that the rough surface is in contact with a flat plane, at a distance d from the surface's reference plane, i.e., the plane from which z_s is measured, one can relate summit height and separation d with the penetration δ by

$$\delta = z_s - d\,. \tag{4.8}$$

Since contact occurs only at summits higher that d, the total contact area can be computed by integrating Equation (4.7) from d to infinity,

$$A(d) = \int_d^{\infty} \pi R N_{\text{sum}}(z_s - d)\varphi_{\text{sum}}(z_s)\mathrm{d}z_s\,. \tag{4.9}$$

At this stage, it proves convenient to rewrite Equation (4.9) in terms of the real contact area fraction. Therefore, by introducing the density of summits per unit area $\mathcal{D}_{\text{sum}} = N_{\text{sum}}/A$, and taking the constants out of the integral, one can write that

$$\frac{A_c(d)}{A} = \pi R\mathcal{D}_{\text{sum}} \int_d^{\infty} (z_s - d)\varphi_{\text{sum}}(z_s)\mathrm{d}z_s\,. \tag{4.10}$$

Additionally, denoting $\hat{z}_s = z_s/\sigma_s$ and $\hat{d} = d/\sigma_s$ as the summit height and separation non-dimensionalized by the standard deviation of summit heights σ_s, and referring to $\hat{\varphi}_{\text{sum}}(\hat{z}_s) = \sigma_s\varphi_{\text{sum}}(\hat{z}_s\sigma_s)$ as the PDF of the dimensionless summit heights, Equation (4.10) becomes

$$\frac{A_c(\hat{d})}{A} = \pi\sigma_s R\mathcal{D}_{\text{sum}} \int_{\hat{d}}^{\infty} (\hat{z}_s - \hat{d})\hat{\varphi}_{\text{sum}}(\hat{z}_s)\mathrm{d}\hat{z}_s\,. \tag{4.11}$$

A similar reasoning can be applied to nominal contact pressure p_0 and load F. The mean contact pressure p_m of a circular contact is (K. L. Johnson, 1987)

$$p_m = \frac{4E^*}{3\pi}\sqrt{\frac{\delta}{R}}\,. \tag{4.12}$$

[4] Summits have negative curvature, by definition.

The infinitesimal load supported by peaks of height z_s equals the mean contact pressure, multiplied by the infinitesimal contact area related with those asperities,

$$\mathrm{d}F = \frac{4}{3}\delta^{3/2} R^{1/2} E^* N_{\mathrm{sum}} \varphi_{\mathrm{sum}}(z_s)\mathrm{d}z_s \,. \tag{4.13}$$

Again, integrating from d to infinity, introducing the dimensionless variables and dividing both members by the nominal contact area, it comes

$$p_0(\hat{d}) = \frac{F(\hat{d})}{A} = \frac{4}{3} R^{1/2} \sigma_s^{3/2} E^* \mathcal{D}_{\mathrm{sum}} \int_{\hat{d}}^{\infty} \left(\hat{z}_s - \hat{d}\right)^{3/2} \hat{\varphi}_{\mathrm{sum}}(\hat{z}_s)\mathrm{d}\hat{z}_s \,. \tag{4.14}$$

Equations (4.11) and (4.14) relate the real contact area fraction and nominal pressure with the dimensionless separation $\hat{d}$ between the rough surface and a flat plane. Thus, in general, the real contact area and load can be related indirectly, by this variable. The required inputs for this theory are the asperities radius of curvature R, effective Young modulus E^*, density of summits $\mathcal{D}_{\mathrm{sum}}$ and the probability distribution of surface summits φ_{sum}. The input φ_{sum} can be divided into two contributions, being the standard deviation of summit heights σ_s and the PDF of dimensionless summit heights $\hat{\varphi}_{\mathrm{sum}}$. Greenwood J. A. and Williamson J. B. P. (1966) considered that summit heights were normally distributed, following

$$\hat{\varphi}_{\mathrm{sum}}(\hat{z}_s) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\hat{z}_s^2}{2}\right) . \tag{4.15}$$

The radius of curvature R was computed as the inverse of the profile RMS curvature, the summit heights standard deviation σ_s could approximated by the profile RMS roughness, and the density of summits $\mathcal{D}_{\mathrm{sum}}$ was estimated from the density of profile peaks.

Remark 4.2 on the reference for measuring separation.
Equation (4.15) *implies that the average summit height is zero, i.e., the average summit height is the same than mean surface height, which is also zero, by definition. However, mean summit height does not necessarily equal mean surface height. Then, with Equation* (4.15) *it is actually being considered that separation d is measured relative to summit mean height, and not relative to surface mean height. This issue is paramount in comparing with other models, which measure separation from the surface mean plane.*

McCool's incorporation of spectral properties

Holding the hypothesis of the summit heights distribution being Gaussian, and by recalling the work of Nayak (1971) and Bush, Gibson, and Keogh (1976) on isotropic Gaussian surfaces, McCool (1986) proposed a new strategy for computing some of the previous quantities. He suggested that the asperities radius R, standard deviation of summit heights σ_s and density of summits $\mathcal{D}_{\mathrm{sum}}$ could be estimated by profile spectral moments

m_n and power spectrum breadth α as

$$\frac{1}{R} = \frac{8}{3}\sqrt{\frac{m_4}{\pi}}\,, \tag{4.16}$$

$$\sigma_s^2 = \left(1 - \frac{0.8968}{\alpha}\right) m_0\,, \tag{4.17}$$

$$\mathcal{D}_{\text{sum}} = \frac{1}{6\pi\sqrt{3}}\frac{m_4}{m_2}\,. \tag{4.18}$$

From these results, GW-McCool model comes

$$\frac{A_c(\hat{d})}{A} = \frac{1}{48}\sqrt{3\pi}(\alpha - 0.8968)^{1/2}\int_{\hat{d}}^{\infty}(\hat{z}_s - \hat{d})\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{\hat{z}_s^2}{2}\right)\mathrm{d}\hat{z}_s\,; \tag{4.19}$$

$$\frac{F(\hat{d})}{AE^*\sqrt{m_2}} = \frac{4}{\pi^{3/4}18\sqrt{8}}(\alpha - 0.8968)^{3/4}\int_{\hat{d}}^{\infty}(\hat{z}_s - \hat{d})^{3/2}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{\hat{z}_s^2}{2}\right)\mathrm{d}\hat{z}_s\,. \tag{4.20}$$

For a isotropic surface, the profile spectral moment m_2 equals the surface RMS slope divided by $\sqrt{2}$. Hence one can readily rewrite Equation (4.20) with RMS slope, instead. For anisotropic surfaces, this theory can still be applied, by using Expressions (2.39). The results predicted by GW-McCool, which inherently assume that surfaces are Gaussian and isotropic, are plotted in Figure 4.7. In particular, the area-load curve, and load versus separation are presented, for physically relevant ranges of real contact area. These results depend strongly on Nayak's parameter α, and no truly linear relation exists, even though for $\alpha = 100$ the area-load curve seems to approach a line.

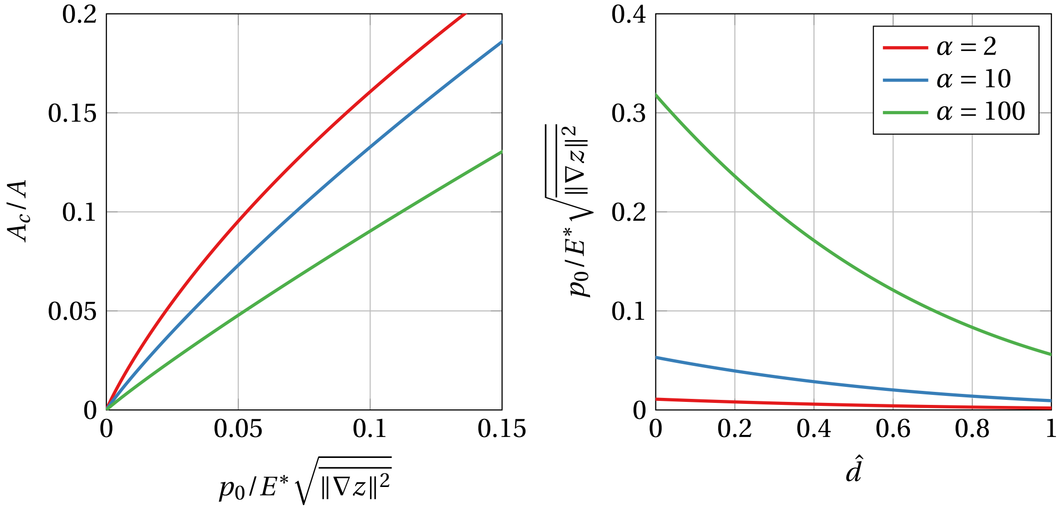


Figure 4.7: Contact area, load and separation from GW-McCool model, for different values of spectrum breadth α. The curves change considerably with varying α, evidencing an inconvenient feature of multiasperity theories.

Nayak (1971) showed that the summit heights distribution for a Gaussian isotropic surface depends on the spectrum breadth $\alpha = m_0 m_4/m^2$, and it only approaches a Gaussian curve for very large values of α. The general case follows[5]

$$\begin{aligned}\hat{\varphi}_{\mathrm{sum}}(t = z_s/\sigma_z) = & \frac{3}{2\pi}\frac{\sqrt{2\alpha-3}}{\alpha}\,t\exp(-C_1 t^2)+ \\ & + \frac{3^{3/2}}{2\alpha\sqrt{2\pi}}(t^2-1)\exp(-t^2/2)\left(1+\mathrm{erf}(\beta)\right)+ \,, \\ & + \frac{\sqrt{\alpha}}{\sqrt{2\pi(\alpha-1)}}\exp\left(-\frac{\alpha t^2}{2(\alpha-1)}\right)\left(1+\mathrm{erf}(\gamma)\right) \,,\end{aligned} \tag{4.21}$$

where

$$C_1 = \frac{\alpha}{2\alpha-3}\,, \tag{4.22}$$

$$\beta = t\sqrt{\frac{3}{2(2\alpha-3)}} \tag{4.23}$$

and

$$,\gamma = t\sqrt{\frac{\alpha}{2(\alpha-1)(2\alpha-3)}}\,. \tag{4.24}$$

In the analytical distribution in Equation (4.21), separation is non-dimensionalized by the standard deviation of the surface heights σ_z, such that the dimensionless separation comes $t = \hat{d}/\sigma_z$. In this case, $\hat{d}$ is measured from the surface mean plane, and not from the mean summit height. Hence, $\hat{d}$ and t are not proportional, in general, and one can argue that a more accurate version of GW-McCool model can be formulated, by using Equation (4.21). However, the attractiveness of the GW-McCool model lives, precisely, in its simplicity, and in the abstraction of a complex mathematical formulation.

4.2.2 General formulation of GW models

At this point, it is also interesting to reformulate GW, without applying explicit relations based on contact theories, e.g., between the contact area radius and penetration, from Hertz theory. One can rewrite Equations (4.7) and (4.13) as

$$\mathrm{d}A_c = \pi a^2(\delta) N_{\mathrm{sum}}\varphi_{\mathrm{sum}}(z_s)\mathrm{d}z_s\,, \tag{4.25}$$

$$\mathrm{d}F = \pi a^2(\delta)\, p_m(\delta) N_{\mathrm{sum}}\varphi_{\mathrm{sum}}(z_s)\mathrm{d}z_s\,. \tag{4.26}$$

The integration of this two equations results in

$$\frac{A_c}{A} = \mathcal{D}_{\mathrm{sum}}\int_d^{\infty}\pi a^2(z_s-d)\varphi_{\mathrm{sum}}(z_s)\;\mathrm{d}z_s\,; \tag{4.27}$$

$$\frac{F}{A} = \mathcal{D}_{\mathrm{sum}}\int_d^{\infty}\pi a^2(z_s-d)\,p_m(z_s-d)\varphi_{\mathrm{sum}}(z_s)\;\mathrm{d}z_s\,. \tag{4.28}$$

Equations (4.27) and (4.28) express the results of GW and GW-McCool in a general formulation, without considering elastic deformation and any particular contact theory. This

[5] Apparently, the formula for the probability distribution of summit heights displayed in the original work of Nayak (1971) is misprinted. See J. A. Greenwood (2006) for a presumably correct expression.

can be extremely useful for further extensions of micromechanical theories to more complex situations.

4.2.3 Bush-Gibson-Thomas model

At the time of writing, the most general formulation of multiasperity theories is provided by the BGT model. The assumption of asperities with spherical caps and constant radius of curvature is replaced with a complete description of summit geometry. This model allows different principal curvatures to exist in each summit, i.e., asperities are modeled as elliptic paraboloids, and the distribution of principal curvatures changes with height. Thus, it requires the joint probability distribution of asperity height z_s and principal curvatures in each direction κ_1 and κ_2. This can can be derived from the statistical geometry theory developed by Nayak (1971) and Longuet-Higgins (1957b,a), and it writes

$$\begin{aligned}\Upsilon(z_s,\kappa_1,\kappa_2) = \frac{27}{8\pi}\frac{1}{m_4^2\sqrt{m_0 m_4}}C_1^{1/2}\exp\left[-C_1\left(\frac{z_s}{\sqrt{m_0}}+\frac{3}{2\sqrt{\alpha}}\frac{\kappa_1+\kappa_2}{2\sqrt{m_4}}\right)^2\right]|\kappa_1-\kappa_2| \\ \cdot\kappa_1\kappa_2\exp\left[-\frac{9(\kappa_1+\kappa_2)^2-24\kappa_1\kappa_2}{16m_4}\right],\end{aligned} \tag{4.29}$$

with C_1 given by Equation (4.22). In the original publication of Bush, Gibson, and Thomas (1975), the authors used a rather different form of Equation (4.29), by specifying the joint probability distribution of asperity height and radius of curvature per unit area, instead.[6] Following the path outlined in the derivation of Equations (4.27) and (4.28), similar expressions accounting for the variation of curvature with height can be derived

$$\frac{A_c}{A} = \mathcal{D}_{\text{sum}}\int_d^{\infty}\int_{-\infty}^{0}\int_{-\infty}^{0}\pi ab\Upsilon(z_s,\kappa_1,\kappa_2)\ \mathrm{d}\kappa_1\mathrm{d}\kappa_2\mathrm{d}z_s\,; \tag{4.30}$$

$$\frac{F}{A} = \mathcal{D}_{\text{sum}}\int_d^{\infty}\int_{-\infty}^{0}\int_{-\infty}^{0}\pi abp_m\Upsilon(z_s,\kappa_1,\kappa_2)\ \mathrm{d}\kappa_1\mathrm{d}\kappa_2\mathrm{d}z_s\,. \tag{4.31}$$

The integrals in Equations (4.30) and (4.31) are carried over all summits (negative curvature, in both directions) higher than the separation. Since asperities are now allowed to take ellipsoidal shape, the contact area is also elliptical, with semi-axis a and b. In turn, the dimensions of the elliptical contact area are a function of the curvatures κ_1 and κ_2 and also of the penetration δ. This raises a major complication in this model, since the Hertzian solution for elliptical contact is expressed through implicit relations, which makes the aforementioned integration impractical. By making a change of variables, and after a long analytical process, the authors of BGT model present a solution for the contact area and load that can be computed through numerical integration. The derivation of that solution falls out of the scope of the present work. Appendix B presents a numerical recipe for the computation of real contact area and load, as shown in the original work by Bush, Gibson, and Thomas (1975).[7] This model predicts an asymptotic linear limit, for

[6] The probability distribution of asperity height per unit area can be obtained by multiplying the respective joint probability distribution by the density of summits $\mathcal{D}_{\text{sum}}$ (Nayak, 1971).

[7] Carbone and Bottiglione (2008) identified a misprint in one equation, in the original BGT publication, and presented the correct expression.

very small nominal pressures, that writes

$$\frac{A_c}{A} = \sqrt{2\pi}\frac{p_0}{E^*\sqrt{\overline{\|\nabla z(x,y)\|^2}}}\,. \tag{4.32}$$

Figure 4.8 shows the results for real contact area fraction versus dimensionless load, and dimensionless load as a function of dimensionless separation, for BGT model, as done with GW-McCool. Note that in BGT, the separation d is non-dimensionalized by the standard deviation of surface heights $\sigma_z = \sqrt{m_0}$, resulting in the dimensionless separation t. Since the separation is measured relative to surface mean plane, instead, the direct comparison of results with GW-McCool for the same value of dimensionless separation is not possible (see remark on page 101). Contact area-load curves varies significantly with spectrum breadth α. Namely, with increasing values of α, the curves depart from the asymptotic limit for increasingly smaller and unrealistic values of contact area. Even for small values of α, whose lower bound is 1.5, the linear relation is approximated at very small values of contact area.

It should be remarked that the computer implementation of this model is rather expensive, compared with others, which are practically instantaneous. It requires the numerical integration of one definite double integral, and an indefinite triple integral—the upper boundaries are infinite, see Appendix B. In addition, these integrals need to be evaluated at every value dimensionless separation. In a machine equipped with a quad-core Intel® Core™ i7-7700HQ CPU at 2.8 GHz, and by using the integration routines provided in the numerical library QUADPACK (accessed through `Python`'s library `SciPy`) it takes, on average, between 1 and 2 seconds, *per* value of dimensionless separation, to compute the respective area and load. The convergence of the integrals, and as a consequence, the computation time depends, rather dramatically, on the value of t and α. For example, for very large compressions $t << 0$, the integration process takes longer to finish. This process can be speed up by using a parallelization strategy, where different processes are responsible for computing area and load for individual segments of the dimensionless t vector. In the machine whose specification were referred earlier, which can run 8 processes in parallel, the computation of area and load for 40 values of t, between 0 and 3, took approximately 2 minutes with sequential execution, and this time was reduced to about 30 seconds, with 8 parallel processes.

4.2.4 Greenwood-Williamson simplified elliptic model

Regarding the complexity of multiasperity theories, GW and BGT lie within the two extremes of the complexity spectrum, with GW being the most simple and BGT the most complex. J. A. Greenwood (2006) proposed a contact theory with intermediate intricacy, arguing that the summits of a Gaussian isotropic surface are only mildly ellipsoidal. Thus they can be modeled as spherical, with an equivalent radius of curvature, but keeping the variation of curvature with height. By doing so, the contact area can be considered circular and: on one hand, the solution for Hertzian contact comes simplified; on the other, the problem can be described with only two variables. In the GW-SE model, an ellipsoidal asperity with principal curvatures κ_1 and κ_2 is replaced by a spherical asperity

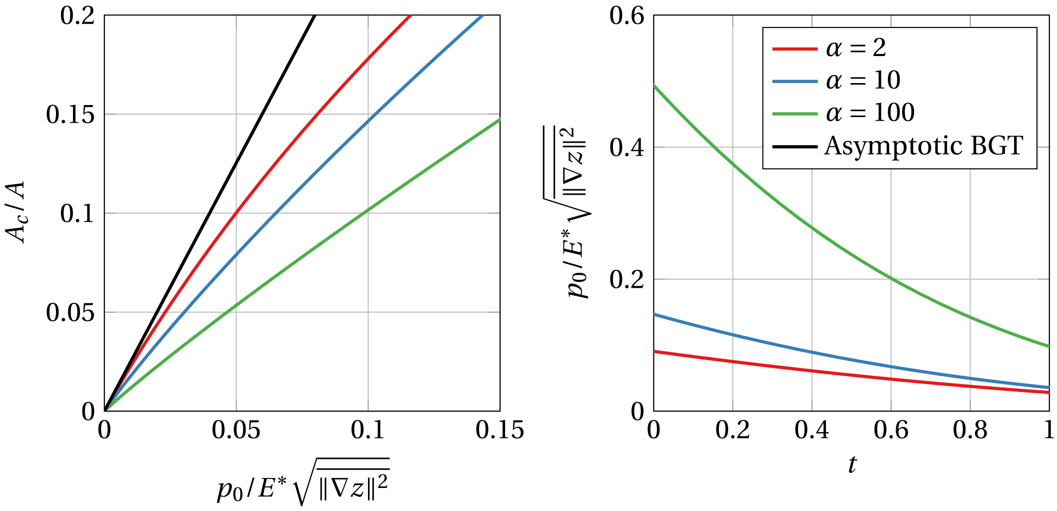


Figure 4.8: Contact area, load and separation from BGT model, for different values of spectrum breadth α. Identically to the results of GW-McCool, BGT predictions for real contact area as a function of load change drastically with increasing α. The higher α, the sooner the curves deviate from the asymptotic linear limit. Note that dimensionless separation t is defined differently from d, since the non-dimensionalization is performed with the standard deviation of heights $\sigma_z = \sqrt{m_0}$, and the reference plane is the surface mean plane instead, hence direct comparison between separation-load of the two models is not feasible.

with curvature $\kappa_G = \sqrt{\kappa_1\kappa_2}$. Without entering in details on the derivation of the theory (see the original work of J. A. Greenwood (2006) for a detailed explanation), and introducing the dimensionless curvature $g = \kappa_G/\sqrt{m_4}$, the contact area and load can be obtained from

$$\frac{A_c(t)}{A} = \frac{3\alpha}{4\sqrt{6\pi(\alpha-1)}}\int_t^\infty\int_0^\infty (\hat{z}_s - t)g^2 f(\hat{z}_s, g)\, \mathrm{d}g\mathrm{d}\hat{z}_s\,; \tag{4.33}$$

$$\frac{F(t)}{AE^*\sqrt{m_2}} = \frac{\alpha^{5/4}}{\pi^{3/2}\sqrt{6(\alpha-1)}}\int_t^\infty\int_0^\infty (\hat{z}_s - t)^{3/2} g^{5/2} f(\hat{z}_s, g)\, \mathrm{d}g\mathrm{d}\hat{z}_s\,. \tag{4.34}$$

The function $f(\hat{z}_s, g)$ is defined by

$$f(\hat{z}_s, g) = \operatorname{erfc}\left[\mu\left(3g - \frac{\hat{z}_s\sqrt{\alpha}}{\alpha-1}\right)\right]\exp\left[\frac{1}{2}\left(3g^2 - \frac{\hat{z}_s^2\alpha}{\alpha-1}\right)\right], \tag{4.35}$$

where erfc$(\cdot)$ denotes the the complementary error function

$$\operatorname{erfc}(x) = \frac{2}{\sqrt{\pi}}\int_x^\infty e^{-a^2}\, \mathrm{d}a \tag{4.36}$$

and μ comes

$$\mu = \sqrt{\frac{\alpha-1}{4\alpha-6}}\,. \tag{4.37}$$

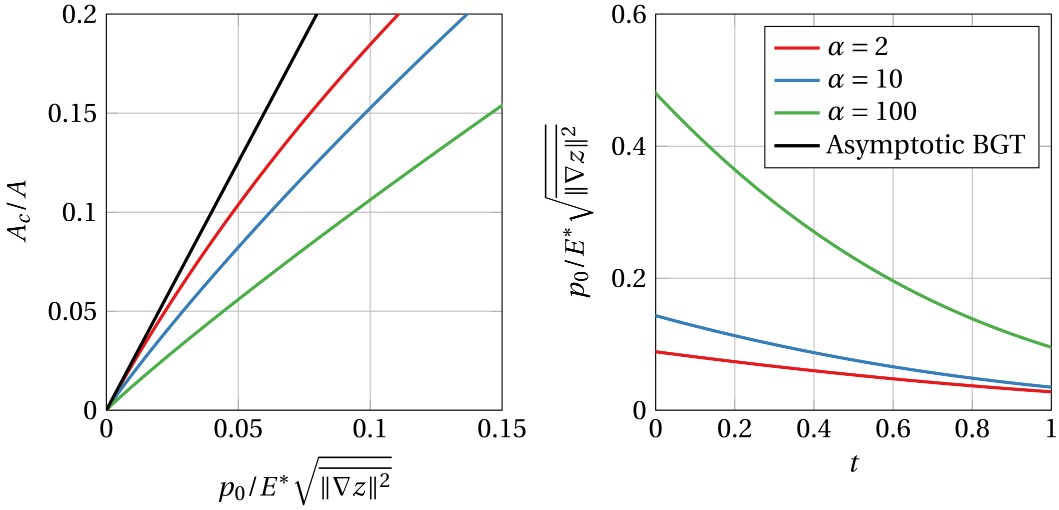


Figure 4.9: Contact area, load and separation from GW-SE model, for different values of spectrum breadth α. Results from this theory match closely the results from BGT, suggesting that it can used as a low cost alternative, for representing multiasperity models.

Following the same structure adopted for the results of GW-McCool and BGT, Figure 4.9 shows the contact area, load and separation predictions from GW-SE. It can be observed that they resemble, closely, the results from BGT, and does also converge for the same asymptotic line, at low loads. This theory requires the numerical computation of only two double integrals, which is accomplished much faster, compared with BGT. After all, it reveals to be a good alternative for representing multiasperity theories reliably and effectively.

4.3 Persson's contact theory

This section is devoted to discussing some theoretical details of Persson's model. Note that in the original work (Persson, 2001a), and even in later presentations of his theory (Persson, Bucher, *et al.*, 2002; Persson, 2001b) the derivation of the diffusive differential equation and other aspects are quite difficult to follow, as already pointed out by other authors (Manners and J. A. Greenwood, 2006). Thus, the following text does not aim at presenting a complete derivation of the theory, but only at clarifying some features instead. For a simplified derivation of the differential equation, the reader is referred to Manners and J. A. Greenwood (2006).

Persson's theory deals with the probability density function of the contact stresses, or contact pressures. Thus, the variance of contact pressures is an extremely useful quantity for its characterization. For this purpose, it is convenient to analyze the pressure distribution which flattens a sinusoidal displacement field. This classical result (K. L. Johnson,

1987) states that a sinusoidal displacement field $w(\boldsymbol{x})$, say in the z direction, given by

$$w(\boldsymbol{x}) = \frac{p_m \lambda}{\pi E^*} \cos(\boldsymbol{k}\cdot\boldsymbol{x}) = \frac{2p_m}{\|\boldsymbol{k}\| E^*} \cos(\boldsymbol{k}\cdot\boldsymbol{x}) \ , \tag{4.38}$$

where $\boldsymbol{x} = (x, y)$, is flattened by a sinusoidal pressure distribution with amplitude p_m, which writes

$$p(\boldsymbol{x}) = p_m \cos(\boldsymbol{k}\cdot\boldsymbol{x}) \ . \tag{4.39}$$

To say that the pressure field given in Equation (4.39) flattens the displacement expressed by Equation (4.38), is equivalent to stating that such pressure field is generated whenever the former displacements are applied. By assuming that the surface contains a certain range of frequencies, from the superposition principle, each infinitesimal contribution,

$$\mathrm{d}w(\boldsymbol{x}) = \frac{4\mathscr{P}(\boldsymbol{k})\mathrm{d}k}{\|\boldsymbol{k}\| E^*} \cos(\boldsymbol{k}\cdot\boldsymbol{x}) \ , \tag{4.40}$$

is flattened by a certain component of the pressure spectrum, with $\mathscr{P}(k)$ denoting the Fourier transform of the pressure distribution. The factor 2 comes from the fact that Fourier transform concerns complex exponentials, while Equation (4.40) is written with a real valued trigonometric function. By integrating Equation (4.40) in the frequency space, i.e., by summing all frequencies, one can relate the Fourier transform of the surface height with the pressure spectrum as

$$\mathscr{F}\{w\} = \frac{2\mathscr{F}\{p\}}{\|\boldsymbol{k}\| E^*} \ . \tag{4.41}$$

Equation (4.41) can be rewritten in terms of the surface power spectrum Φ and also the power spectrum of the pressure field Φ_p as

$$\Phi(\boldsymbol{k}) = \frac{4\Phi_p(\boldsymbol{k})}{\|\boldsymbol{k}\|^2 [E^*]^2} \ . \tag{4.42}$$

Finally, by recalling that the variance σ_z^2 equals the spectral moment m_{00} of the spectrum in question, and that $m_{02} + m_{20}$ equals the variance of the surface slopes, integration Equation (4.42) in the frequency domain yields

$$m_{20} + m_{02} = \sigma_z^2(\|\nabla z\|) = \overline{\|\nabla z\|^2} = \frac{4\sigma_z^2(p)}{[E^*]^2} \ . \tag{4.43}$$

Equation (4.43) relates the surface RMS slope with the variance of contact pressure distribution, in full contact conditions. Note that the surface RMS slope depends on the bandwidth over which the spectrum is integrated. Consider that the surface spectrum is defined between the zero frequency and an upper bound $\zeta\|\boldsymbol{k}_0\|$, with $\|\boldsymbol{k}_0\| = 2\pi/L$. Here, L is the characteristic length of the surface, and ζ is a magnification factor. The magnification ζ is the extent at which the surface spectrum is represented, or in other words, the level of detail in the surface. Then, the variance of pressures σ_p^2 depends on the magnification factor ζ, and comes

$$\sigma_p^2(\zeta) = \frac{[E^*]^2}{4} \overline{\|\nabla z\|^2} \ . \tag{4.44}$$

The proceeding analysis, by Persson, concerns only the variance of contact pressures, and how it changes with increasing magnification. A key point to mention is that if the surface heights is Gaussian, then the contact pressure distribution is also Gaussian, since it results from the superposition of a large number of independent random variables. Denoting $P(p,\zeta)$ as the contact pressure probability distribution at magnification ζ, and assuming that the contact pressure variation $\mathrm{d}p$ is independent of the actual pressure p, the following differential equation holds

$$\frac{\partial P}{\partial \zeta} = G'(\zeta)\, p_0^2\, \frac{\partial^2 P}{\partial p^2}, \tag{4.45}$$

with

$$G(\zeta) = \frac{1}{2}\frac{\sigma_p^2(\zeta)}{p_0^2}\,. \tag{4.46}$$

Equation (4.45) models how $P(p,\zeta)$ evolves as increments of roughness $\mathrm{d}\zeta$ are added to the surface, i.e., as PSD covers wider ranges, while the nominal pressure p_0 remains constant—in the case of full contact. The solution of the diffusive differential equations must verify the boundary conditions:

$$P(p,0) = \delta(p-p_0)\,; \tag{4.47}$$

$$P(\infty,\zeta) = 0\,; \tag{4.48}$$

$$P(0,\zeta) = 0\,. \tag{4.49}$$

The boundary condition Equation (4.47) specifies that when the magnification is null, i.e., when only the zero frequency component is considered, the only value that the contact pressure can take is the nominal pressure (cf. Figure 4.5). Equation (4.48) simply prohibits the existence of infinity large contact stresses, which are physically meaningless. Finally, the boundary condition Equation (4.49) is responsible for modeling partial contact, by imposing that the probability of existing negative contact pressure (traction, since $p > 0$ corresponds to compression) is null. The solution for $P(p,\zeta)$ comes

$$P(p,\zeta) = \frac{1}{\sqrt{2\pi\sigma_p^2(\zeta)}}\left(\exp\left[-\frac{(p-p_0)^2}{2\sigma_p^2(\zeta)}\right] - \exp\left[-\frac{(p+p_0)^2}{2\sigma_p^2(\zeta)}\right]\right). \tag{4.50}$$

The solution provided in Equation (4.50) does not verify unit integral, in partial contact—a necessary condition for a probability density function. In fact, in partial contact, there is a infinite number of points which are not in contact, hence have zero contact pressure. Thus, Equation (4.50) must be complemented by a Dirac delta function at the origin, multiplied by a factor that makes the integral of the PDF unit, as illustrated in Figure 4.5. The fraction of real contact area is necessarily given by the probability of having positive contact pressure

$$\frac{A}{A_c} = \int_{0^+}^{\infty} P(p,\zeta)\,\mathrm{d}p = \mathrm{erf}\left(\frac{p_0}{\sqrt{2\sigma_p^2(\zeta)}}\right), \tag{4.51}$$

and recalling Equation (4.44), one recovers Equation (4.3)

$$\frac{A_c}{A} = \operatorname{erf}\left(\frac{\sqrt{2} p_0}{E^* \sqrt{\overline{\|\nabla z(x,y)\|^2}}} \right). \tag{4.52}$$

The results of Persson's theory for the real contact area and contact pressure probability density function are plotted in Figures 4.10 and 4.11, respectively. From Figure 4.10 it can be seen that Persson's model predictions match closely the linear asymptotic, up to 20% of real contact area fraction. It can also be observed, in the same figure, that as the normalized external pressure increases, the real contact area approaches unit (the full contact solution). When the dimensionless pressure reaches the value around 1.6, the deviation of contact area fraction from 1 is negligible. As for Figure 4.11, the PDF spreading with increasing magnification can readily be seen, together with the different values for the area under each curve, i.e., their integral—this is compensated by the additional Dirca delta at the origin.

Even though the former concepts concerned the three dimensional formulation of Persson's model, the restriction to the contact of rough profiles is straightforward. In fact, the 2D rough contact can be thought as the contact of strongly anisotropic rough surfaces, in which case roughness exists only in one direction. The 2D formulation of Equation (4.52) comes (Carbone, Lorenz, *et al.*, 2009)

$$\frac{A_c}{A} = \operatorname{erf}\left(\frac{\sqrt{2} p_0}{E^* \sqrt{\overline{(\mathrm{d}z/\mathrm{d}x)^2}}} \right). \tag{4.53}$$

4.4 Closing comments

As a final discussion to close this chapter, a brief comparison between the contact area-load predictions for all micromechanical contact theories discussed previously is performed. The results for all the models are presented in Figure 4.12. Persson's asymptotic results is discarded since, for this range of contact area, it is almost indistinguishable from the exact solution. Regarding the multiasperity theories, the results for two different values of spectrum breadth α are plotted.

Persson's multiscale model and the asymptotic BGT are the two major references for linearity between real contact area and load. The two models predict different slopes for the area-load curve, with Persson's predicting the smallest. All multiasperity models are very sensitive to α, and deviate from linearity for unrealistically smaller loads and contact areas as α increases.

Even though it is observed that GW-McCool does not converge to same asymptotic limit, for the physically meaningful range of contact area fractions considered, it predict values very close to other more complex multiasperity models.

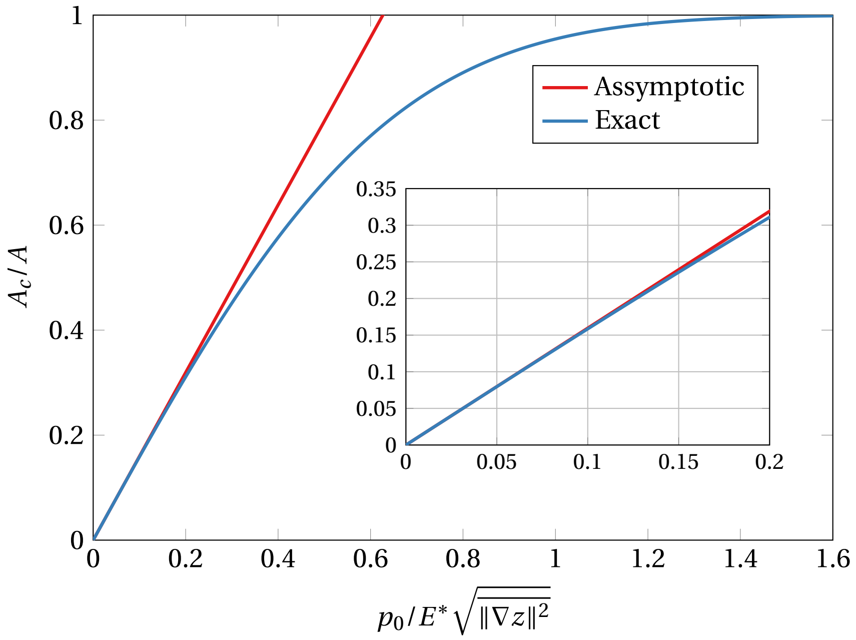


Figure 4.10: Real contact area fraction as a function of dimensionless load predicted by Persson's theory. With increasing nominal pressure, the real contact area fraction goes to 1, as expected for full contact conditions. The linear asymptotic limit for small nominal pressures holds with very small error up to 20% of the nominal contact area.

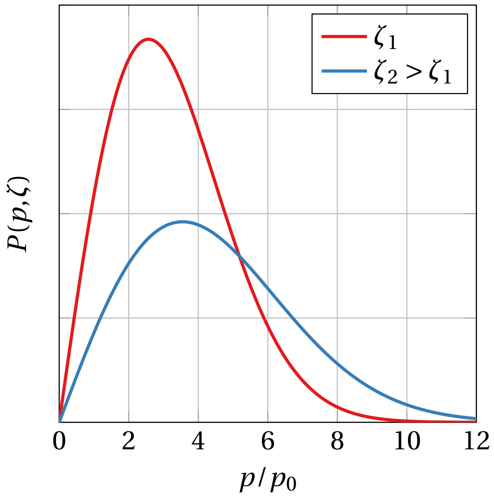


Figure 4.11: Contact pressure probability density function, for two values of magnification, computed from Persson's model. With increasing magnification, the probability distribution spreads over the contact pressure axis. It can be observed that the area below the two curves is not equal. This difference is complemented by adding a Dirac delta function at the origin, multiplied by a coefficient, which depends on magnification.

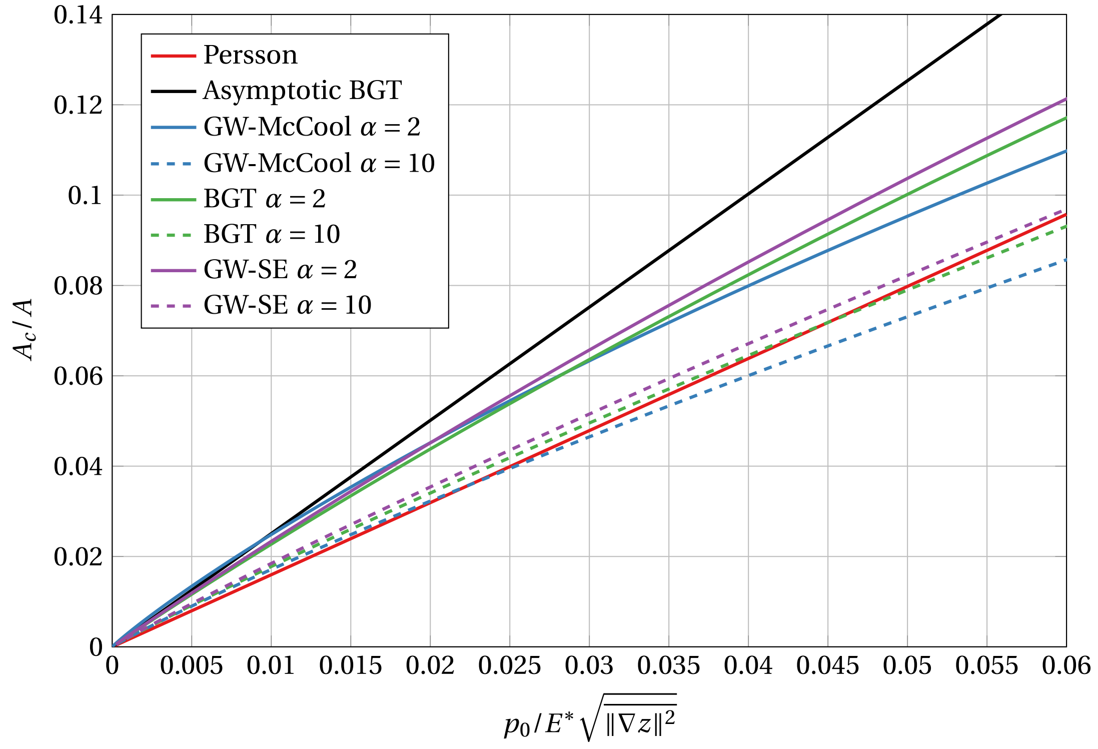


Figure 4.12: Comparison of real contact area fraction *versus* normalized external pressure curves, between different multiasperity contact theories, with different values of spectrum breadth α, and Persson's model. Persson's model and asymptotic BGT provide to reference linear relations. Multiasperity models are sensible to spectrum breadth α, and increasing this value makes multiasperity curves go down, and deviate from the asymptotic limit for smaller loads. All three multiasperity models predict similar real contact area fractions, for the same load, for physically meaningful contact area fractions.

Chapter 5

Single scale dual mortar finite element modeling of rough contact

Analytical modeling of rough contact provides some relatively simple and immediate results for the evolution of contact area with load. However, such approach poses some limitations, even on the solution of micromechanical frictionless and elastic contact problems, since it lacks an ubiquitous answer regarding the accuracy of the currently proposed theories. With the rapid growth of computational resources over the last decades, numerical methods and, in particular, the Finite Element Method (FEM), became increasingly more attractive for the investigation of rough contact. In comparison with analytical theories, these techniques are more flexible and relax the base hypothesis of rough contact formulation. Furthermore, topography realizations can be modeled directly, along with every individual geometrical feature.

Micromechanical contact is intrinsically multiphysical, with numerous phenomena occurring at the contacting interface, see Figure 5.1. While they are strongly coupled in real applications, they can be isolated rather easily by the application of numerical methods, where each individual contribution can be assessed. In fact, in experiments it is difficult to isolate individual effects, yet in the numerical model it is cumbersome, not to say impossible, to account for every possible interaction. Therefore, computational methods and experimental work are, altogether, powerful methods to address the problem of rough contact.

In this chapter, the main numerical approaches to rough contact adopted are reviewed. Special attention is given to the FEM, and to the features on which it shows superior behavior relative to other alternatives. Computational contact algorithms are also subjected to attention, with focus on dual mortar methods. Then, the fundamental theoretical aspects of continuum contact mechanics within the framework of the dual Mortar method are briefly introduced, and the global solution algorithm for the contact problem is presented. Frictionless elastic rough contact finite element simulations within a dual mortar contact algorithm are performed in a single scale framework, in order to establish a statistically *Representative Contact Element* (RCE). All simulations address the 2D contact of a self-affine rough elastic block and a rigid subtract—a *Signorini* contact problem. These preliminary studies are paramount for the future application of multiscale strategies to

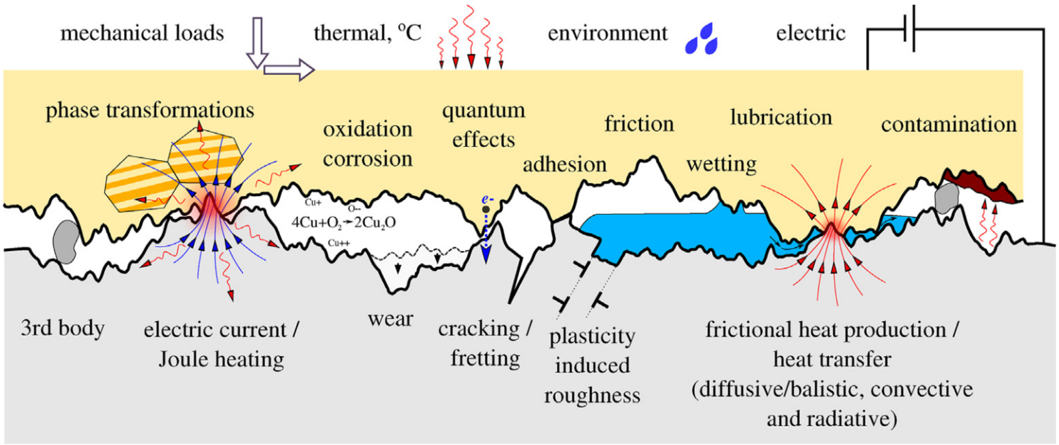


Figure 5.1: Illustration of the multiphysical phenomena involved in the tribological interactions between two rough surfaces in contact. These interactions range from the thermal and electrical behavior of the contact to phase transformations of the material's microstructure. Adapted from Vakis *et al.* (2018).

rough contact, inasmuch that the *micromechanical* problem is well characterized, and the representativeness of the mechanical response is known beforehand.

5.1 State of the art

When it comes to the numerical treatment of rough contact, two topics must be distinguished: the technique for modeling the continuum media and the contact algorithm. These two subjects are covered in the next paragraphs, individually.

5.1.1 Numerical modeling of the continuum

The Finite Element Method and the Boundary Element Method (BEM) are two well established numerical procedures for the solution solid mechanics problems, which have been extensively used in numerical investigations of rough contact. A review of the recent numerical investigations on rough contact is provided in Yastrebov, Anciaux, *et al.* (2015). The interested reader is also referred to Vakis *et al.* (2018) for a more comprehensive review and comparison of numerical methods used to model rough contact.

The application of the BEM requires only the discretization of the surfaces of contacting bodies. The bulk is modeled with fundamental (analytical) solutions, which are only provided for simple situations, e.g., for elastic half-spaces. These can be still applied to rough contact, as long as the RMS slope is small. For this reason, the associated computational cost is drastically reduced, allowing the simulation of rough surfaces with a large number of degrees of freedom. In fact, one the finest meshes used lately in rough contact simulation, containing 4096 elements in each direction (about 50 millions degrees of freedom), was solved with a BEM approach, in Campañá, Müser, and Robbins

(2008). Typically, published works regarding BEM as the modeling technique for rough contact practice a minimum of approximately 1000 elements in each direction. Two major families of BEM strategies commonly employed in micromechanical contact analysis can be distinguished, namely the Green's Function Molecular Dynamics (GFMD) and a FFT based boundary element.

The GFMD technique, initially developed by Campañá and Müser (2006), are boundary element strategies based on molecular dynamics concepts. In a molecular dynamics approach, each point is considered as a particle, which interacts with its neighboring particles through some potential field. This branch of boundary element methods has successfully been applied to the normal contact of two elastic solids by Campañá and Müser (2006), Campañá and Müser (2007), Campañá, Müser, and Robbins (2008), and Prodanov *et al.* (2014) to both normal and transverse loading in Campañá, Persson, *et al.* (2011).

The aforementioned FFT based boundary element methodology, proposed by Stanley and Kato (1997), makes use of the highly efficient FFT algorithms to solve the rough contact problem in the frequency domain. This method has been used recently by Yastrebov, Anciaux, *et al.* (2012, 2015, 2017) in extensive numerical investigations, and by Jackson and Green (2011) for comparison with analytical strategies.

With regard to the FEM, its expression in the general picture of frictionless elastic contact modeling is far less evident, because the BEM provides a cheaper but still reliable solution for this class of problems. The works of Hyun, Pei, *et al.* (2004) and Hyun and Robbins (2007) on elastic contact of self-affine surface are widely known for the application of the FEM, in this context. Yet, their results have been criticized due to the lack of smoothness of the artificial surfaces considered for the simulations (Yastrebov, Anciaux, *et al.*, 2012, 2015). Other noteworthy application of the FEM to elastic rough contact modeling is the work of Yastrebov, Durand, *et al.* (2011), where single asperity simulations and a large scale simulation were performed.

As one moves way from the simple case of normal elastic and frictionless contact, by adding sources of nonlinearity to the model, such as large deformations and nonlinear constitutive laws, FE based approaches dominate every other method in the realm of rough contact, due to its high versatility. It should be emphasized that the much-publicized BEM is not naturally fit to handle such complex problems, then engineers and researchers must necessarily opt to more flexible alternatives. Pei *et al.* (2005) revisited the problem of contact between self-affine surfaces, including plasticity in the numerical model. Also Bandeira, Wriggers, *et al.* (2004) and Bandeira, Pimenta, *et al.* (2008) used 3D elastic and elastoplastic FE models, respectively, to establish contact interface laws. Several studies on rubber friction, which is mainly due to the hysteretic viscoelastic behavior of rubber, and where a large deformation formulation is mandatory, have been carried out with the FEM, e.g., by Reinelt and Wriggers (2010), De Lorenzis and Wriggers (2013), and Wagner, Wriggers, Klapproth, *et al.* (2015). All in all, FE techniques can deal with arbitrarily complex problems, at the cost of more expensive computational resources. As a side note, it should also be mention that commercial FE packages are more abundant than boundary element software.

Remark 5.1 on the application of a FE formulation for the present work.
The current work concerns only the elastic and frictionless contact of self-affine topography, through the application of the finite element method. At this stage, it may appear that a BEM formulation would be more profitable, instead. One of main objectives of this work is the development of a single and multiscale numerical framework for rough contact. This is intended to provide a numerical tool for further investigations, on which more complex phenomena shall be included. Thus, since the present work represents the first step in such development, the simplest case of the elastic contact is a natural starting point.

5.1.2 Contact algorithms

Several computational techniques have been employed in order to incorporate contact in FE frameworks. Initial methodologies by Francavilla and Zienkiewicz (1975) and Hughes *et al.* (1976) modeled contact constraints on a purely nodal approach, and was restricted to meshes with matching nodes, undergoing small deformations. The last two conditions, and specially the requirement of a node-matching meshes, are very restrictive, since the contact interface must be determined as part of the solution, in general.

Node-To-Segment (NTS) contact algorithms are the most widely used discretization techniques in computational contact mechanics, and are widespread in commercial FE codes. The contact constraints are enforced in a point-wise fashion, related to some segment/surface on the opposite boundary, thus allowing for dissimilar meshes to be used. Applications of this algorithm have evolved from simple (Hallquist, 1979) to general contact scenarios (Laursen and Simo, 1993). The NTS algorithm poses some limitations, as demonstrated, e.g., in Papadopoulos and Taylor (1992), which lead to the employment of higher order contact interpolations of the contacting surfaces in order to overcome such difficulties (M. A. Puso and Laursen, 2002).

The so called *Mortar-based contact formulations* have been developed during the last two decades as more robust alternatives to model contact constraints. These were preluded by similar ideas of the *Segment-To-Segment* algorithms, on which the contact interface is partitioned into individual segments for numerical integration, see Simo *et al.* (1985) and Papadopoulos and Taylor (1992). The mortar method, introduced by Bernardi *et al.* (1993), allows for the variationally consistent treatment of contact constraints. This is, the contact constraints enter directly in the weak formulation of the contact problem, and are imposed on an optimal weak sense, by introducing Lagrange multipliers. These must be carefully chosen to preserve the accuracy of the solution. Mortar finite element technologies have been successfully applied, for example, by Belgacem *et al.* (1998) and M. Puso (2004).

In the current state of the art mortar methodology for contact problems, the dual basis for Lagrange multipliers is commonly adopted. On the theoretical foundation of the dual Lagrange multipliers is the *bi-orthogonality* condition, which allows the condensation of some elements of the mortar matrices, without compromising the accuracy (B. Wohlmuth, 2000). This approach has been applied to small deformations, e.g., in Flemisch and B. I. Wohlmuth (2007), and was extended to the general finite deformation realm by Popp, Gee, *et al.* (2009) and Popp, Gitterle, *et al.* (2010). Moreover, dual Lagrange multipliers are naturally fit for the application of Primal-Dual Active Set Strategies (PDASS), which area well established techniques from constrained optimization (Alart and Curnier, 1991). The cornerstone of PDASS is the regularization of the inequality contact constrains by using *Nonlinear Complementarity* (NCP) Functions. These allow the application of Newton-Raphson type algorithms to solve for both the displacement field and the contact interface (active set) within a single loop.

It is noteworthy that several alternative contact discretization techniques exist, apart from the previously discussed ones. In particular, the contact domain method proposed in Oliver *et al.* (2009) and Hartmann *et al.* (2009). This method appeared in the advent of mortar methods, which ended up absorbing all attention from the scientific community.

5.2 Contact modeling with the dual mortar method

In the following sections, the mathematical formulation of continuum contact mechanics problems is presented, followed by the FE discretization, within the framework of the dual mortar method. The following description does not claim to be exhaustive, and shall only go through fundamental concepts and equations. For the sake of completeness, a list of references is provided for each topic discussed here, where more comprehensive introductions to the subjects can be found:

- Nonlinear continuum mechanics - Holzapfel (2000) and Bonet and Wood (2008);
- Contact mechanics - Wriggers (2006) and Wriggers and Laursen (2008);
- Finite element method - Zienkiewicz *et al.* (2000a,b);
- Dual mortar methods for contact problems - Popp (2012) and Pinto Carvalho (2018).

5.2.1 Continuum mechanics and governing equations

The general continuum mechanics framework for deformable bodies departs from the geometrical description of motion and deformation kinematics, which are illustrated in Figure 5.2. The following analysis considers the classical continuum description for every configuration. Additionally, and without loss of generality, all configurations are considered to share the same Cartesian coordinate system. The Lipschitz open set $\Omega_0 \subset \mathbb{R}^d$ denotes the deformable body in the reference configuration, and some point $\mathrm{P} \in \Omega_0$ is referred by the position vector $\boldsymbol{X}$. The symbol d represents the number of spatial dimensions of the problem—for the cases of interest in this work $d = 2, 3$. Each point $\boldsymbol{X}$ is mapped from the reference configuration Ω_0 to the current configuration Ω_t by a bijective nonlinear deformation map φ at each time instant t, which writes

$$\boldsymbol{x} = \varphi(\boldsymbol{X}, t). \tag{5.1}$$

The displacement of the material point tracked by point P is

$$\boldsymbol{u}(\boldsymbol{X}, t) = \boldsymbol{x}(\boldsymbol{X}, t) - \boldsymbol{X}(t)\,. \tag{5.2}$$

The function φ is a one-to-one mapping of material points between the reference and current configuration, thus not allowing the superposition of material points and opening of gap within the material. The independent variable of the formulation is the position in the reference configuration $\boldsymbol{X}$, which is known *a priori*, and $\boldsymbol{x}$ is treated as the dependent variable through the displacement field $\boldsymbol{u}$. The point $\boldsymbol{X}$ is associated with a *Lagrangian description*, since it tracks an individual material point, while $\boldsymbol{x}$ relates to an *Eulerian description*, on which a specific fixed point in space is monitored.

The boundary of the deformable body in the reference configuration is denoted by $\partial\Omega_0$, and is divided in two open disjoint subsets, namely the Neumann partition Γ_σ and the Dirichlet partition Γ_u. At these regions, stresses and displacements are prescribed as boundary conditions, respectively. The disjointness property writes

$$\Gamma_\sigma \cup \Gamma_u = \partial\Omega_0\,, \tag{5.3a}$$

$$\Gamma_\sigma \cap \Gamma_u = \emptyset\,. \tag{5.3b}$$

The counterparts of the Neumann and Dirichlet partition in the current configuration are denoted by γ_σ and γ_u.

The deformation gradient, denoted by $\boldsymbol{F}$, is a fundamental measure of deformation and strain of the body. It is a second order two-point tensor, defined as the partial derivative of the current configuration position $\boldsymbol{x}$ to the relative quantity in the reference configuration

$$\boldsymbol{F} = \frac{\partial \boldsymbol{x}(\boldsymbol{X},t)}{\partial \boldsymbol{X}} = \boldsymbol{I} + \frac{\partial \boldsymbol{u}(\boldsymbol{x},t)}{\partial \boldsymbol{X}}. \tag{5.4}$$

Here, $\boldsymbol{I}$ denotes the second-order identity tensor. For clarity, Equation (5.4) can be rewritten as

$$F_{ij} = \frac{\partial x_i}{\partial X_j} = \delta_{ij} + \frac{\partial u_i}{\partial X_j}\,, \tag{5.5}$$

with δ_{ij} denoting the Kronecker delta. The Jacobian J is defined as the determinant of the gradient tensor

$$J = \det \boldsymbol{F}\,, \tag{5.6}$$

and it relates the volumes in the reference and current configuration, denoted V_0 and V respectively, by

$$\mathrm{d}V = J\mathrm{d}V_0\,. \tag{5.7}$$

Strain measures

Alternative *strain measures* can be derived from the deformation gradient. For example, it may prove convenient to have a strain measure which only depends on the reference configuration—recall that the deformation gradient is a two-point tensor and, thus, depends on both configurations. This is satisfied by the right Cauchy-Green tensor

$$\boldsymbol{C} = \boldsymbol{F}^{\mathrm{T}}\boldsymbol{F}\,. \tag{5.8}$$

Other important property of the right Cauchy-Green strain tensor (and others) is that it is an objective measure, i.e., it discards any rigid body rotations that are present in

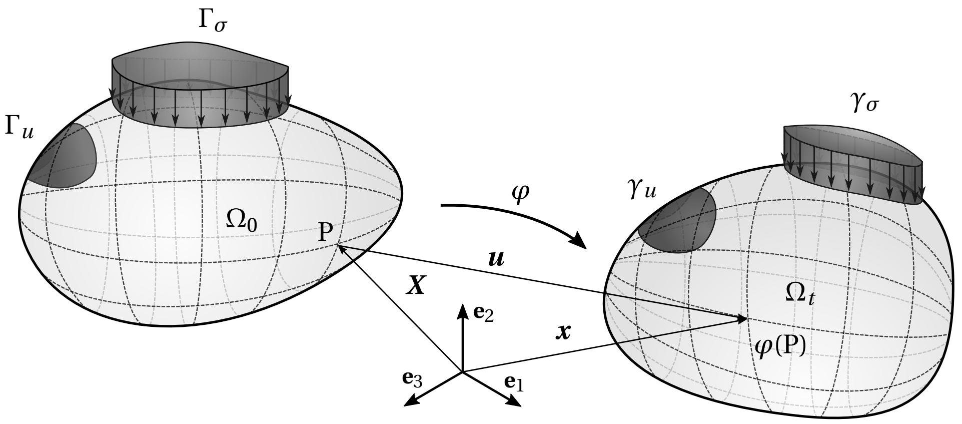


Figure 5.2: Deformable bodies in the reference and current configuration, and respective nomenclature. Adapted from Pinto Carvalho (2018).

the deformation gradient.[1] To ensure that a zero strain state occurs at the reference, or undeformed, configuration, one can define additional strain tensors such as the Green-Lagrange strain tensor $\boldsymbol{E}$

$$\boldsymbol{E} = \frac{1}{2}\left(\boldsymbol{C} - \boldsymbol{I}\right) . \tag{5.9}$$

Similarly, strain measures which depend solely on the current configuration can also be defined, e.g., the left Cauchy-Green tensor. The references listed in Section 5.2 provide a wider view of strain measures and their physical interpretation.

Stress measures

In parallel with strain measures, also several *stress measures* can be defined in nonlinear solid mechanics. The conventional Cauchy stress tensor $\boldsymbol{\sigma}$, widely known from the theory of infinitesimal deformations, is a stress measure which maps the current configuration surface element area to the *true* internal force in the body

$$\mathrm{d}\boldsymbol{f} = \boldsymbol{\sigma} \cdot \mathrm{d}A\boldsymbol{n} , \tag{5.10}$$

with $\boldsymbol{n}$ denoting the outward normal vector to the surface element of area $\mathrm{d}A$, in the current configuration. Alternatively, one can map the surface element area in the reference configuration to the true internal force in the current configuration by using the first Piola-Kirchoff tensor $\boldsymbol{P}$

$$\boldsymbol{P} = J\boldsymbol{\sigma} \cdot \boldsymbol{F}^{-\mathrm{T}} , \tag{5.11}$$

and introducing the outward normal $\boldsymbol{N}$ to the surface element of area $\mathrm{d}A_0$ in the reference configuration, the mapping writes

$$\mathrm{d}\boldsymbol{f} = \boldsymbol{P} \cdot \mathrm{d}A_0\boldsymbol{N} . \tag{5.12}$$

[1] From the polar decomposition theorem, the deformation gradient can be decomposed in two multiplicative quantities: a volume-preserving rigid body rotation tensor, and a volume-changing stretch contribution.

In Equation (5.11), $\boldsymbol{F}^{-\mathrm{T}}$ denoted the inverse of the deformation gradient transposed. The Cauchy stress tensor is symmetric, while the first Piola-Kirchoff is not. Other convenient stress measure is the second Piola-Kirchoff $\boldsymbol{S}$

$$\boldsymbol{S} = \boldsymbol{F}^{-1} \cdot \boldsymbol{P} = J\boldsymbol{F}^{-1} \cdot \boldsymbol{\sigma} \cdot \boldsymbol{F}^{-\mathrm{T}}\,, \tag{5.13}$$

which recovers the symmetry property, but does not have a clear interpretation like the previous ones.

Despite the variety of stress and strain measures, it should be kept in mind that these cannot be combined arbitrarily. There are stress-strain pairs, defined based on work conjugacy, which guarantee that the internal work is the same across combinations. For example, the first Piola-Kirchoff $\boldsymbol{P}$ must be used together with the deformation gradient $\boldsymbol{F}$, while the second Piola-Kirchoff must be combined with the Green-Lagrange strain tensor $\boldsymbol{E}$ in energy considerations.

Constitutive laws

Stresses and strains are not independent of each other. In fact, materials are characterized by constitutive laws, which specify the relation between these two physical quantities. Constitutive models can be expressed in terms of the strain energy function Ψ—also termed elastic potential. This function must satisfy some physical requirements, such as independence from rotation, and verification of the second law of thermodynamics. For example, the relation between the second Piola-Kirchoff tensor and the Green-Lagrange strain tensor writes

$$\boldsymbol{S} = \frac{\partial \Psi}{\partial \boldsymbol{E}}\,. \tag{5.14}$$

The strain energy function allows the definition of the fourth-order constitutive tensor $\mathscr{C}$, which then specifies the relation between increments in stress and strain by

$$\mathscr{C} = \frac{\partial \boldsymbol{S}}{\partial \boldsymbol{E}}\,. \tag{5.15}$$

The specific nature of Ψ or $\mathscr{C}$ comes from a the particular constitutive model adopted, for example, hyperelasticity or viscoelasticity.

Further than the continuous descriptions, mechanical systems must be characterized by their equilibrium conditions. In particular, conservation of mass, equilibrium of linear and angular momentum and energy balances must be verified. Equilibrium of angular momentum simply reduces to the conditions of symmetry of the Cauchy and second Piola-Kirchoff stress tensors. As for energy balance, since purely mechanical systems are considered in this work, it is redundant with linear momentum equilibrium. The laws of conservation of mass and equilibrium of linear momentum are presented next.

Conservation of mass

From a physical view frame, the mass of a given particle in the reference configuration must be conserved after deformation, even if its volume changes. Denoting ρ_0 and ρ the body density at the reference and current configurations, and m as the mass of the body, the conversation of mass equation writes

$$\frac{\mathrm{d}m}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}\int_{\Omega_0} \rho_0 \,\mathrm{d}V_0 = \frac{\mathrm{d}}{\mathrm{d}t}\int_{\Omega_t} \rho \,\mathrm{d}V = 0\,. \tag{5.16}$$

By applying Reynold's transport theorem, and recalling that both the reference configuration and its density ρ_0 do not depend on time, the conservation of mass reduces to the local verification of

$$\dot{\rho} + \rho\,\mathrm{div}\dot{\boldsymbol{u}} = 0\,; \tag{5.17a}$$

$$\dot{\rho}_0 = 0\,. \tag{5.17b}$$

In Expressions (5.17), the notation $(\dot{\bullet})$ stands for the total time derivative, and $\mathrm{div}\dot{\boldsymbol{u}}$ denotes the divergence of the vector field $\dot{\boldsymbol{u}}$ and reads

$$\mathrm{div}\boldsymbol{u} = \frac{\partial \dot{u}_1}{\partial x_1} + \frac{\partial \dot{u}_2}{\partial x_2} + \frac{\partial \dot{u}_3}{\partial x_3}\,. \tag{5.18}$$

Equilibrium of linear momentum

Newton's second law describes the equilibrium between linear momentum and external forces. This works is only concerned with quasi-static problems, therefore, the linear momentum is null. Under this condition, the balance of linear momentum is reduced to the equilibrium between external forces in the current configuration. Denoting the body forces in the current configuration by $\boldsymbol{b}$ (forces per unit volume of the body in the current configuration), and the surface traction in the current configuration boundary by $\boldsymbol{t}$ (forces per unit area of the boundary of the body in the current configuration), the equilibrium of forces comes

$$\int_{\Omega_t} \boldsymbol{b}\,\mathrm{d}V + \int_{\partial\Omega_t} \boldsymbol{t}\,\mathrm{d}A = 0\,. \tag{5.19}$$

Introducing Gauss divergence theorem, the local formulation of the equilibrium of forces in the current configuration follows

$$\mathrm{div}\boldsymbol{\sigma} + \boldsymbol{b} = \boldsymbol{0}\,. \tag{5.20}$$

In Equation (5.20), $\boldsymbol{0}$ is a null vector, and the divergence of a second-order tensor is a first-order tensor which verifies

$$\boldsymbol{e}_i \cdot \mathrm{div}\boldsymbol{\sigma} = \frac{\partial \sigma_{i1}}{\partial x_1} + \frac{\partial \sigma_{i2}}{\partial x_2} + \frac{\partial \sigma_{i3}}{\partial x_3}\,. \tag{5.21}$$

5.2.2 Strong form of nonlinear solid mechanics problems

Based on the governing equations established previously, and on the definition of the Neumann and Dirichlet partitions of the domain's boundary, the strong form of the *Initial Value Boundary Problem* (IBVP) of finite deformation of solids can now be stated:

Problem 5.1 (Strong form of the IBVP of nonlinear solid mechanics).
For every solid sub-domain Ω^i_t, the deformed solution must verify the system of equations which encompasses both the momentum balance and the boundary conditions of the problem (Neumann and Dirichlet):

$$\operatorname{div}\boldsymbol{\sigma}^i + \boldsymbol{b}^i = \boldsymbol{0}, \quad \text{in } \Omega^i_t \times [0,T], \tag{5.22a}$$

$$\boldsymbol{u}^i = \bar{\boldsymbol{u}}^i, \quad \text{in } \gamma^i_u \times [0,T], \tag{5.22b}$$

$$\boldsymbol{\sigma}^i \boldsymbol{n}^i = \bar{\boldsymbol{t}}^i, \quad \text{in } \gamma^i_\sigma \times [0,T], \tag{5.22c}$$

where $\bar{\boldsymbol{u}}^i$ and $\bar{\boldsymbol{t}}^i$ denote the prescribed displacements and surface tractions at the current configuration Dirichlet and Neumann boundaries, respectively, and T the total simulation time.

Despite the fact that the above strong formulation does not includes dynamic equilibrium, the total simulation time T is included in order to account for static time dependent phenomena, such as plasticity. For the same reason, the *initial boundary conditions* are not explicitly present in the formulation, and can be thought as part of the Dirichlet set.

5.2.3 Contact mechanics

So far in this chapter, the problem of solid mechanics undergoing quasi-static finite deformations has been established, assuming that the Neumann and Dirichlet partitions of the boundary were known beforehand. In addition to geometry and material laws, contact brings another source of nonlinearity to the problem, called boundary nonlinearity. This comes from the fact that the contacting boundary is not known *a priori* and must be determined as part of the solution itself—in opposition to what happens to the other boundary partitions. Despite the diversity of categories on which contact problems can be sorted (such as Signorini contact between an elastic surface a rigid wall, contact between multiple bodies and self contact), it can be formulated without loss of generality for the general case of two deformable bodies in unilateral contact.

The Figure 5.2 can be recovered with some modifications for the representation of this problem, and is illustrated in Figure 5.3. Henceforth, two solid bodies are considered, and these shall be referred as the non-mortar and mortar bodies, identified by the superscripts s and m, respectively. Each kinematic quantity discussed earlier can be attributed to each body individually. The boundary of each body is now partitioned in three disjoint open sets. Apart from the aforementioned Neumann and Dirichlet, a new partition Γ_c is introduced in each domain boundary, called the potential contact boundary, cf. Figure 5.3. For each body, it must verify

$$\Gamma^i_\sigma \cup \Gamma^i_u \cup \Gamma^i_\mathrm{c} = \partial\Omega^i_0\,, \tag{5.23a}$$

$$\Gamma^i_\sigma \cap \Gamma^i_u = \Gamma^i_\sigma \cap \Gamma^i_\mathrm{c} = \Gamma^i_\mathrm{c} \cap \Gamma^i_u = \emptyset\,. \tag{5.23b}$$

In fact, the designation *potential contact surface* itself suggests that it does not match the so called *active contact surface* $\Gamma_\mathrm{a} \subseteq \Gamma_\mathrm{c}$, because it must be found together with the displacement solution. Regions which belong to the potential contact surface but fall out

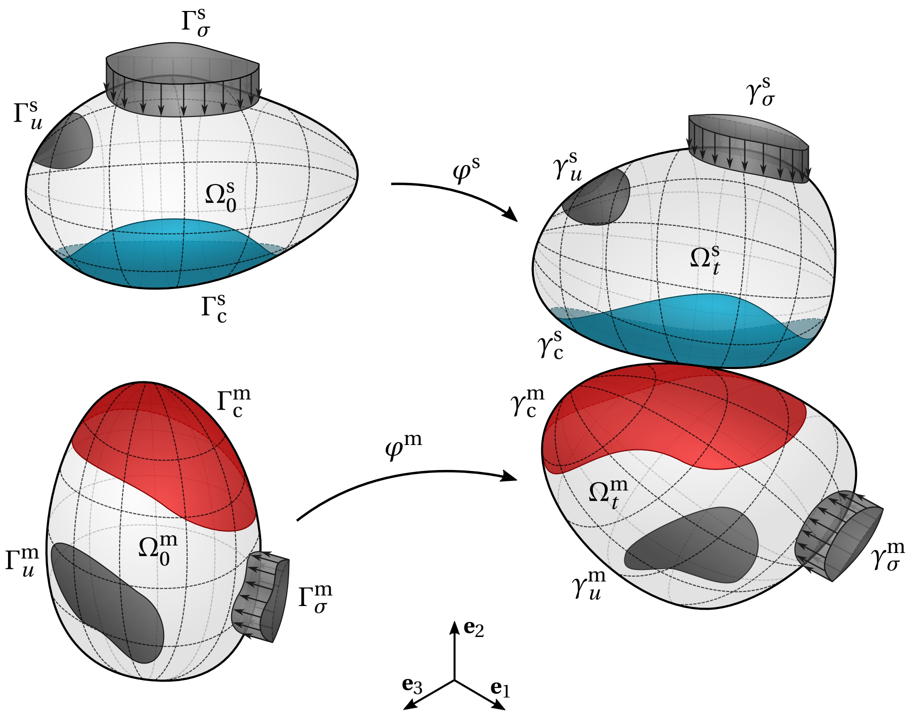

$\Gamma^{\mathrm{s}}_{\sigma}$
Γ^{s}_{u}
Ω^{s}_{0}
$\Gamma^{\mathrm{s}}_{\mathrm{c}}$
φ^{s}
$\gamma^{\mathrm{s}}_{\sigma}$
γ^{s}_{u}
Ω^{s}_{t}
$\gamma^{\mathrm{s}}_{\mathrm{c}}$
$\Gamma^{\mathrm{m}}_{\mathrm{c}}$
$\gamma^{\mathrm{m}}_{\mathrm{c}}$
φ^{m}
Ω^{m}_{t}
Γ^{m}_{u}
Ω^{m}_{0}
$\Gamma^{\mathrm{m}}_{\sigma}$
$\gamma^{\mathrm{m}}_{\sigma}$
γ^{m}_{u}
$\mathbf{e}_2$
$\mathbf{e}_3$
$\mathbf{e}_1$

Figure 5.3: Schematic illustration of a two deformable bodies in unilateral contact, along with respective nomenclature. Adapted from Pinto Carvalho (2018).

of the active contact boundary must be considered as part of the Neumann boundary:

$$\Gamma_{\mathrm{c}} \setminus \Gamma_{\mathrm{a}} \subset \Gamma_{\sigma} \,. \tag{5.24}$$

The current configuration counterparts of the potential contact boundaries $\Gamma^{\mathrm{m}}_{\mathrm{c}}$ and $\Gamma^{\mathrm{s}}_{\mathrm{c}}$ are denoted by $\gamma^{\mathrm{m}}_{\mathrm{c}}$ and $\gamma^{\mathrm{s}}_{\mathrm{c}}$, respectively—cf, Figure 5.3.

Contact kinematics

The formulation of contact mechanics problems relies on the definition of kinematic quantities responsible for describing the potential interaction between the contacting bodies. Such description must have the ability to characterize both normal and tangential contact. Furthermore, it is convenient to use one of the potential contact boundaries to parametrize contact related quantities. Here, the non-mortar potential contact boundary $\gamma^{\mathrm{s}}_{\mathrm{c}}$ is chosen for that purpose.

As a fundamental measure of proximity, potential contact and penetration of two bodies, it is convenient to define the gap function g at some point $\boldsymbol{x}^{\mathrm{s}} \in \gamma^{\mathrm{s}}_{\mathrm{c}}$ as the distance between $\boldsymbol{x}^{\mathrm{s}}$ and its projection on the mortar side, along the unit normal of the non-mortar interface $\boldsymbol{\eta}(\boldsymbol{x}^{\mathrm{s}}, t)$. The projected point at the mortar side is denoted by $\hat{\boldsymbol{x}}^{\mathrm{m}} \in \gamma^{\mathrm{m}}_{\mathrm{c}}$. For a graphical illustration of the gap function, see Figure 5.4. Formally, the gap reads

$$g(\boldsymbol{x}^{\mathrm{s}}, t) = -\boldsymbol{\eta}(\boldsymbol{x}^{\mathrm{s}}, t) \cdot \left[\boldsymbol{x}^{\mathrm{s}} - \hat{\boldsymbol{x}}^{\mathrm{m}}(\boldsymbol{x}^{\mathrm{s}}, t)\right] \,. \tag{5.25}$$

Alternatively, the gap vector can be defined as

$$\boldsymbol{g}(\boldsymbol{x}^{\mathrm{s}},t) = \boldsymbol{x}^{\mathrm{s}} - \hat{\boldsymbol{x}}^{\mathrm{m}}(\boldsymbol{x}^{\mathrm{s}},t)\,. \tag{5.26}$$

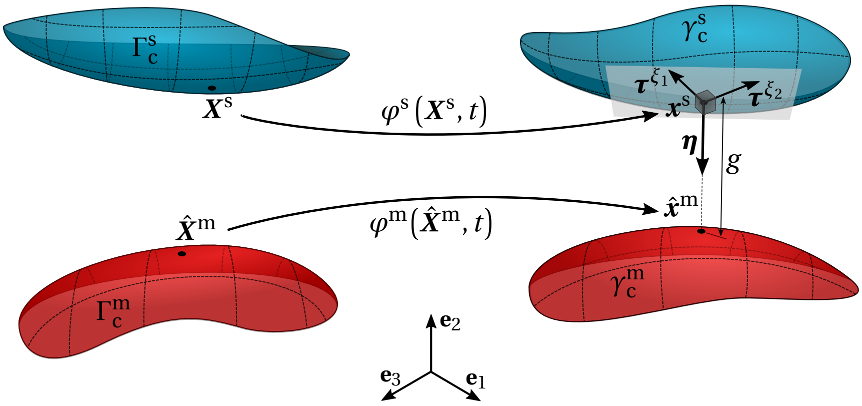

$\Gamma_{\mathrm{c}}^{\mathrm{s}}$
$\gamma_{\mathrm{c}}^{\mathrm{s}}$
$\boldsymbol{\tau}^{\xi_1}$
$\boldsymbol{\tau}^{\xi_2}$
$\boldsymbol{X}^{\mathrm{s}}$
$\varphi^{\mathrm{s}}\left(\boldsymbol{X}^{\mathrm{s}},t\right)$
$\boldsymbol{x}^{\mathrm{s}}$
$\boldsymbol{\eta}$
g
$\hat{\boldsymbol{x}}^{\mathrm{m}}$
$\hat{\boldsymbol{X}}^{\mathrm{m}}$
$\varphi^{\mathrm{m}}\left(\hat{\boldsymbol{X}}^{\mathrm{m}},t\right)$
$\gamma_{\mathrm{c}}^{\mathrm{m}}$
$\Gamma_{\mathrm{c}}^{\mathrm{m}}$
$\mathbf{e}_2$
$\mathbf{e}_3$
$\mathbf{e}_1$

Figure 5.4: Graphical definition of the gap function. The gap is defined as the distance between some point in the non-mortar interface and its projection on the mortar surface along the outward unit normal of the former boundary. All these quantities are evaluated in the current configuration. Adapted from Pinto Carvalho (2018).

Gap related quantities regard the contact description in the normal direction. For tangential contact, the primary kinematic variable is the relative tangential velocity. It can be formulated by two different approaches, either using slip advected bases or by difference of material velocities. Adopting the latter formulation, for simplicity, the relative tangential velocity $\boldsymbol{v}^{\tau}$ is the projection in the tangential direction of the time derivative of the gap vector

$$\boldsymbol{v}^{\tau} = \left(\boldsymbol{I} - \boldsymbol{\eta}\otimes\boldsymbol{\eta}\right)\dot{\boldsymbol{g}}\,, \tag{5.27}$$

where the dyadic product between two first-order tensors follows $\boldsymbol{\eta}\otimes\boldsymbol{\eta} = \boldsymbol{\eta}\cdot\boldsymbol{\eta}^{\mathrm{T}}$. It should be noted that the tangential relative velocity, as defined in Equation (5.27), is only true if the points are in contact. Nevertheless, it is common practice to used this approach to quantify tangential relative movement.

Contact constraints

The contact kinematic quantities allow the establishment of physical meaningful constraints for the contact problem. Before introducing this formulation, it is paramount to emphasize the decomposition of contact tractions into the normal and tangential components. Taking the surface traction at the non-mortar contact interface, it writes

$$\boldsymbol{t}_{\mathrm{c}}^{\mathrm{s}}(\boldsymbol{x}^{\mathrm{s}},t) = p^{\eta}\boldsymbol{\eta} + \boldsymbol{t}^{\tau}\,, \tag{5.28}$$

where p^{η} denotes the contact normal pressure, which is the only component of the contact traction in frictionless contact, and $\boldsymbol{t}^{\tau}$ is the tangential contact traction. By applying the conservation of linear momentum to the interface, it comes

$$\boldsymbol{t}_{\mathrm{c}}^{\mathrm{m}}(\hat{\boldsymbol{x}}^{\mathrm{m}},t) = -\boldsymbol{t}_{\mathrm{c}}^{\mathrm{s}}(\boldsymbol{x}^{\mathrm{s}},t)\,. \tag{5.29}$$

The contact constraints in the normal direction must guarantee that bodies do not penetrate each other, and that only compressive stresses are originated at the contact interface—adhesive contact is neglected. By inspecting the definition of the gap function Equation (5.25), it can be seen that this function is non-negative for all points. For $g = 0$, points are in the active contact boundary, and must necessarily verify $p^{\eta} < 0$. In opposition, points which verify $g > 0$ are not part of the active contact boundary, and shall meet $p^{\eta} = 0$. Contact constraints in the normal direction can be formulated from the *Karush-Kuhn-Tucker* (KKT) conditions, often designated *Hertz-Signorini-Moreau* (HSM), which read

$$g(\boldsymbol{x}^{\mathrm{s}}, t) \geq 0\,, \tag{5.30a}$$

$$p^{\eta}(\boldsymbol{\eta}, \boldsymbol{t}_{\mathrm{c}}^{\mathrm{s}}) \leq 0\,, \tag{5.30b}$$

$$p^{\eta}(\boldsymbol{\eta}, t)\, g(\boldsymbol{x}^{\mathrm{s}}, t) = 0\,. \tag{5.30c}$$

The KKT conditions must be verified for every point in the non-mortar surface, i.e., for $\boldsymbol{x}^{\mathrm{s}} \in \gamma_{\mathrm{c}}^{\mathrm{s}}$. The first condition imposes the condition of non-penetration. The second condition forces contact pressures to assume only negative values—compressive stresses. The third condition, commonly termed the *complementarity condition*, guarantees that if bodies are in contact ($g = 0$) then the contact pressure is necessarily negative, whereas if points are not contacting ($g > 0$) then contact normal pressure vanishes.

The contact constraints in the tangential direction are responsible for modeling frictional contact. The phenomenological Coulomb's friction law is often adopted to model such constraints by introducing the well-known coefficient of friction μ. It considers two different states, namely the *stick* and *slip* state. While in the stick state, the tangential contact stress can increase up to a certain limiting value μp^{η}, and the relative tangential velocity remains null. If the tangential contact stress reaches μp^{η}, then $\boldsymbol{v}^{\tau}$ can occur in the opposite direction to the tangential stress, with a certain magnitude. These conditions can be formulated as

$$\psi(\boldsymbol{t}^{\tau}, p^{\eta}) \equiv \| \boldsymbol{t}^{\tau}(\boldsymbol{x}^{\mathrm{s}}, t) \| - \mu | p^{\eta}(\boldsymbol{x}^{\mathrm{s}}, t) | \leq 0\,, \tag{5.31a}$$

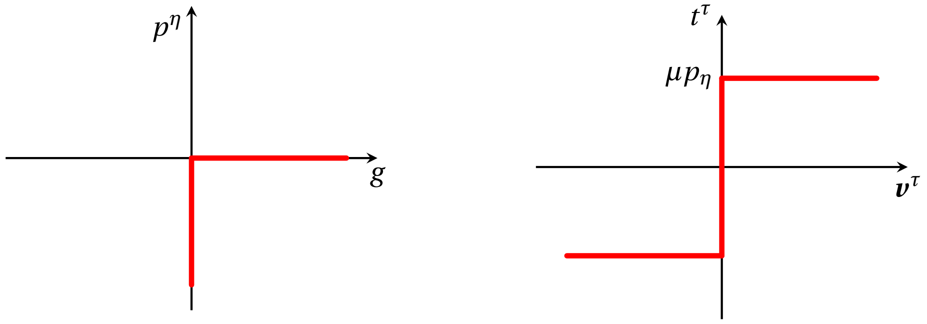
$$\boldsymbol{v}^{\tau}(\boldsymbol{x}^{\mathrm{s}}, t) + \beta \boldsymbol{t}^{\tau}(\boldsymbol{x}^{\mathrm{s}}, t) = \boldsymbol{0}\,, \tag{5.31b}$$

$$\beta \geq 0\,, \tag{5.31c}$$

$$\psi(\boldsymbol{t}^{\tau}, p^{\eta})\beta = 0\,. \tag{5.31d}$$

In the tangential contact constraints, $\psi(\boldsymbol{t}^{\tau}, p^{\eta})$ is designated the slip function, and β is a positive scalar parameter. It can be seen that when $\psi < 0$, the magnitude of the tangential contact traction is smaller than the limiting value μp^{η}, which forces $\beta = 0$, and thus there is no tangential relative velocity—stick state. When $\psi = 0$, the tangential contact traction has reached the limiting value, and tangential relative motion is free to occur, but it is ensured that it is collinear with the tangential contact traction, and in the opposite direction—slip state.

A graphical representation of the contact constrains is given in Figure 5.5. It should not be overlooked that both contact constraints are non-smooth and multivalued at the origin.



(a) KKT conditions

(b) Coulomb's friction law

Figure 5.5: Graphical representation of contact constraints in the normal and tangential direction. Both conditions are non-smooth and multivalued at the origin.

5.2.4 Strong form of the finite deformation frictional contact

Attending to the formulation of the contact constraints, continuum contact mechanics problems can be formulated as a constrained classical solid mechanics problem. The IBVP for the general scenario of finite deformation frictional contact follows:

Problem 5.2 (Strong form of the IBVP of finite deformation frictional contact). *For every solid sub-domain Ω_t^i, the deformed solution must verify the system of equations which encompasses both the momentum balance and the boundary conditions of the problem (Neumann and Dirichlet):*

$$\operatorname{div}\boldsymbol{\sigma}^i + \boldsymbol{b}^i = \boldsymbol{0}, \quad \text{in } \Omega_t^i \times [0,T], \tag{5.32a}$$

$$\boldsymbol{u}^i = \bar{\boldsymbol{u}}^i, \quad \text{in } \gamma_u^i \times [0,T], \tag{5.32b}$$

$$\boldsymbol{\sigma}^i \boldsymbol{n}^i = \bar{\boldsymbol{t}}^i, \quad \text{in } \gamma_\sigma^i \times [0,T], \tag{5.32c}$$

and the contact constraints in the normal and tangential directions

$$g \geq 0\,, \quad p^\eta \leq 0\,, \quad p^\eta g = 0\,, \quad \text{in } \gamma_{\mathrm{c}}^{\mathrm{s}} \times [0,T]\,, \tag{5.33a}$$

$$\psi \leq 0\,, \quad \boldsymbol{v}^\tau + \beta \boldsymbol{t}^\tau = \boldsymbol{0}\,, \quad \beta \geq 0\,, \quad \psi\beta = 0\,, \quad \text{in } \gamma_{\mathrm{c}}^{\mathrm{s}} \times [0,T]\,. \tag{5.33b}$$

5.2.5 Weak form of the contact problem

The strong formulation of the contact problem within the context of finite deformation solid mechanics must be conveyed to a weak formulation, prone to the application of the FEM. The weak form is derived from the application of the *Principle of Virtual Work* (PVW), by introducing a kinematically admissible virtual displacement field $\delta\boldsymbol{u}$. The solution spaces $\mathcal{U}^i$ and $\mathcal{V}^i$ for the displacement and virtual displacement fields are defined, respectively, as

$$\mathcal{U}^i \equiv \left\{ \boldsymbol{u}^i \in \left[H^1 \left(\Omega_t^i \right) \right]^d \mid \boldsymbol{u}^i = \bar{\boldsymbol{u}}^i \text{ in } \gamma_u^i \right\} , \tag{5.34}$$

$$\mathcal{V}^i \equiv \left\{ \delta \boldsymbol{u}^i \in \left[H^1 \left(\Omega_t^i \right) \right]^d \mid \delta \boldsymbol{u}^i = \boldsymbol{0} \text{ in } \gamma_u^i \right\} , \tag{5.35}$$

where $H^1 \left(\Omega_0^i \right)$ denotes the space of all square integrable functions over the domain—the so called Sobolev space. With the purpose of simplifying the notation, the product spaces $\mathcal{U} \equiv \mathcal{U}^{\mathrm{s}} \times \mathcal{U}^{\mathrm{m}}$ and $\mathcal{V} \equiv \mathcal{V}^{\mathrm{s}} \times \mathcal{V}^{\mathrm{m}}$ are introduced, and the superscripts are dropped from the physical and virtual displacements. The PVW states that

$$\delta \Pi_{\text{int}} \left(\boldsymbol{u}, \delta \boldsymbol{u} \right) - \delta \Pi_{\text{ext}} \left(\delta \boldsymbol{u} \right) + \delta \Pi_{\text{c}} \left(\boldsymbol{u}, \delta \boldsymbol{u} \right) = 0 , \quad \forall \delta \boldsymbol{u} \in \mathcal{V} . \tag{5.36}$$

In the above, $\delta \Pi_{\text{int}}$ is the internal virtual work due to internal forces, which reads

$$\delta \Pi_{\text{int}} \left(\boldsymbol{u}, \delta \boldsymbol{u} \right) = - \sum_{i \in \{\mathrm{s},\mathrm{m}\}} \left[\int_{\Omega_t^i} \boldsymbol{\sigma}^i : \nabla_x \left(\delta \boldsymbol{u}^i \right) \, \mathrm{d}\Omega_t^i \right] . \tag{5.37}$$

The operator $(\bullet) : (\bullet)$ represents the tensor double contraction, in this case, for two second order tensor, which writes

$$\boldsymbol{A} : \boldsymbol{B} \equiv \sum_i \sum_j A_{ij} B_{ij} , \tag{5.38}$$

and $\nabla_x (\bullet)$ denotes the spatial gradient of a vector field

$$\nabla_x \left(\delta \boldsymbol{u} \right) = \sum_{i=1}^{3} \sum_{j=1}^{3} \frac{\partial \delta u_j}{\partial x_i} \boldsymbol{e}_i \otimes \boldsymbol{e}_j . \tag{5.39}$$

Regarding the virtual work from external forces, this classical result writes

$$\delta \Pi_{\text{ext}} \left(\delta \boldsymbol{u} \right) = - \sum_{i \in \{\mathrm{s},\mathrm{m}\}} \left[\int_{\Omega_t^i} \boldsymbol{b}^i \cdot \delta \boldsymbol{u}^i \, \mathrm{d}\Omega_t^i + \int_{\gamma_\sigma^i} \boldsymbol{t}^i \cdot \delta \boldsymbol{u}^i \, \mathrm{d}\gamma_\sigma^i \right] . \tag{5.40}$$

Finally, the virtual work due to contact interactions comes

$$\delta \Pi_{\text{c}} \left(\boldsymbol{u}, \delta \boldsymbol{u} \right) = - \int_{\gamma_{\text{c}}^{\text{s}}} \boldsymbol{t}_{\text{c}}^{\text{s}} \left(\delta \boldsymbol{u}^{\text{s}} - \delta \hat{\boldsymbol{u}}^{\text{m}} \right) \, \mathrm{d}\gamma_{\text{c}} . \tag{5.41}$$

In Equation (5.41), the conservation of momentum at the contact interface expressed in Equation (5.29) is introduced. The symbol $\delta \hat{\boldsymbol{u}}^{\text{m}}$ denotes the virtual displacement of the projected point at the mortar boundary—as defined in Section 5.2.3.

Enforcement of the contact constraints

In the formulation of mortar finite element methods, the contact constraints are treated via the introduction of a Lagrange multiplier vector $\boldsymbol{\lambda}$. Here, it is set to the negative contact traction vector on the non-mortar boundary, i.e.

$$\boldsymbol{\lambda} = - \boldsymbol{t}_{\text{c}}^{\text{s}} . \tag{5.42}$$

Analogously to the decomposition of the surface traction vector in the normal an tangential direction (see Equation (5.28)), also the Lagrange multiplier can be similarly decomposed as

$$\boldsymbol{\lambda} = \lambda^{\eta}\boldsymbol{\eta} + \boldsymbol{\lambda}^{\tau}\,. \tag{5.43}$$

At this stage, the solution space for the Lagrange multipliers must be defined, so that it can be included in the weak form. In the context of the dual mortar method, the solution space $\mathcal{M}(\boldsymbol{\lambda})$ is defined as the dual space of the restriction of the solution space $\mathcal{U}^{\mathrm{s}}$ to the potential contact boundary $\gamma_{\mathrm{c}}^{\mathrm{s}}$. For a rigorous mathematical definition of $\mathcal{M}(\boldsymbol{\lambda})$, the reader is referred to Hüeber (2008). This solution space satisfies the KKT optimality conditions and the Coulomb's laws of friction in the weak sense.

By writing the contact constrains as variational inequities (see the work of Kikuchi and J. T. Oden (1988)), the weak form of the IBVP problem of finite deformation frictional contact can be summarized as follows:

Problem 5.3 (Weak form of the IBVP of finite deformation frictional contact). *Find the kinematically admissible displacement field* $\boldsymbol{u} \in \mathcal{U}$ *and the Lagrange multiplier vector* $\boldsymbol{\lambda} \in \mathcal{M}(\boldsymbol{\lambda})$*, such that, for all* $t \in [0, T]$*, the PVW holds*

$$\delta\Pi_{\text{int}}\left(\boldsymbol{u},\delta\boldsymbol{u}\right) - \delta\Pi_{\text{ext}}\left(\delta\boldsymbol{u}\right) + \int_{\gamma_{\mathrm{c}}^{\mathrm{s}}} \boldsymbol{\lambda}\left(\delta\boldsymbol{u}^{\mathrm{s}} - \delta\hat{\boldsymbol{u}}^{\mathrm{m}}\right)\,\mathrm{d}\gamma_{\mathrm{c}} = 0\,, \quad \forall \delta\boldsymbol{u} \in \mathcal{V}\,, \tag{5.44}$$

$$\int_{\gamma_{\mathrm{c}}^{\mathrm{s}}} g\left(\delta\lambda^{\eta} - \lambda^{\eta}\right)\,\mathrm{d}\gamma_{\mathrm{c}}^{\mathrm{s}} \geq 0\,, \quad \forall \delta\boldsymbol{\lambda} \in \mathcal{M}\left(\boldsymbol{\lambda}\right)\,, \tag{5.45}$$

$$\int_{\gamma_{\mathrm{c}}^{\mathrm{s}}} \boldsymbol{v}^{\tau} \cdot \left(\delta\boldsymbol{\lambda}^{\tau} - \boldsymbol{\lambda}^{\tau}\right)\,\mathrm{d}\gamma_{\mathrm{c}}^{\mathrm{s}} \leq 0\,, \quad \forall \delta\boldsymbol{\lambda} \in \mathcal{M}\left(\boldsymbol{\lambda}\right)\,. \tag{5.46}$$

Remark 5.2 on the restriction of the finite element formulation to normal contact. *So far in the current chapter, both the normal and tangential contact constrains were referred, in order to present the general formulation of finite deformation frictional contact. Henceforth, in particular, for the finite element approximation, the frictional constraints will be omitted, since the numerical simulations were performed assuming frictionless contact. In order to preserve some coherence between the mathematical formulation and the numerical work, the next section shall focus only on normal contact.*

5.2.6 Mortar finite element discretization

The weak formulation of the finite deformation contact problem can be discretized by recurring to the finite element method. The fundamental idea of the FEM is the utilization of the finite dimensional spaces $\mathcal{U}^h \subset \mathcal{U}$ and $\mathcal{V}^h \subset \mathcal{V}$ for the solution spaces of the displacement and virtual displacement fields. The problem domain $\Omega = \Omega_t^{\mathrm{s}} \cup \Omega_t^{\mathrm{m}}$ is divided into n^e sub-domains $\Omega_e \subset \Omega^h$ (the superscript h denotes a FE discretized variable), such that

$$\Omega \approx \Omega^h \equiv \bigcup_{e=1}^{n^e} \Omega_e\,. \tag{5.47}$$

The finite elements are connected at nodes, which form a finite element mesh. The basis functions of the finite dimensional spaces $\mathcal{U}^h$ and $\mathcal{V}^h$ have compact support, meaning they are zero everywhere, except in the elements immediately surrounding a given node. These interpolation, or shape, functions are usually specified in an element-basis, using standardized parameter spaces

$$\boldsymbol{\xi} = (\xi_1, \ldots, \xi_d)\ . \tag{5.48}$$

Adopting an isoparametric approach, the same shape functions are used to interpolate both the displacement field and the geometry at every sub-domain. The finite discretization of the bulk is completely independent of the mortar approach adopted for modeling contact.

The boundary geometry and field variables are also interpolated the using shape function, with dimension $d-1$. The boundary geometry interpolation writes

$$\boldsymbol{x}^{\mathrm{s}} \approx \left\{\boldsymbol{x}^{\mathrm{s}}\right\}^h\Big|_{\{\gamma_{\mathrm{c}}^{\mathrm{s}}\}^h} = \sum_{k=1}^{n^{\mathrm{s}}} N_k^{\mathrm{s}}\left(\boldsymbol{\xi}^{\mathrm{s}}\right)\mathbf{x}_k^{\mathrm{s}}\,, \tag{5.49a}$$

$$\boldsymbol{x}^{\mathrm{m}} \approx \left\{\boldsymbol{x}^{\mathrm{m}}\right\}^h\Big|_{\{\gamma_{\mathrm{c}}^{\mathrm{m}}\}^h} = \sum_{l=1}^{n^{\mathrm{m}}} N_l^{\mathrm{m}}\left(\boldsymbol{\xi}^{\mathrm{m}}\right)\mathbf{x}_l^{\mathrm{m}}\,, \tag{5.49b}$$

and similarly the field variables at the boundaries

$$\boldsymbol{u}^{\mathrm{s}} \approx \left\{\boldsymbol{u}^{\mathrm{s}}\right\}^h\Big|_{\{\gamma_{\mathrm{c}}^{\mathrm{s}}\}^h} = \sum_{k=1}^{n^{\mathrm{s}}} N_k^{\mathrm{s}}\left(\boldsymbol{\xi}^{\mathrm{s}}\right)\mathbf{d}_k^{\mathrm{s}}\,, \tag{5.50a}$$

$$\boldsymbol{u}^{\mathrm{m}} \approx \left\{\boldsymbol{u}^{\mathrm{m}}\right\}^h\Big|_{\{\gamma_{\mathrm{c}}^{\mathrm{m}}\}^h} = \sum_{l=1}^{n^{\mathrm{m}}} N_l^{\mathrm{m}}\left(\boldsymbol{\xi}^{\mathrm{m}}\right)\mathbf{d}_l^{\mathrm{m}}\,. \tag{5.50b}$$

In the above, n^{s} and n^{m} represent, respectively, the number of nodes in the non-mortar and mortar boundaries, $\left\{\gamma_{\mathrm{c}}^{i}\right\}^h$ denotes the discretized boundaries, $N(\boldsymbol{\xi}^i)$ the shape function at the boundaries, $\mathbf{x}$ the nodal coordinates and $\mathbf{d}$ the nodal displacements. Accordingly, the Lagrange multipliers are interpolated from the finite dimensional set $\mathcal{M}^h \subset \mathcal{M}$

$$\boldsymbol{\lambda} \approx \boldsymbol{\lambda}^h = \sum_{j=1}^{n^{\lambda}} \Phi_j\left(\boldsymbol{\xi}^{\mathrm{s}}\right)\mathbf{z}_j\,, \tag{5.51}$$

where n^{λ} denotes the number of non-mortar nodes carrying Lagrange multipliers, Φ_j the Lagrange multiplier interpolation function and $\mathbf{z}_j$ the discrete nodal Lagrange multipliers.

With the introduction of the mortar finite element discretization in the weak form of the contact problem (cf. Equation (5.44)), the virtual work of the contact tractions results in the so called *first mortar coupling matrix* and the *second mortar coupling matrix*, here termed **D** and **M**, respectively. The elements of both these matrices are defined by

$$\mathrm{D}_{jk} = \int_{\{\gamma_{\mathrm{c}}^{\mathrm{s}}\}^h} \Phi_j(\boldsymbol{\xi}^{\mathrm{s}}) N_k^{\mathrm{s}}(\boldsymbol{\xi}^{\mathrm{s}})\ \mathrm{d}\gamma_{\mathrm{c}}^{\mathrm{s}}\,, \quad \text{for } j = 1, \ldots, n^{\lambda}\,,\ k = 1, \ldots, n^{\mathrm{s}}\,; \tag{5.52}$$

$$\mathrm{M}_{jl} = \int_{\{\gamma_{\mathrm{c}}^{\mathrm{s}}\}^h} \Phi_j(\boldsymbol{\xi}^{\mathrm{s}}) N_k^{\mathrm{m}}(\hat{\boldsymbol{\xi}}^{\mathrm{m}})\ \mathrm{d}\gamma_{\mathrm{c}}^{\mathrm{s}}\,, \quad \text{for } j = 1, \ldots, n^{\lambda}\,,\ l = 1, \ldots, n^{\mathrm{m}}\,. \tag{5.53}$$

The first mortar matrix involves only the integration of quantities related to the non-mortar boundary, while the second mortar matrices requires the integration over the discrete non-mortar boundary of quantities which relate to both boundaries—including projected quantities $\hat{\boldsymbol{\xi}}^{\mathrm{m}}$. Numerical schemes for the integration of the mortar matrices, such as segmentation, can be found in Popp (2012), Farah *et al.* (2015), and Pinto Carvalho (2018).

5.2.7 Dual Lagrange multipliers

When the *standard basis* for the Lagrange multipliers is adopted, the boundary displacements interpolation functions N_j^{s} are also chosen for Φ_j, regarding the interpolation of the Lagrange multipliers. This leads to strong coupling relations between displacements and the Lagrange multipliers—the first mortar matrix is densely populated. By adopting the dual Lagrange multipliers, proposed in B. Wohlmuth (2000), the *bi-orthogonality condition* applies, viz.

$$\int_{\{\gamma_{\mathrm{c}}^{\mathrm{s}}\}^h} \Phi_j(\boldsymbol{\xi}^{\mathrm{s}}) N_k^{\mathrm{s}}(\boldsymbol{\xi}^{\mathrm{s}})\ \mathrm{d}\gamma_{\mathrm{c}}^{\mathrm{s}} = \delta_{jk} \int_{\{\gamma_{\mathrm{c}}^{\mathrm{s}}\}^h} N_k^{\mathrm{s}}(\boldsymbol{\xi}^{\mathrm{s}})\ \mathrm{d}\gamma_{\mathrm{c}}^{\mathrm{s}}\,. \tag{5.54}$$

On one hand, this diagonalizes the first mortar matrix and, thus, localizes the coupling between Lagrange multipliers and displacements. Furthermore, it decouples the weak form of the normal contact constraints in Equation (5.45), which can now be written as a set of point wise conditions (Hüeber, 2008)

$$\tilde{g}_j \geq 0\,, \tag{5.55a}$$

$$z_j^{\eta} \geq 0\,, \tag{5.55b}$$

$$\tilde{g}_j z_j^{\eta} = 0\,, \tag{5.55c}$$

at all non-mortar nodes, i.e., for $j = 1, \ldots, n^{\mathrm{s}}$. The discrete weighted gap $\tilde{g}$ writes

$$\tilde{g}\left(\boldsymbol{\xi}^{\mathrm{s}}\right) = \int_{\{\gamma_{\mathrm{c}}^{\mathrm{s}}\}^h} \Phi_j\left(\boldsymbol{\xi}^{\mathrm{s}}\right) g^h\left(\boldsymbol{\xi}^{\mathrm{s}}\right)\ \mathrm{d}\gamma_{\mathrm{c}}^{\mathrm{s}}\,. \tag{5.56}$$

It must be noted that for the case of frictionless contact considered here, the only component of the Lagrange multiplier vector is λ^{η}, hence $\boldsymbol{\lambda}^{\tau} = \mathbf{0}$. All things considered, the discretization of the weak form of the frictionless contact problem by using a dual mortar finite element approach results in a system of generally nonlinear equations, derived from Equation (5.44), and a set of point-wise inequalities, in Equations (5.55).[2]

5.2.8 Primal-Dual active set strategy

The latter discrete formulation of the normal contact problem is still not amenable for an efficient numerical treatment, due to the necessity to treat inequalities. To cope with this difficulty, the problem can be regularized by the introduction of a *Nonlinear Complementarity* (NPC) function. Regarding frictionless contact, the NPC function for normal constraints follows (Hüeber and B. I. Wohlmuth, 2005)

$$C_j^{\eta}(\mathbf{d}, \mathbf{z}) = z_j^{\eta} - \max\left\{0,\ z_j^{\eta} - c^{\eta}\tilde{g}_j\right\}, \quad c^{\eta} > 0\,. \tag{5.57}$$

[2] For simplicity, the system of equations is not explicitly presented.

The point-wise verification of the *inequalities* in Equations (5.55) can be transposed to the point-wise verification of the following *equalities*

$$C_j^\eta = 0, \quad j = 1, \dots, n^{\mathrm{s}} \,. \tag{5.58}$$

The normal complementarity parameter c^η is an algorithmic parameter, which has been thought to influence only the convergence rate and not the accuracy of the method. Nonetheless, it has not been verified yet to impact any of the previous characteristics. The NCP function for the normal direction is plotted in Figure 5.6, and the resemblance with the KKT conditions can readily be seen. In its branched structure, there is a natural distinction between the active and inactive contact nodes.

With the introduction of the NCP function for the normal direction, the problem can be regularized and expressed as a set of nonlinear equalities. This is due to the nature of the maximum function max(•), which is semi-smooth and allows the computation of directional derivatives. Therefore, the application of semi-smooth Newton-Raphson type algorithms can be used to include all sources of nonlinearities of the problem, within a single loop.

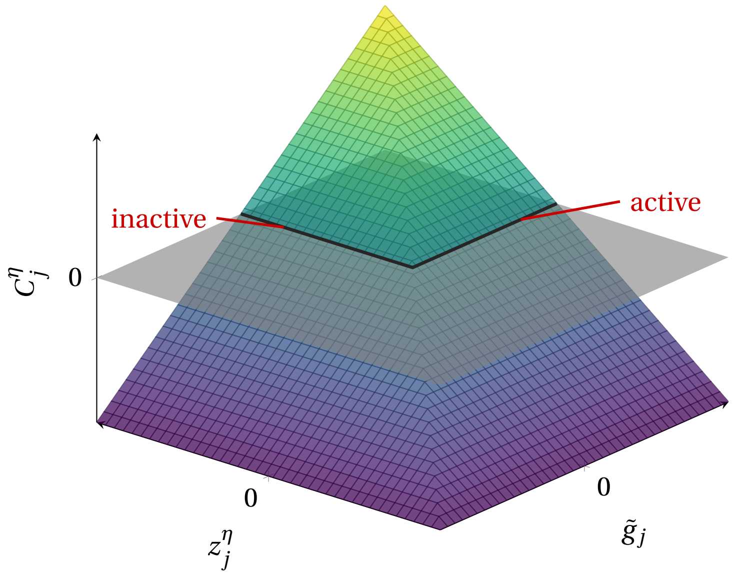


Figure 5.6: Nodal nonlinear complementarity function C_j^η for the normal contact constraints, with $c^\eta = 1$.

In brief words, the global solution algorithm consists in finding the primal-dual pair $\left(\Delta\mathbf{d}, \mathbf{z}^{k+1}\right)$ by solving the regularized system of nonlinear equations ($\Delta\mathbf{d}$ is the displacement increment), followed by the update of the displacement vector and the active set (set of nodes in contact, as evaluated by Equations 5.55). When the active set stops changing and the system of equations residuals meets some user defined tolerance, the iterations have converged, and a new load step can be started.

5.3 A FEM approach to rough contact

The present work, namely, the current chapter, concerns the single scale numerical modeling of rough contact by means of the finite element method, coupled with dual mortar technology. The numerical framework was built on top of a in-house `Fortran` program called `LINKS` (Large Strain Non-linear Analysis of Solids Linking Scales), developed by CM2S (Computational Multi-Scale Modeling of Solids and Structures) at the Faculty of Engineering of University of Porto. `LINKS` is a finite element code for implicit small and large strain analysis of several types of materials, such as elastic and elasto-plastic, equipped with a dual mortar contact formulation, introduced in Pinto Carvalho (2018).

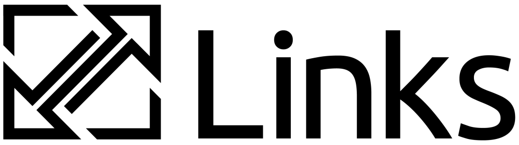


Figure 5.7: `LINKS` logo.

A preprocessing toolbox for rough contact has been implemented as part of this thesis, focusing on rough topography synthesis and respective finite element mesh generation. The code was written in the programming language `Python`, in order to take advantage of its extensive scientific libraries and its versatility in merging the different programming frameworks necessary for the execution of `LINKS` analysis.

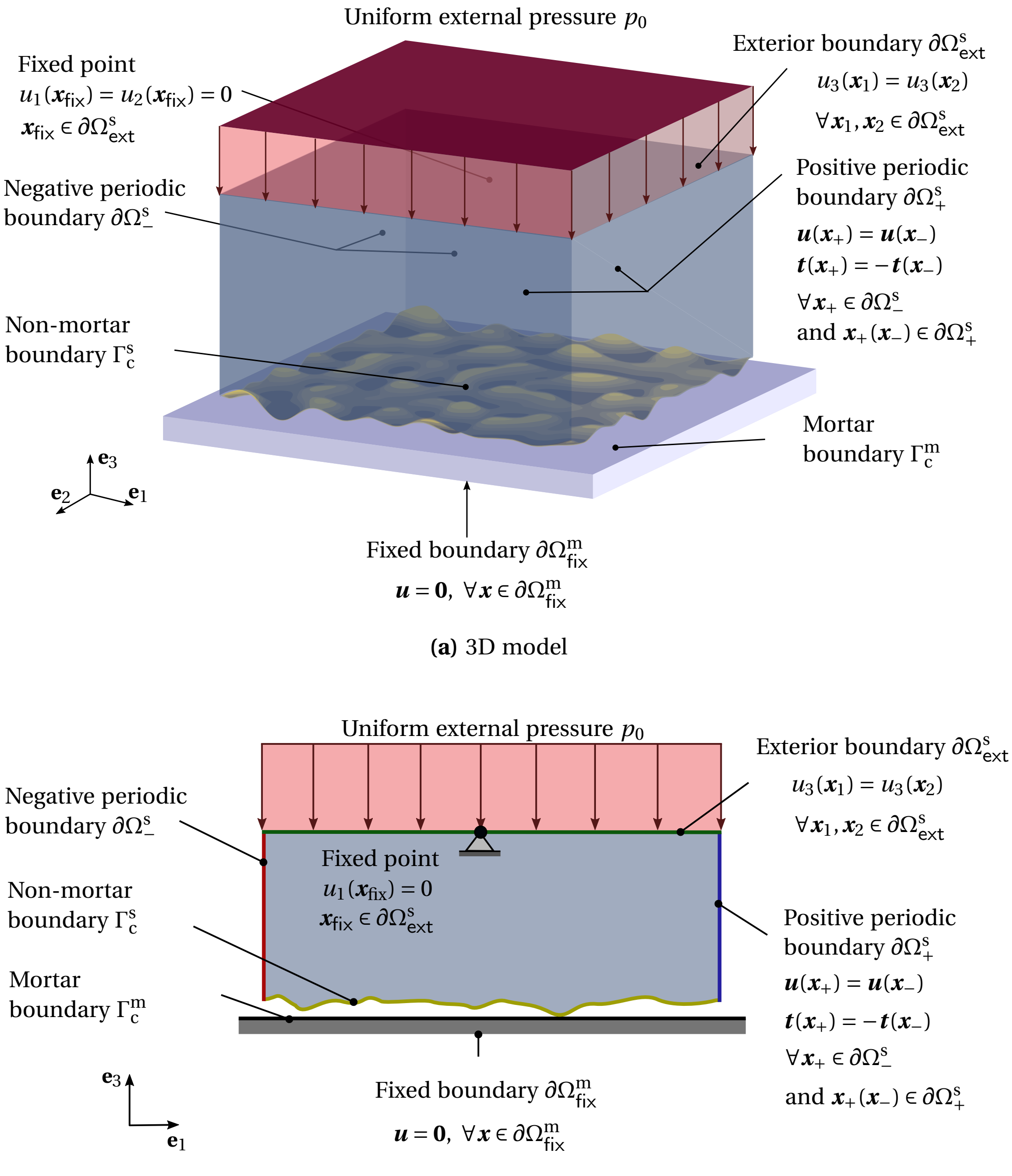
5.3.1 Numerical model setup

For modeling micromechanical contact, a new analysis type was specified in the code, named a *Representative Contact Element* (RCE). It is meant to characterize the contact interface only at the roughness level, ignoring the macroscale geometrical structure of the body.[3] For the definition of the numerical micromechanical model, one needs, firstly, to setup the geometry and materials for the contacting bodies. Then, a set admissible boundary conditions must be specified, together with physically reasonable loads. For computational convenience, only two dimensional single scale FE simulations are addressed in this chapter, due the fast scaling of computational resources required to model the problem—as observed in later sections of this chapter. Regardless, the problem definition for both 2D and 3D analysis follows, for the sake of completeness. Figure 5.8 illustrates the micromechanical problems for both cases.

General geometry description

Starting off with the contact geometry, here only Signorini type problems will be solved. This is a two body contact problem, where one of the bodies is deformable and elastic, and the other is infinitely rigid. In the following, the rough block (body 1) is considered the deformable body. Its geometry is defined, naturally, by a numerically generated *pe-*

[3] In fact, the future practical application of the methods described below requires this additional modeling feature, which must then be coupled with the henceforth discussed micromechanical contact analysis.


Uniform external pressure p_0
Fixed point
$u_1(\boldsymbol{x}_{\mathrm{fix}}) = u_2(\boldsymbol{x}_{\mathrm{fix}}) = 0$
$\boldsymbol{x}_{\mathrm{fix}} \in \partial\Omega^{\mathrm{s}}_{\mathrm{ext}}$
Exterior boundary $\partial\Omega^{\mathrm{s}}_{\mathrm{ext}}$
$u_3(\boldsymbol{x}_1) = u_3(\boldsymbol{x}_2)$
$\forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in \partial\Omega^{\mathrm{s}}_{\mathrm{ext}}$
Negative periodic boundary $\partial\Omega^{\mathrm{s}}_{-}$
Positive periodic boundary $\partial\Omega^{\mathrm{s}}_{+}$
$\boldsymbol{u}(\boldsymbol{x}_+) = \boldsymbol{u}(\boldsymbol{x}_-)$
$\boldsymbol{t}(\boldsymbol{x}_+) = -\boldsymbol{t}(\boldsymbol{x}_-)$
$\forall \boldsymbol{x}_+ \in \partial\Omega^{\mathrm{s}}_{-}$
and $\boldsymbol{x}_+(\boldsymbol{x}_-) \in \partial\Omega^{\mathrm{s}}_{+}$
Non-mortar boundary $\Gamma^{\mathrm{s}}_{\mathrm{c}}$
Mortar boundary $\Gamma^{\mathrm{m}}_{\mathrm{c}}$
$\mathbf{e}_3$
$\mathbf{e}_2$
$\mathbf{e}_1$
Fixed boundary $\partial\Omega^{\mathrm{m}}_{\mathrm{fix}}$
$\boldsymbol{u} = \boldsymbol{0}, \ \forall \boldsymbol{x} \in \partial\Omega^{\mathrm{m}}_{\mathrm{fix}}$
Uniform external pressure p_0
Exterior boundary $\partial\Omega^{\mathrm{s}}_{\mathrm{ext}}$
$u_3(\boldsymbol{x}_1) = u_3(\boldsymbol{x}_2)$
$\forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in \partial\Omega^{\mathrm{s}}_{\mathrm{ext}}$
Negative periodic boundary $\partial\Omega^{\mathrm{s}}_{-}$
Fixed point
$u_1(\boldsymbol{x}_{\mathrm{fix}}) = 0$
$\boldsymbol{x}_{\mathrm{fix}} \in \partial\Omega^{\mathrm{s}}_{\mathrm{ext}}$
Non-mortar boundary $\Gamma^{\mathrm{s}}_{\mathrm{c}}$
Positive periodic boundary $\partial\Omega^{\mathrm{s}}_{+}$
$\boldsymbol{u}(\boldsymbol{x}_+) = \boldsymbol{u}(\boldsymbol{x}_-)$
$\boldsymbol{t}(\boldsymbol{x}_+) = -\boldsymbol{t}(\boldsymbol{x}_-)$
$\forall \boldsymbol{x}_+ \in \partial\Omega^{\mathrm{s}}_{-}$
and $\boldsymbol{x}_+(\boldsymbol{x}_-) \in \partial\Omega^{\mathrm{s}}_{+}$
Mortar boundary $\Gamma^{\mathrm{m}}_{\mathrm{c}}$
$\mathbf{e}_3$
$\mathbf{e}_1$
Fixed boundary $\partial\Omega^{\mathrm{m}}_{\mathrm{fix}}$
$\boldsymbol{u} = \boldsymbol{0}, \ \forall \boldsymbol{x} \in \partial\Omega^{\mathrm{m}}_{\mathrm{fix}}$

(a) 3D model

(b) 2D model

Figure 5.8: Numerical model setup, with emphasis on the boundary conditions, for simulating micromechanical contact with the FEM, in two and three dimensions.

riodic topography at one side, and flat boundary, parallel to the rough boundary mean plane. In sum, the rough block is a parallelepiped/rectangle where one of the faces was replaced by a rough boundary. The required height and length for the rough block are not known beforehand, and must be established with further investigations. The analysis is restricted to self-affine rough profiles and surfaces. The rigid flat base (body 2) is modeled simply as a rectangular box, whose planar dimensions, i.e., at the contact interface, are slightly larger than the rough block's, such that expansion in the plane directions can be accommodated without causing dropping edges. Its height is set to the mesh size at the non-mortar interface. This is, the flat rigid block is realized by a single layer of square elements, thus its height changes with the discretization of the rough boundary. However, this does not impact the results, since there are virtually no displacements in this body.

Rough boundary features

The geometry of the self-affine rough boundary is characterized by its Hurst exponent H, the large and short cut-off wavelengths λ_l and λ_s, respectively, the roll-off frequency λ_r and a scaling factor. As a starting point, only Gaussian topographies will be considered throughout this work. In order to remove one degree of freedom from the numerical studies, only topography without roll-off frequency shall be analyzed, i.e., $\lambda_r = \lambda_l$. By fixing all previous parameters and varying only the scaling factor, one can control specific properties of the topography, such as the RMS slope and, thus, the overall smoothness of the topography. This is crucial for the convergence of the numerical method, since the field of unit outward normals for very rough profiles (very high RMS slope) changes drastically within the FE mesh. In fact, the selection of the scaling factor can be interpreted as a normalization of the generated surface height. For example, if the profile was prescribed a value for the RMS slope, one could simply compute this RMS parameter by a finite differences formula (see Equations (2.6)) or by the discrete spectral moments (see Equations (2.76) and (2.78)), and normalize the topography heights through an arithmetic division. However, as pointed out by Yastrebov, Anciaux, *et al.* (2015), this approach introduces an implicit dependence on the discretization, and often results in underestimated RMS properties.

To circumvent this difficulty, one can stray from discretized operations by computing the required scaling factor that prescribes a specific spectral moment directly from the analytical expression of the PSD—and then relate it to the RMS parameter. The *discrete* PSD of a self-affine profile follows

$$\hat{\Phi}^{\theta}\left[k = \frac{p}{N}\Omega_s\right] = \begin{cases} \hat{C}_0' & , \quad k_l \leq k < k_s \\ \hat{C}_0' \left(\dfrac{k_r}{k}\right)^{1+2H} & , \quad k_r \leq k \leq k_s \\ 0 & , \quad \text{elsewhere}\,. \end{cases} \tag{5.59}$$

Note that here the discrete PSD is used, in contrast with the previous presented formula in Equation (2.57), and $\hat{C}_0'$ denotes the discrete PSD scaling factor. Additionally, for the sake of completeness, the expressions are derived for the general case of topography with roll-off frequency. By relating the discrete PSD with its continuous version via Equation (2.74), and from the definition of spectral moments in Equation (2.33), the analytical formulas for the spectral properties of self-affine profiles come

$$m_0 = \frac{l_s \hat{C}_0' k_r}{\pi}\left(1-\xi+\frac{1-\zeta^{-2H}}{2H}\right); \tag{5.60a}$$

$$m_2 = \frac{l_s \hat{C}_0' k_r^3}{\pi}\left(\frac{1-\xi^3}{3}+\frac{\zeta^{2-2H}-1}{2-2H}\right); \tag{5.60b}$$

$$m_4 = \frac{l_s \hat{C}_0' k_r^5}{\pi}\left(\frac{1-\xi^5}{5}+\frac{\zeta^{4-2H}-1}{4-2H}\right); \tag{5.60c}$$

$$\alpha = \left(1-\xi+\frac{1-\zeta^{-2H}}{2H}\right)\left(\frac{1-\xi^5}{5}+\frac{\zeta^{4-2H}-1}{4-2H}\right)\Big/\left(\frac{1-\xi^3}{3}+\frac{\zeta^{2-2H}-1}{2-2H}\right)^2 . \tag{5.60d}$$

In Equations (5.60a), $\xi = \lambda_r/\lambda_l$ and $\zeta = \lambda_r/\lambda_s$. When the roll-off frequency is not regarded, $k_r = k_l$, $\xi = 1$ and $\zeta = \lambda_l/\lambda_s$. Equations (5.60a) allow the computation of the scaling factor $\hat{C}_0'$, which can be passed to the topography generator, in order to generate profiles whose analytical RMS parameter have a given value (cf. Equations (2.35) and (2.36)). By doing so, the scaling factor is uniquely determined and does not depend on the discretization of the derivative or integral operations. The same procedure can be applied to self-affine rough surfaces, whose discrete spectrum writes

$$\hat{\Phi}\left[\boldsymbol{k} = \left(\frac{q}{M}\Omega_{s_y},\, k = \frac{p}{N}\Omega_{s_x}\right)\right] = \begin{cases} \hat{C}_0 & , \quad k_l \le \|\boldsymbol{k}\| < k_s \\ \hat{C}_0\left(\dfrac{k_r}{\|k\|}\right)^{2(H+1)} & , \quad k_r \le \|k\| \le k_s \\ 0 & , \quad \text{elsewhere}\,, \end{cases} \tag{5.61}$$

and the spectral properties similarly come

$$m_{00} = \frac{l_{s_x} l_{s_y} \hat{C}_0 k_r^2}{2\pi}\left(\frac{1-\xi^2}{2}+\frac{1-\zeta^{-2H}}{2H}\right); \tag{5.62a}$$

$$m_{20} = \frac{l_{s_x} l_{s_y} \hat{C}_0 k_r^4}{4\pi}\left(\frac{1-\xi^4}{4}+\frac{\zeta^{2-2H}-1}{2-2H}\right); \tag{5.62b}$$

$$m_{40} = \frac{3 l_{s_x} l_{s_y} \hat{C}_0 k_r^6}{16\pi}\left(\frac{1-\xi^6}{6}+\frac{\zeta^{4-2H}-1}{4-2H}\right); \tag{5.62c}$$

$$\alpha = \frac{3}{2}\left(1-\xi^2+\frac{1-\zeta^{-2H}}{H}\right)\left(\frac{1-\xi^6}{3}+\frac{\zeta^{4-2H}-1}{2-H}\right)\Big/\left(\frac{1-\xi^4}{4}+\frac{\zeta^{2-2H}-1}{1-H}\right)^2 . \tag{5.62d}$$

In recent numerical investigations, surfaces with RMS slope around 0.1 have been employed, which is verified to be physically reasonable (Yastrebov, Anciaux, *et al.*, 2012, 2015; Pei *et al.*, 2005). In this work, rough profiles are normalized by the previous procedure in order to fix the RMS slope at 0.2.

Materials

The deformable rough block is modeled with an elastic constitutive model, and the numerical values of the properties selected from steel's conventional elastic properties—these values were chose inasmuch that to keep a practical point of view consistent with the experimental roughness measurements presented in Chapter 3— mostly steel components. As for the rigid block, the *infinite stiffness* is approximated by setting a large value for the elastic modulus, compared with the equivalent property of the rough block. Furthermore, for completeness, it should be mentioned that $\nu_2 = 0$, for all the examined cases. Table 5.1 shows the elastic properties of both bodies. The chosen value Young modulus for the flat body was verified to satisfy both the stiffness criterion, i.e., that displacements are virtually zero in this body, and also to provide acceptable numerical stability in the simulations.

Table 5.1: Elastic properties of the flat and rough block, used throughout the tests.

Body	E /GPa	ν
Rough (1)	210	0.3
Flat (2)	5000	0.0

Boundary conditions

With regard to the boundary conditions, several instances must be referred. The mortar and non-mortar boundaries, where the contact boundary conditions apply, are set to the upper face of the flat block and the rough boundary of the rough block. By doing so, the values of the Lagrange multipliers at the rough contact are explicitly obtained in the solution, which allows for a direct treatment of the pressure distribution at this boundary. The bottom face of the flat block, here termed $\partial\Omega^{\mathrm{m}}_{\mathrm{fix}}$, is fixed

$$\boldsymbol{u}(\boldsymbol{x}) = \boldsymbol{0}, \quad \text{for } \boldsymbol{x} \in \partial\Omega^{\mathrm{m}}_{\mathrm{fix}} \,. \tag{5.63}$$

Since the micromechanical rough contact is intended to represent a particular feature of a macroscopic contact situation, occurring, typically, at considerably different scales, periodic boundary conditions are adopted for the side faces of the rough block. Note that this is admissible as long as the micromechanical problem is solved in a small scale compared with the macroscopic contact. This can be thought, for example, as the contact of nominally flat surfaces, or as if the micromechanical problem is contained within the nominal contact area of an Hertz contact problem. Rough topography must also be periodic, in order to have a consistent formulation of this boundary condition. This issue is easily handled at this stage, since the implemented rough topography generator is naturally fit to generate periodic topography (see Chapter 3). If such condition is not met, an artificial strategy should be adopted in order to guarantee periodicity, such as spline reconstruction (Wagner, Wriggers, Veltmaat, *et al.*, 2017). The periodicity implies the equality of the displacements at matching points in the positive and negative boundaries, respectively $\partial\Omega^{\mathrm{s}}_{+}$ and $\partial\Omega^{\mathrm{s}}_{-}$, and writes

$$\boldsymbol{u}(\boldsymbol{x}_{-}) = \boldsymbol{u}(\boldsymbol{x}_{+}), \quad \text{for } \boldsymbol{x}_{+} \in \partial\Omega^{\mathrm{s}}_{+}, \text{ and } \boldsymbol{x}_{-}(\boldsymbol{x}_{+}) \in \partial\Omega^{\mathrm{s}}_{-} \,, \tag{5.64}$$

and that the surface traction must be antisymmetric at the same matching points in both boundaries

$$\boldsymbol{t}(\boldsymbol{x}_{-}) = -\boldsymbol{t}(\boldsymbol{x}_{+}), \quad \text{for } \boldsymbol{x}_{+} \in \partial\Omega^{\mathrm{s}}_{+}, \text{ and } \boldsymbol{x}_{-}(\boldsymbol{x}_{+}) \in \partial\Omega^{\mathrm{s}}_{-} \,. \tag{5.65}$$

At the top boundary of the rough block, herein called exterior boundary $\partial\Omega^{\text{s}}_{\text{ext}}$, all points have the same vertical displacement $\bar{u}_3$, which is not prescribed and is part of the solution. In order to prevent rigid body motions, which would occur for slight force unbalances in the plane directions, an arbitrary point $\boldsymbol{x}_{\text{fix}}$ (node of the finite element mesh) is blocked in these directions—it can only move in the vertical direction. Additionally, a uniform pressure p_0 is applied on this boundary, following an incremental strategy. At each load step, the equilibrium equation is solved until the final load is reached. The specification of load steps from the beginning of the simulation requires that contact must exist at the initial configuration. This is accomplished in a simple preprocessing step, where the rough geometry is translated in order to close the gap at the maximum summit/peak of the topography—which is trivial, because one of the contacting boundaries is flat. Otherwise, displacement steps should be specified until the active set is non-null, and from there on pressure steps can be applied (Wagner, 2018). The 2D simulations are regarded as plane strain cases—the rough profile can be tough as taken across the lay of a strong anisotropic surface.

Remark 5.3 on the boundary conditions at corners of the non-mortar boundary. *Some points belonging to the non-mortar boundary are also part of the periodic boundaries. Since the contact specification within the dual mortar method is formulated in a vectorial fashion, i.e., one works directly with the Lagrange multiplier vector (surface traction), the constraints for all degrees of freedom of a non-mortar point are inherently included. Therefore, if in addition to the contact boundary condition, a periodic boundary condition was prescribed at these corner points, the problem would be overconstrained, and no solution could be found. A physically reasonable set of boundary conditions would be the periodicity of the plane displacements, and contact on the vertical displacement. The displacement of the corner nodes would be necessarily equal in the 2D problem, since they would become active at the same time. Nevertheless, this would require a slight deviation from the original formulation, and considerable changes to the currently available implementation. This issue is often overlooked in the literature, and practically no references exist regarding it, possibly because it does not seem to impact the results considerably. Here, the contact boundary condition prevails relative to the periodic conditions, which are discarded at these points.*

5.3.2 Preprocessing and mesh generation

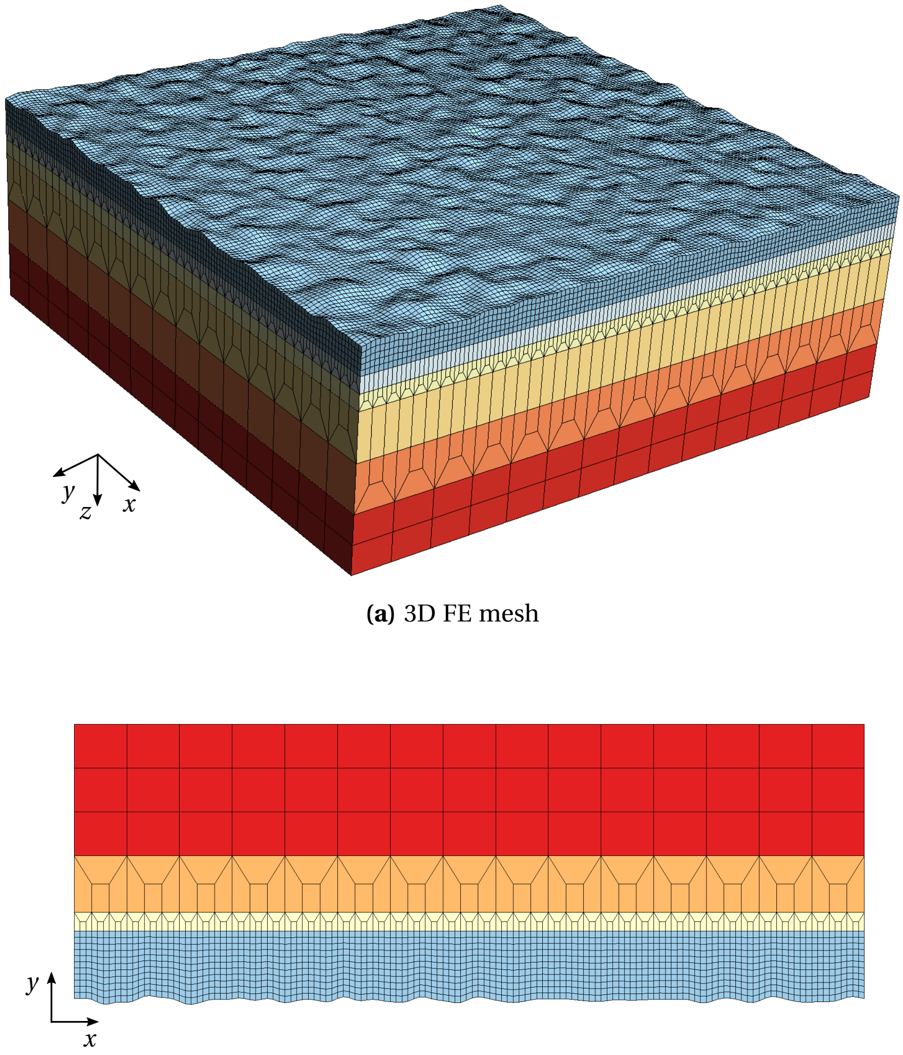
The fundamental preprocessing steps preceding the rough contact analysis with the FEM are the surface topography discretization and finite element mesh generation. The former can be obtained either from experimental measurements or by the application of a randomly rough topography generation algorithm. Naturally, the second alternative is adopted here, as already discussed thoroughly in Chapter 3. The topography generation algorithm provides the coordinates of the discrete points in a vector, or structured grid. This list of coordinates is then fed into the mesh generator together with the rough block height, in order to generate the required finite element mesh.

Rough contact is primarily an interface problem, then all relevant physical phenomenon occur in a relatively thin layer near the surface, requiring a fine discretization. In opposition, the height of the rough block subtract must be sufficiently large in order to model the mechanical response of the bulk properly and, thus, provide enough stiffness to the block. In regions close to the exterior boundary, the effect of roughness decreases, and only bulk phenomena are considerable. Therefore, for the case of micromechanical contact, the FE mesh can coarse with increasing distance to the rough boundary, without any significant harm to the results. With the purpose of reducing the computational costs in each simulation, a mesh transition strategy shall be adopted, on which the size of the finite elements increase as one moves away from rough boundary. This can be accomplished simply by applying a gradient in the element characteristic length along the height of the block, producing a unstructured grid of irregularly shaped elements. Regular transition schemes with well defined geometries are often employed, such as in Stupkiewicz (2007), Yastrebov (2011), and Yastrebov, Durand, *et al.* (2011). Systematic combinations of finite elements with different geometries are used in these works, in order to merge a cell of 9 elements with a single element, with larger dimensions, in 3D problems. Also transition from cell of 4 elements to only 1 are employed. Transition strategies for 2D problems are readily formulated from the more complex 3D versions, and may allow the reduction of the number of elements by 3 or 2 in a single transition layer.

In this contribution, a variation of the strategy used in Stupkiewicz (2007) and Yastrebov, Durand, *et al.* (2011) and was implemented, and is illustrated for both 2D and 3D meshes in Section 5.4.2. A high resolution region is defined near the rough boundary, on which a structured FE mesh of quadrilateral elements is generated. Several layers of elements are generated with a transfinite interpolation between the rough boundary and a reference flat plane, within this region. The height of the elements is slightly larger than their width, in order to maintain an approximately square shape during the compressive deformation. Then, a certain number of transition layers reduces the number of elements by a factor of 3 at each level. The remaining region is filled with a conforming coarse structured grid. For the 3D case, the transitions layers must be applied at different directions successively, so that the aspect ratio is approximately preserved after the transition. This meshing strategy enables, on one hand, the reduction of the number of elements compared with a uniform grid with high resolution, and on the other hand, the periodic boundary conditions are easier to enforce, since there are matching nodes at both the positive and negative boundaries—otherwise, mortar methods should be applied to connect the non-conforming meshes.

The open source software `Gmsh`, by Geuzaine and Remacle (2009), was used for the establishment of the numerical framework, as the mesh generator. `Gmsh` provides an *Application Programming Interface* (API) for several programming languages, including `Python`. This facilitates the definition of the operations flow, by allowing to perform the mesh generation of the rough topography within `Python` itself, in a single sequence of instructions. The implementation of the current meshing scheme is based on the definition of different transfinite interpolation curves, surfaces and volumes, within `Gmsh` self philosophy. For further insights into the functionalities of `Gmsh`, the respective manual should be consulted in Geuzaine and Remacle (2019).

Since only elastic material laws are considered, both blocks in 2D are meshed with four-node quadrangular bilinear elements (QUAD4-FBAR), and for the 3D scenario, 8-



(a) 3D FE mesh

(b) 2D FE mesh

Figure 5.9: Example of the finite element meshes of the rough blocks generated by `Gmsh`. These serve mainly to illustrate the mesh transition strategy, and are not necessarily representative of the one used in the course of this work.

node hexahedral elements are used (HEXA8-FBAR). The *F-bar* finite element technology is used in this work, in order to prevent *volumetric locking* typically associated with low order standard finite elements in large strain formulations, and to guarantee accurate results (Souza Neto *et al.*, 1996). Such spurious locking of the solution could also be eliminated by considering high order elements, however, the inherent simplicity of low order elements is predominantly attractive. The *F-bar* methodology consists in splitting the deformation gradient into an isochoric (volume preserving) component and a volumetric (purely dilatational) contribution. The isochoric component is computed in the Gauss point where the stress tensor is to be established, and the volumetric component is computed at the *centroid* of the element. Regarding the mortar and non-mortar interface, the same mesh size is prescribed at both, in order to improve accuracy and stability of the numerical method.

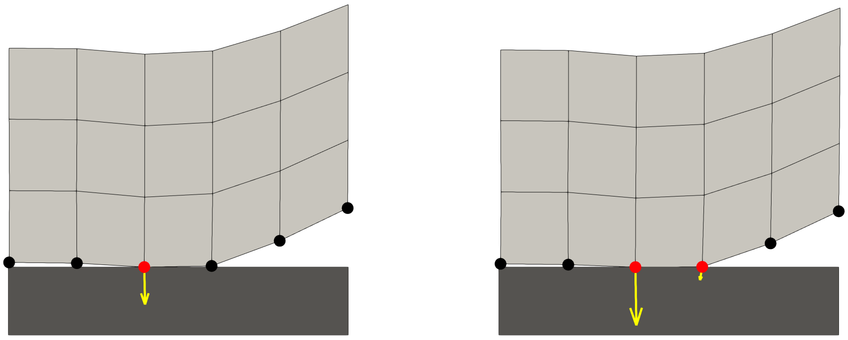
Remark 5.4 on the smoothness of the numerical mesh.
The finite element mesh at the rough boundary must be smooth enough, in order to favor the application of the FEM. One could be tempted to generate a coarse surface discretization with the topography generator algorithm, and then generate a fine FE mesh over it, by specifying spline interpolations between the discrete points. This procedure, however, distorts the topography PSD, carefully preserved during the numerical generation of the topography. In order to maintain a consistent philosophy across the numerical framework, only points resulting from numerical generation procedure shall be used as nodes for the FE mesh. The smoothness of the topography and numerical mesh shall be uniquely guaranteed by the topography features, namely, the cut-off wavelengths λ_l and λ_s, and not forced with numerical interpolation.

5.3.3 Numerical computation of the real contact area

Another fundamental aspect of the present work regards the numerical evaluation of the real contact area. This quantity is to be computed at each load step from the finite element results, and can follow either a geometrical or physical argument. From a geometrical perspective, the contact area can be computed simply as the fraction of nodes with non-zero contact pressure. A more robust scheme considers that when two consecutive nodes are active, the area bounded by those active nodes is part of the real contact area. The foundations of this methodology are arguably stronger, since it computes, in fact, the discrete contact area of the numerical model. However, points which are almost entering the active set, and which are not already part of it because the discretization does not capture the continuous growth of contact clusters, are neglected in the results. In order to account for these situations, the real contact area can be computed from a equilibrium-based physical argument. From the definition of the mean contact pressure, it follows that the real contact area fraction is the ratio between the nominal exterior pressure (at each load increment) and the arithmetic mean of the Lagrange multipliers, i.e., the contact pressures.

While the geometric argument is based uniquely on the current active set, the scheme relying of Lagrange multipliers, which result directly from the dual mortar formulation, have incorporated information of the neighboring elements. For example, a contact patch constituted by a single element with 2 active nodes does not include the information that one node may be compressed more than the other and, therefore, a contribution for the real contact area may be discarded. However, it should be noted that the Lagrange multipliers are not necessarily vertical, due to the averaged normal vector field used within the mortar integration scheme. Therefore, in the contact area computed from equilibrium considerations, one may sum the magnitude of Lagrange multipliers that are not necessarily collinear, which might introduce errors in the numerical approach. This aspect is better illustrated in Figure 5.10. At a first moment, only one node is active, and by the geometrical principle, the contact area is null, while from the equilibrium based alternative, the contact area is not zero. Only when the neighbor node becomes active, the contact spot, as thought from a geometrical perspective, grows, even though the value of the Lagrange multiplier at that node is very small. The geometrical contact area can be computed with a finite resolution related with the mesh size, and is not sensible to the value of the contact normal pressure.

In principle, the equilibrium based approach provides a richer information of the contact status, having in mind the downside related with the non-vertical Lagrange multiplier vectors. It is not clear what method provides more realistic results. This issue shall be resumed regarding multiscale analysis, where some extra conclusions can be extracted. For the following analysis, the geometrical argument is adopted.



(a) 1 active node

(b) 2 active nodes

Figure 5.10: Illustration of the contact area increments in the numerical model. The active nodes are represented by the red points, and the inactive set by the black points. The Lagrange multiplier vector is denoted by the yellow arrows. When only one is active, the geometrical contact area is null, but in the respective *continuum* model the contact area would not be zero. From the geometrical perspective, the contact spot grows only when the neighbor node becomes active, even though the value of the Lagrange multiplier is very small.

5.3.4 Numerical framework

The general FE-based approach to rough contact combines all the building blocks described previously in a single workflow, see Figure 5.11. Within a Python master script, a discrete topography is generated with the algorithms discussed in Chapter 3. The coordinates of the randomly generated points are used directly as the finite element mesh nodes. The full geometry of both the rough and flat block is discretized in finite elements with Gmsh, via its Python's API. A *.msh file is produced by Gmsh, which is then parsed in order to extract the nodal coordinates, table of connectivities and identify the nodes under different boundary conditions. Before entering the automatic input file generation routine, the order of the nodes in the table of connectives must be corrected to match that implemented in LINKS. Following the previous sequence, input files for different topography features, or even for different realizations of the same topography characteristics (by varying the seed passed to the random number generator), can be readily generated. Several RCE are run in parallel, in order to profit from the computational resources at disposal. The contact area fraction and list of nodal contact pressures at each load increment are given by LINKS. These output files are further post-processed and used for plotting, again within a Python environment.

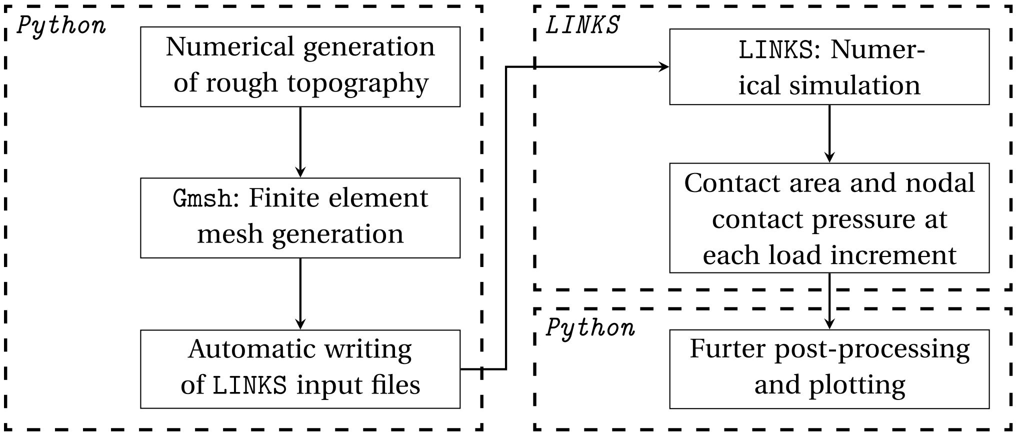

Python
Numerical generation of rough topography
Gmsh: Finite element mesh generation
Automatic writing of LINKS input files
LINKS
LINKS: Numerical simulation
Contact area and nodal contact pressure at each load increment
Python
Furter post-processing and plotting

Figure 5.11: General framework of numerical tools used to process rough contact.

5.4 Definition of a statistically Representative Contact Element

The periodic boundary conditions intend to restrict the simulation of a large rough surface to a single contacting element. This *Representative Contact Element* (RCE) shall provide a statistically representative mechanical response of the system. Even though the issue of representativeness for rough contact has been investigated in the literature, general rules for the definition of a RCE are still not well-established. The best example of this type of studies can be found in Yastrebov, Anciaux, *et al.* (2012), who performed numerical studies on *Representative Self-affine Surface Element* (RSSE) within a BEM framework. The analysis has been based on both the height distribution and on the contact area evolution curves for self-affine rough surface, with different topography parameters. In the context of the FEM, representativeness studies, namely on the influence of the height of contacting blocks, were performed in Temizer and Wriggers (2008) and De Lorenzis and Wriggers (2013), yet on slightly different contexts. Additionally, representativeness analysis of microscale problems for differer research areas are commonly found, e.g., on RVE's for polycrystalline aggregates (Vieira, 2018).

Aiming at establishing the rules for the definition of a 2D RCE for a self-affine topography, several parametric investigations have been carried by varying different fundamental RCE parameters, unknown beforehand. The goal of the ongoing discussion is the determination of RCE characteristics, such as length and height, which will allow the computation of a statistically representative contact area response from a rough topography with given Hurst exponent H and cut-off wavelengths λ_l and λ_s. The only fixed topography parameter for all cases is the RMS slope, which is set to $\sqrt{m_2} = 0.2$. The material properties are fixed, as well (see Section 5.3.1). This value for the RMS slope, guarantees that the generated rough topographies are fairly smooth, therefore prone to the application of the FEM, while staying within the range of physically reasonable values. Additionally, due to the inherent statistical nature of the rough contact, 10 different topography realizations are considered for each case. The RCE representativeness tests encompass the following investigations

- Mesh convergence;
- Length of the RCE;
- Height of the RCE;
- Influence of the number of realizations.

5.4.1 Initial estimation of the RCE dimensions

The dimensions of the rough block shall necessarily be estimated before starting any of the previously mentioned analysis. These must already be closely representative, i.e., in excess relative to the minimum required—in the representativeness point of view. The numerical experiments must be carried with such excessive values, in order to assure that the conclusions extracted for each case will still be valid for an RCE with minimum requirements. At a certain point during the analysis, these oversized parameters will be allowed to relax, until divergence from the representative response is observed, leading to the establishment of a minimum requirement for the RCE.

Three main variables must be estimated at this stage, viz., the rough block length L (relative to the long cut-off wavelength λ_l), the height of the rough substrate (rough block) H_{sub} and the height of the refined, high resolution region of the finite element mesh, near the non-mortar interface, denoted by H_{ref}. The starting iteration for the length of the RCE can be determined based on the results of Yastrebov, Anciaux, *et al.* (2012, 2015), and also recalling the tests on the Gaussianity of the artificially generated rough topographies, in Section 3.3.3. Looking first at latter statistics study carried in this thesis, it was verified that the Gaussian generator produces normally distributed heights only when the ratio L/λ_l is sufficiently high. By visual inspection of Figure 3.8, it can be stated that a minimum of $L/\lambda_l = 4$ is required in order to obtain reasonably Gaussian topographies. Similar results are reported in the aforementioned publications, yet these authors consider a smaller tolerance for the error on the heights distribution, relative to a Gaussian reference, and claim that the minimum ratio is $L/\lambda_l = 16$. The value $L/\lambda_l = 4$ provides a good trade-off between accuracy and computational convenience and, thus, will be considered the starting lower threshold for the RCE length.[4]

Focusing now on the rough substrate height and the extent of the high resolution mesh, there are no results available in the literature regarding self-affine rough topography and, therefore, these shall be determined based on preliminary numerical results. The rough

[4] If the ratio proposed by Yastrebov, Anciaux, *et al.* (2012, 2015) was adopted, a large limitation would be set on the bandwidth (λ_l/λ_s) range considered for the analysis. The size of the finite element problem increases rapidly with the ratio L/λ_s, then if L/λ_l is already large, only small values for λ_l/λ_l could be considered, such that a large number of simulations could be run and analyzed, in the mean time.

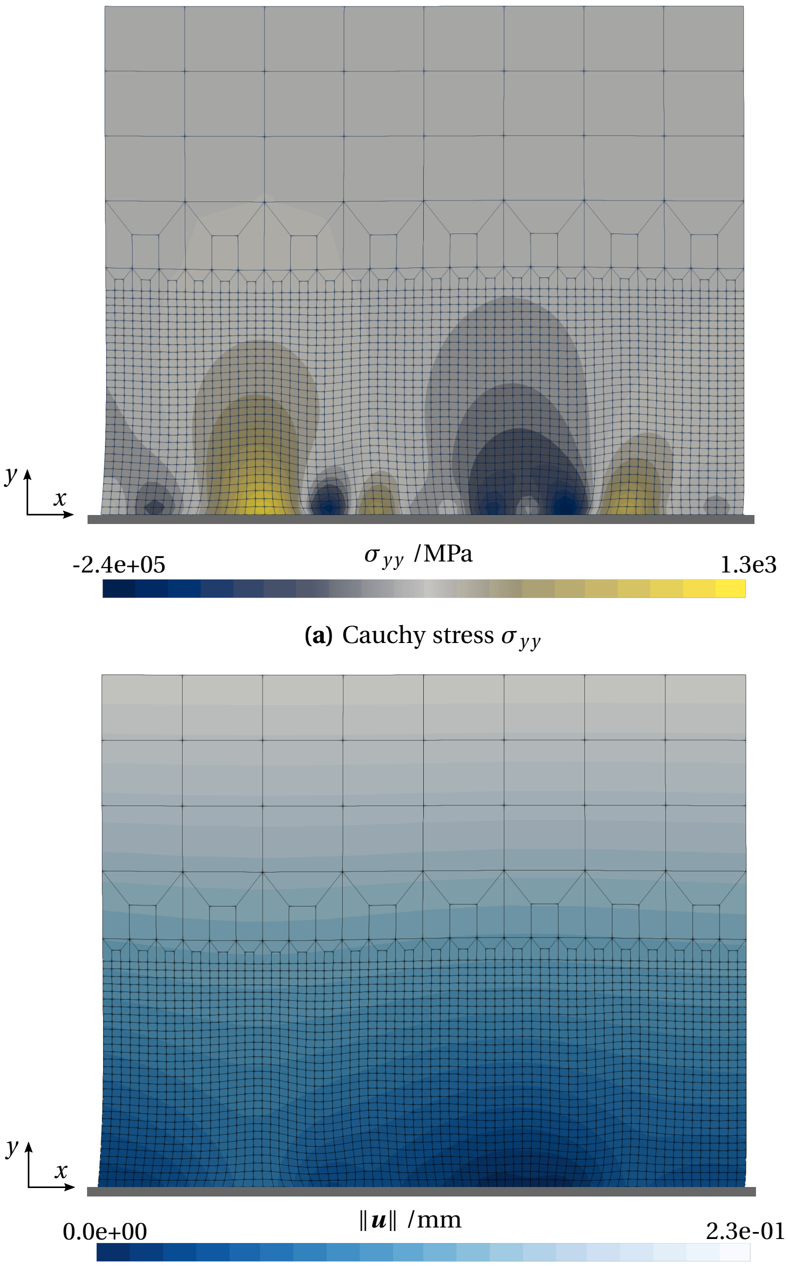
block must be sufficiently high, so that it represents accurately the stiffness of the bulk of the material. Furthermore, it must give room for the boundary layer of stress, strain and displacements to develop near the rough boundary, and converge to the bulk field far from this boundary. In fact, the rough topography shall only influence results in a thin layer near the contact interface, and at distant points its presence shall not be felt. Away from the rough boundary, the RCE behaves simply like a block in compression, with uniformly distributed stresses in the cross sections, and the displacement field varies linearly with the distance from the contact interface.

As initial estimates for these dimensions, the height of the substrate was set to the length of the block, and the height of the high resolution mesh was set as 80 times the RMS height of the profile. The assessment of the validity of these values was made by performing some numerical simulations, and evaluating the results by visual inspection. For the simulations, the rough block length was set to 5 mm, the cut-off ratios follow $L/\lambda_l = 4$ and $L/\lambda_s = 32$ and $H = 0.8$. The rough boundary was discretized with 288 elements. A small ratio for L/λ_l and a short bandwidth were chosen in order to increase the RMS height, such that it poses harsh conditions for the estimate values of the heights. The RCE is loaded up to nearly full contact conditions, where the stress and displacement gradients are the most considerable.

The numerical results for the Cauchy stress σ_{yy} and magnitude of the displacement vector $\|\boldsymbol{u}\|$ are shown in Figure 5.12. Only a region with one fifth of the total height (length of the block) is shown. It can readily be seen that all stress gradients are inside the high resolution region, within a fairly large margin. The stress peaks are concentrated very near the rough boundary, and the stress field quickly merges into a unique stress value. The same happens with the displacement field, where all relevant variations are captured by the fine mesh, and a linear evolution can be observed after the transition layers. It should be remarked that the displacement field at the flat base is zero. For completeness, the previously discussed fields are shown in full extent in Figure 5.13.

In conclusion, by setting the rough block height equal to its length, and the high resolution region to 80 times the RMS height, the large gradients near the boundary are captured correctly, and the solution for the bulk of the body is approached, both in terms of the material (stress) and kinematical (displacement) quantities. Until any word in contrary, these values shall be assumed for the following numeral tests.

Remark 5.5 on the order of magnitude of contact stresses.
Inspecting Figure 5.12 closely, it can be observed that compressive Cauchy stresses σ_{yy} around 2.6×10^5 MPa exist neat the contact interface, in full contact conditions. Intuitively, at this level of elastic stress, plastic deformation should necessarily be included, in order to provide an accurate and realistic description of the phenomenon. In the following, one sticks to the purely elastic material law, independently of the practical applicability of the results. This issue is no longer commented in the remaining sections, and the physical validity of the results must be addressed in future works.



(a) Cauchy stress σ_{yy}

(b) Magnitude of displacement vector $\boldsymbol{u}$

Figure 5.12: Stress and displacement field in a region of rough block, with initially estimated dimensions. A portion of the block with approximately only one fifth of its height is represented. The opacity of stress values around the mean value are reduced in order to enhance the visual perception of the extreme values and nearby gradients. The stress field reduces to a constant value for points sufficiently far from the rough boundary, where the displacement field is observed to change linearly with the distance to contact interface.

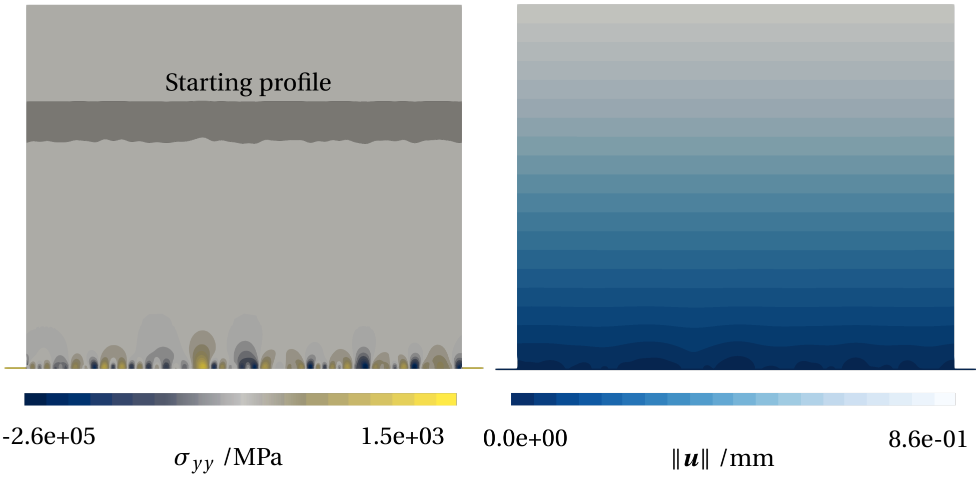


Figure 5.13: Full field representation of the Cauchy stress σ_{yy} and magnitude of the displacement vector, for the RCE with the initially estimated values for the length, substrate height and high resolution region.

5.4.2 Mesh convergence

Within a FE mesh, distinct frequency contributions are discretized with different levels of quality. Intuitively, the large wavelengths are smoother than the short wavelengths, for a given mesh size—there are more elements in each period of the large wavelength harmonics. Thus, the shortest wavelength in the topography is the most poorly discretized one. As the mesh step Δx decreases, i.e., the number of elements in each period increases, for every frequency, it is conjectured that there is a point where improvements on the results due to the discretization are residual.

In the present section, the convergence of the contact area evolution curves with progressively finer meshes is addressed, regarding different topography cases. In order to measure the quality of the mesh relatively to the topography features, the *number of nodes per asperity* is used to provide a quantitative description of the former variable. The minimum wavelength that can be resolved by a discrete profile is equal to two times the mesh spacing $2\Delta x$ (see Appendix A). In that case, the harmonic is approximated only by two nodes in each period, resulting in a *one node per asperity* scheme—λ_s is equal to $2\Delta x$. The *minimum* number of nodes in each asperity can then be written as

$$\text{Minimum number of nodes } \textit{per} \text{ asperity} \equiv \frac{\lambda_s}{2\Delta x}\,. \tag{5.66}$$

The main goal of the mesh convergence study is to establish the minimum value of $\lambda_s/2\Delta x$ that guarantees a converged area-load curve. For that purpose, four combinations of $H \in [0.40, 0.8]$ and $L/\lambda_l \in [4, 8]$ were considered. For each pair, the study comprised four different values of the ratio $L/\lambda_s \in [32, 64, 128, 256]$ (the bandwidth can be determined combining both ratios L/λ_l and L/λ_s), and for each ratio, four different levels of discretization were tested. Meshes with approximately 1, 2, 4 and 8 nodes *per* asperity were used in the numerical simulations. The length of the block was set to 5 mm, and for each set of topography characteristics, 10 different RCE realizations have been tested. The final result is determined by averaging the curves of all realizations. In Figure 5.14, two successive levels of discretization of one of the tested RCEs are shown, namely, regarding the cases concerning one and two nodes *per* asperity. The number of non-mortar elements used in each case can be found in Table 5.2. Note that the number of nodes *per* asperity is prescribed in an approximate fashion, since the number of non-mortar elements must be a multiple of 3, in order to apply the FE mesh transition layers.

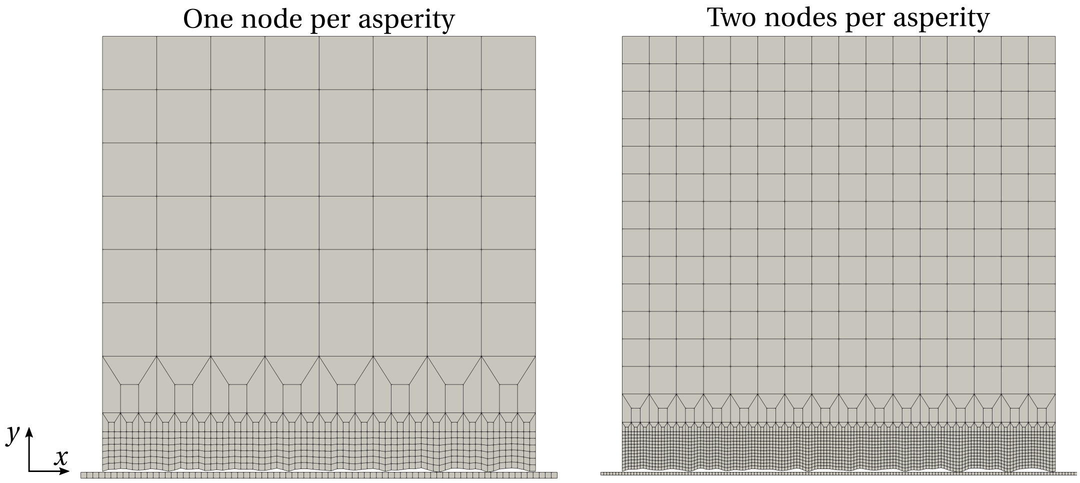


Figure 5.14: Two finite element meshes with different resolutions, used in the mesh convergence study.

Table 5.2: Number of non-mortar elements for different ratios L/λ_s and levels of discretization, used for the mesh convergence tests.

	L/λ_s			
$\lambda_s/2\Delta x$	32	64	128	256
1	72	135	270	567
2	144	270	540	1134
4	288	540	1080	2268
8	576	1080	2160	4536

The results of the mesh convergence numerical tests are plotted in Figure 5.15. The horizontal axis refers to the nominal external pressure, normalized by the RMS slope and the effective Young modulus, and the vertical axis refers to the real contact area fraction. The markers represent the average of contact areas across all realizations, and the error bars measure one standard deviation for each side. The inset plot emphasizes the curves for small yet physically reasonable loads and contact area fractions. Only the results for $H = 8$ and $L/\lambda_l = 8$ are presented, since the curves for the three remaining pairs are similar, and the conclusions extracted from Figure 5.15 are equally valid.

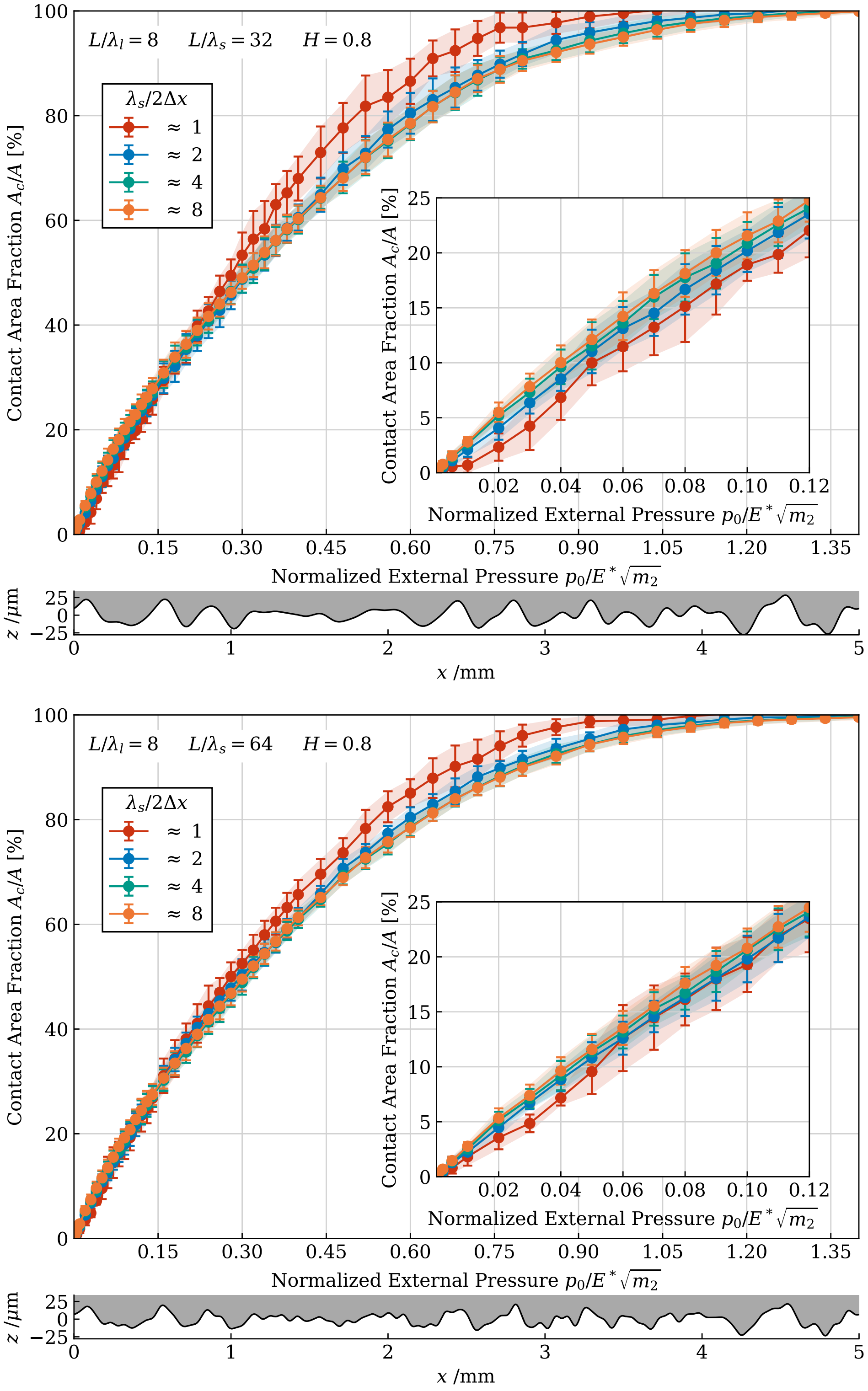


Figure 5.15: Mesh convergence study on the contact evolution curve, for different bandwidth ratios. The Hurst exponent is fixed at 0.8, the RMS slope at 0.2 and $L/\lambda_l = 8$.

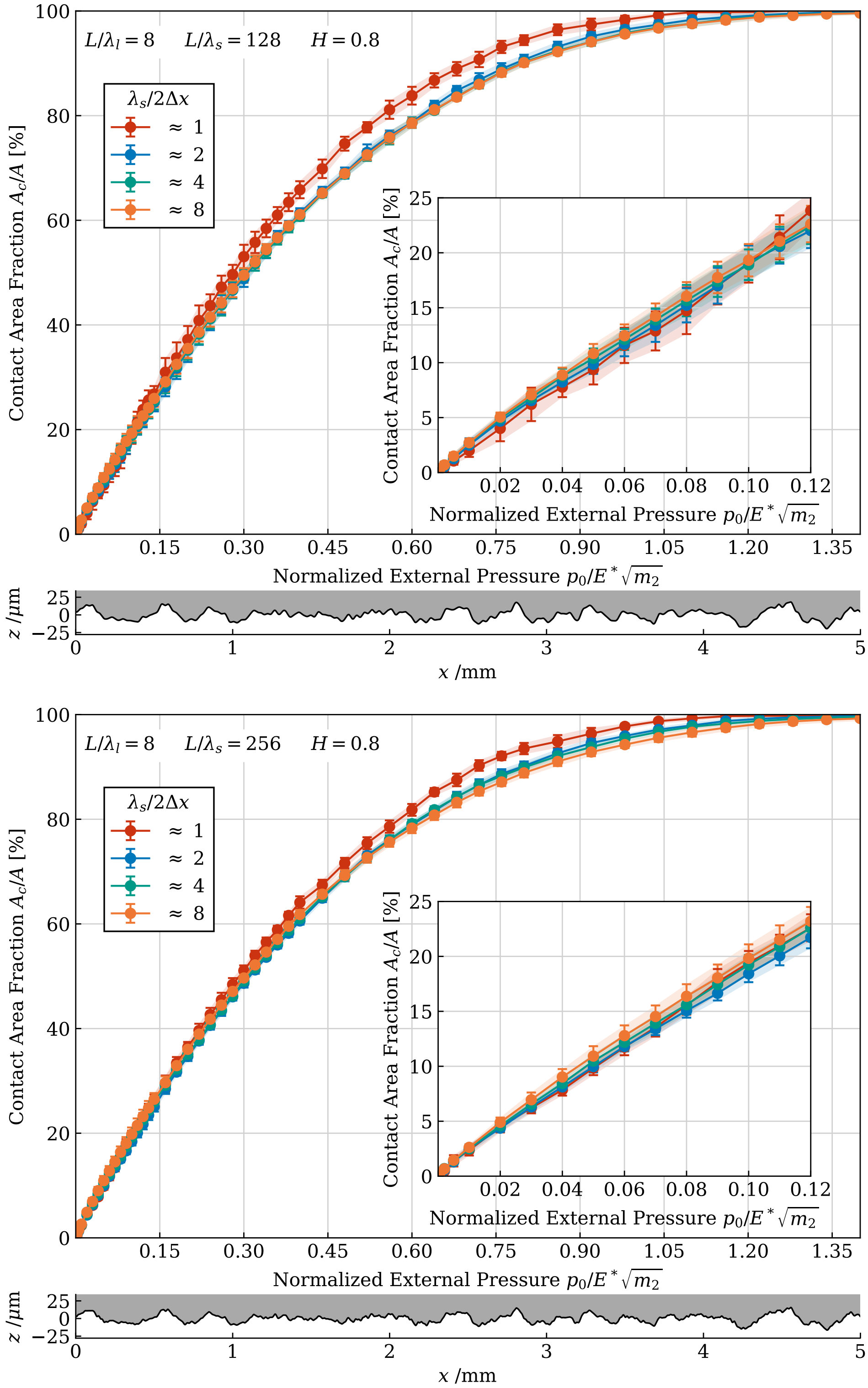


Figure 5.15: Mesh convergence study on the contact evolution curve, for different bandwidth ratios. The Hurst exponent is fixed at 0.8, the RMS slope at 0.2 and $L/\lambda_l = 8$ (continued).

A striking similarity in the area-load curves for all bandwidths is the deviation of the results for the one node *per* asperity scheme ($\lambda_s/2\Delta x = 1$). In fact, all the curves are relatively close to each other with the exception of the aforementioned case. This effect is visible both at small and large loads, and the maximum absolute gap between the curves occurs between 0.45 and 1.05, in the normalized external pressure axis. Some landmark works on elastic contact, such as by Hyun, Pei, *et al.* (2004) and Hyun and Robbins (2007), preserved the self-affine nature down to the discretization scale—i.e., keeping a minimum of one node per asperity. Yastrebov, Anciaux, *et al.* (2012) mentioned that such poorly discretized topography could not accurately resolve the mechanics of the problem. The results presented in this work corroborate such claim, once the curve for $\lambda_s/2\Delta x = 1$ predicts different mechanical response relative to the other mesh sizes.

By increasing the number of elements in the topography, the overall convergence of the curves is observed, for all tested bandwidths. The curve referring to 4 and 8 nodes *per* asperity are almost coincident at every point. For the largest ratio $L/\lambda_s = 256$ (bandwidth $\lambda_l/\lambda_s = 32$), however, some small differences between the curves for $\lambda_s/2\Delta x = 4$ and $\lambda_s/2\Delta x = 8$ can be detected, in opposition to the previous cases, where the results were practically coincident. Nevertheless, the observed gap is very narrow, and convergence can also be accepted for this case. It should be remarked that the case $L/\lambda_s = 256$ with 8 nodes *per* asperity was only tested for $H = 0.8$ and $L/\lambda_l = 8$, due to extremely large simulation times. Each realization of such cases took, on average, 12 hours to complete.

The increase of the number of elements in the contact interface is followed by a decrease of the standard deviation, for every bandwidth—which also contributes to the representativeness of the RCE. This is intrinsically associated with the geometrical scheme used for the numerical evaluation of the contact area (cf. Section 5.3.3). The decrease on the mesh spacing reduces the minimum increment of contact area and, therefore, also the relative standard deviation.

Referring some brief qualitative observations, it should be noted that the variation of real contact area with external pressure is nearly linear up to 25%-30% of the real contact fraction. From that point onward, the behavior is noticeably nonlinear, and full contact is reached when the normalized external pressure reaches approximately 1.4.

All in all, it can be stated that the area-load converges with increasingly finer meshes. When each asperity is discretized by at least 4 nodes it can be assumed that the area response has converged. Eventually, this conclusion can be questioned for $L/\lambda_l = 256$, due to slight deviations between the results for 4 and 8 nodes *per* asperity, and further investigations shall be designed in order to clarify this issue. All the points observed before were verified to hold true for the remaining combinations of H and L/λ_l, and these are not plotted here for simplicity.

5.4.3 Length of the block

The length of the RCE can be interpreted as the low frequency counterpart of the mesh size, regarding RCE representativeness. In the mesh convergence test, focus has been placed on finding the minimum number of elements required to discretize the highest frequency of the spectrum. For the study on the influence of the RCE length, the minimum number of periods contained in the RCE which gives a representative mechanical response is addressed. Undoubtedly, the largest wavelength harmonic λ_l is the contribution with less periods *per* RCE, then it shall be used to parametrize the following investigation. This issue reports to the numerical tests on the Gaussianity of numerically generated rough profiles, as mentioned in the beginning this section. It was verified that the length of the rough block must be at least four times (approximately) the long wavelength cut-off, such that the heights distribution is seemingly Gaussian. This topic is now assessed based on the contact area curve for several RCE realizations, with different topography characteristics.

The long cut-off wavelength was fixed at $\lambda_l = 5\,\mathrm{mm}$, and four combinations of Hurst exponents $H \in [0.4,\ 0.8]$ and bandwidth ratios $\lambda_l/\lambda_s \in [4,\ 16]$ were tested, in order to explore both the influence of the Hurst exponent and roughness spectrum in the results. For each case, RCEs with four different lengths $L/\lambda_l \in [1,\ 4,\ 8,\ 16]$ were generated. With the purpose of minimizing the effect of resolution on the numerical computation of the contact area associated with different levels of discretization, for the same RCE length L/λ_l, both profiles concerning $\lambda_l/\lambda_s = 4$ and $\lambda_l/\lambda_s = 16$ are meshed with the same number of elements in the non-mortar boundary. By doing so, it is assured that the same discrete frequencies exist in all cases. The number of elements at the non-mortar interfaces was chosen in order to verify the 4 nodes *per* asperity criterion established in the previous section, regarding the topography with $\lambda_l/wl_s = 8$—the other topography case, being meshed with the same number of elements, this condition is necessarily verified, as well. The results from ten RCE with different topography realizations were averaged, for each unique set of topography parameters.

Figure 5.16 shows four distinct RCE rough boundaries, for the different values of lengths tested, regarding the case $\lambda_l/\lambda_s = 4$ and $H = 0.8$. It can readily be seen that the block for $L/\lambda_l = 1$ can intuitively be embedded in any other topography, or in other words, that a similar pattern to that of $L/\lambda_l = 1$ can be found in longer RCEs. In fact, Figure 5.16 shows that by increasing the RCE length, a wider variety of geometrical features is added to the topography, such that the particular effect of each one is averaged out. The problem consists in establishing what length is enough to introduced a sufficiently wide range of distinct geometrical features.

The results of the finite element analysis are presented in Figure 5.17, for $H = 0.8$ (again, these are similar for $H = 0.4$). The most eye-catching difference between the two graphs is the difference between the results between the two spectrum bandwidths. The overall standard deviation, but specially that associated with $L/\lambda_l = 1$, is noticeably larger for the shorter spectrum $\lambda_l/\lambda_s = 4$. This may precisely owe to the previous discussion around Figure 5.16. Different rough profiles configurations can verify the input PSD, and by modeling each one individually, the variability of the mechanical response increases. With increasing RCE length, the standard deviation of the contact area is drastically reduced, since each realization already contains a wide variety of geometrical features.

Still regarding to the case $L/\lambda_l = 1$, also the average contact area is observed to diverge considerably from the rest, both at light and full contact. Thinking on this from the perspective of topography generation, by restricting the spectrum to a given range,

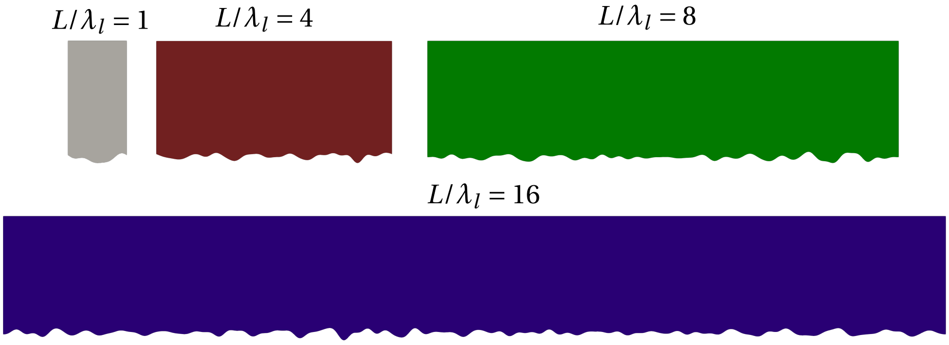


Figure 5.16: RCEs with different length, yet holding the same topographies characteristics. The profiles were generated with $H = 0.8$, $\lambda_l/\lambda_s = 4$, $\lambda_l = 5\,\text{mm}$, and RMS slope equal to 0.2. For convenience in the graphical representation, the scale along the RCE length is half of height scale. Only part of the RCE is represented.

the sum of harmonics is truncated and, thus, the underlaying preposition of the Gaussian generator—the sum of several random variables is normally distributed—is violated. The distortion of the contact area curve is attenuated with increasing RCE length, first, because more topography features are considered. Second, with increasing RCE length, the frequency resolution in the spectrum increases, meaning that the sum of harmonics includes more frequencies and the underlying hypothesis of the Gaussian roughness generator is recovered. For the four different lengths evaluated, the curves converge with increasing length, yet with some visible differences—for the bandwidth $\lambda_l/\lambda_s = 4$ currently under discussion. The curve for $L/\lambda_l = 1$ is surely apart from the others, and $L/\lambda_l = 4$ and $L/\lambda_l = 8$ are very similar only at light contact. For $L/\lambda_l = 8$ and $L/\lambda_l = 16$, the results are very similar throughout the load range, except at small loads, where a small gap can readily be identified. Despite the apparent convergence, it is not straightforward to state that the results for $L/\lambda_l = 16$ have completely converged for this bandwidth, since small differences still exist between the longest lengths.

By extending the roughness bandwidth, for $\lambda_l/\lambda_s = 16$ the standard deviation is globally reduced, in particular for $L/\lambda_l = 1$. With a wider spectrum, more harmonics are summed to synthesize the surface. This contributes, on the one hand, for the Gaussianity of the heights distributions, and, on the other, to increase topography variability within a given period and, thus, reduce the fundamentally possible different geometrical configurations. In contrast with the previous bandwidth, the convergence of the average contact area with length can be observed. The curves for $L/\lambda_l = 8$ and $L/\lambda_l = 16$ are almost coincident, suggesting that extending the RCE length even further will, most likely, not bring any accuracy improvement.

As has been noted, if the the length of the RCE is at least 8 times the largest wavelength in the profile, a nearly representative contact area evolution curve can be obtained, for all the examined bandwidths and Hurst exponents. This value guarantees converged results with small standard deviation. In comparison with the ratio $L/\lambda_l = 16$, using 8 times λ_l for the RCE length proves computational advantageous, since it allows the simulation

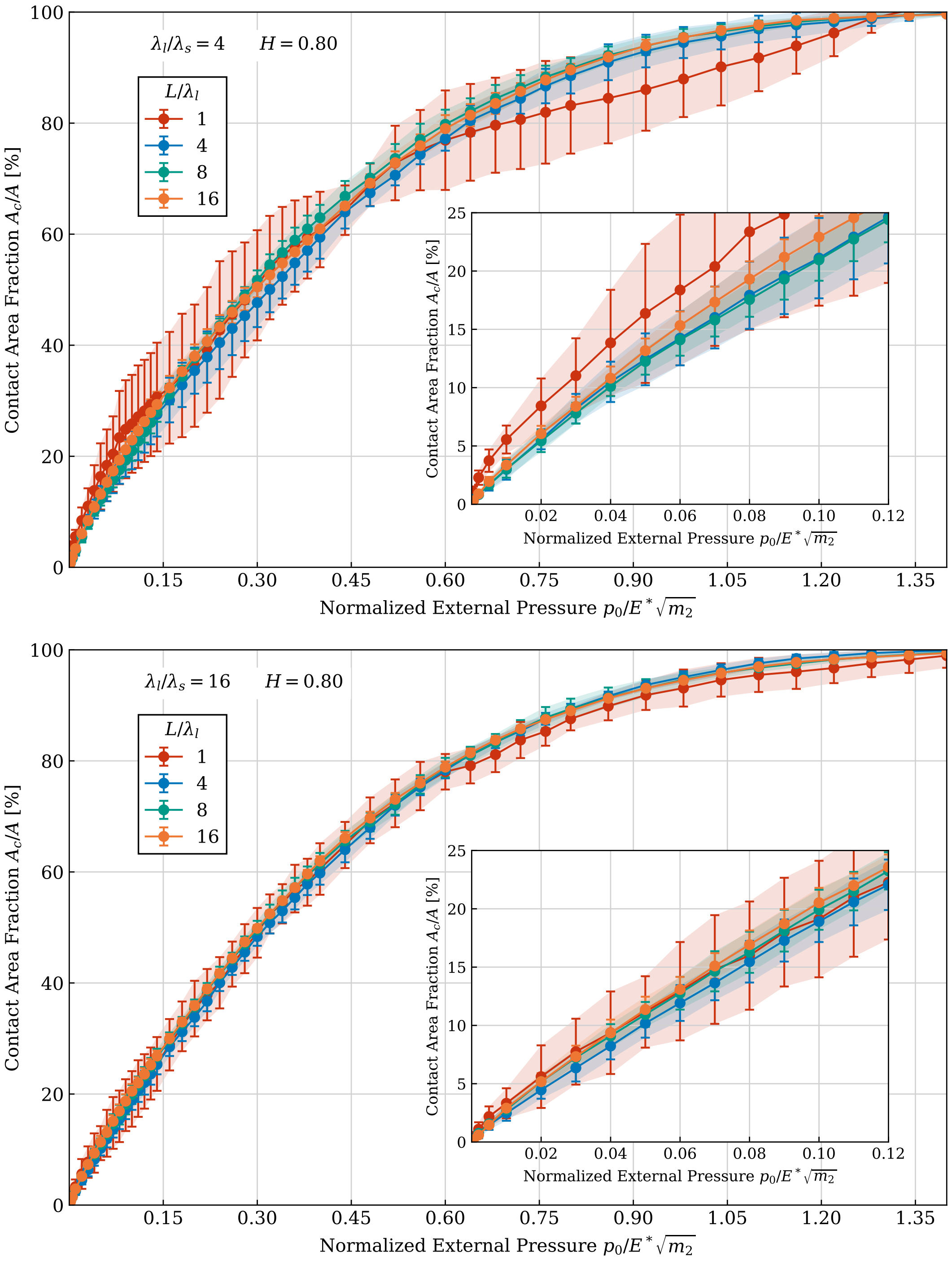


Figure 5.17: Influence of the RCE length in the contact area of 2D self-affine profiles, for different topography parameters. Two bandwidths $L/\lambda_l = 4$ and $L/\lambda_l = 16$ are plotted for $H = 0.8$. The results for $H = 0.4$ are similar, and are not shown here. The RMS slope is fixed at 0.2.

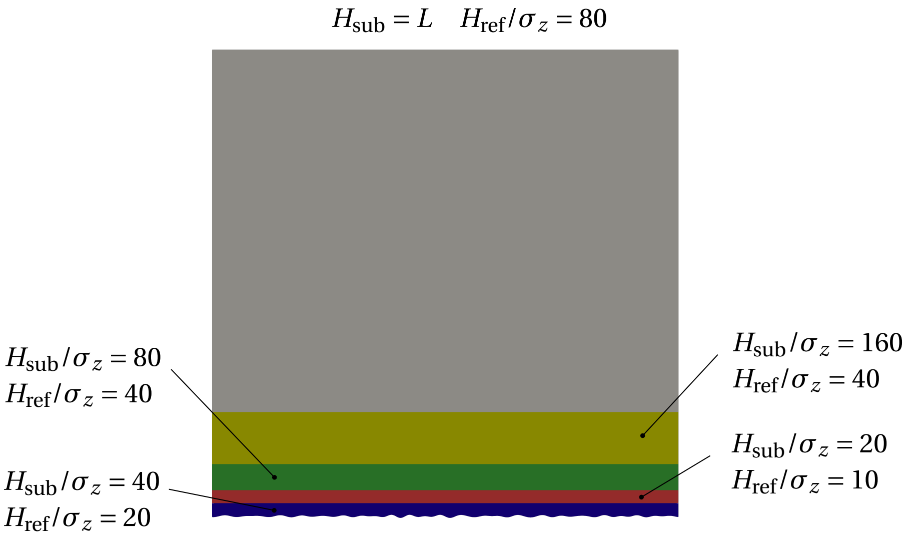

$H_{\text{sub}} = L \quad H_{\text{ref}}/\sigma_z = 80$
$H_{\text{sub}}/\sigma_z = 80$
$H_{\text{ref}}/\sigma_z = 40$
$H_{\text{sub}}/\sigma_z = 160$
$H_{\text{ref}}/\sigma_z = 40$
$H_{\text{sub}}/\sigma_z = 20$
$H_{\text{ref}}/\sigma_z = 10$
$H_{\text{sub}}/\sigma_z = 40$
$H_{\text{ref}}/\sigma_z = 20$

Figure 5.18: RCEs with the same topography and varying substrate height. The ratio between the substrate height and the high resolution region was kept equal to 2, with the exception of the reference RCE (it was mentioned that $H_{\text{ref}} = 80\sigma_z$) and also for the case $H_{\text{sub}} = 160\sigma_z$, where it was defined as $H_{\text{ref}} = 40\sigma_z$.

of profiles with wider spectra with smaller meshes, and very small error relative to the longer lengths. In fact, if needed, the length can even be defined as 4 times λ_l, once its error relative to the curves referring to longer lengths is still within an acceptable range—specially for wide spectra.

5.4.4 Height of the substrate

From Section 5.4.1 up the the current section, the RCE height was set equal to its length, and the high resolution mesh to 80 times the RMS height, based on the visual interpretation of the stress and displacement fields. Having established the mesh size and length required for the RCE, the topic of the RCE height can be resumed. It is paramount to reduce both the RCE height and high resolution region down to minimum values, in order to reduce the size of the FE mesh and, consequently, the computational resources required to solve its equilibrium problem.

A set of four values for the substrate height have been tested for that purpose, namely $H_{\text{sub}}/\sigma_z \in [20,\ 40,\ 80,\ 160]$. A constant ratio $H_{\text{sub}}/H_{\text{ref}} = 2$ was kept for every case, with exception of the largest height $H_{\text{sub}} = 160\sigma_z$, where the high resolution region was capped at $H_{\text{ref}} = 40\sigma_z$. Figure 5.18 illustrates the size of the substrate height considered for the different RCEs, in comparison with the initially estimated value. As in previous tests, two values were assumed for the Hurst exponent $H \in [0.4,\ 0.8]$. Four combinations of the long and short cut-off wavelengths were selected, namely $L/\lambda_l \in [4,\ 16]$ and $L/\lambda_s \in [32,\ 64]$, with $L = 5\,\text{mm}$, in order to include the influence of the length and spectrum bandwidth in the study. Each combination is tested with ten different topography realizations.

The numerical results for the assessment of the RCE height are plotted in Figure 5.19, in comparison with the reference solution, computed with the initially estimated block height. The results for all combinations of cut-off and Hurst exponents are very similar, and it was chosen only to present those in Figure 5.19. Identically to the RCE length test, all profiles were discretized with the same number of elements in the contact interface, dictated by the largest L/λ_s. It can readily be seen that the results for all RCE height fall relatively close to each other, and are practically insensitive to the spectrum bandwidth and length. For $H_{\text{sub}} = 20\sigma_z$, it can be visually identified that the curve is apart from the rest of the results, which are clustered around the reference solution. With increasing RCE height, the results converge to the reference solution, approximately at $H_{\text{sub}} = 80\sigma_z$, and improvements with $H_{\text{sub}} = 160\sigma_z$ are barely noticeable.

The height of the RCE can be equally assessed by other arguments than the real contact area curve. Recalling the boundary condition on the exterior boundary $\partial\Omega_{\text{ext}}$ (see Figure 5.8), the vertical displacement must be equal at all nodes. This condition would naturally be satisfied if the RCE height was sufficiently large, case where this boundary would be far from the contact interface and, thus, the boundary would behave as in simple compression. Due to such boundary conditions, the vertical displacements will invariably be equal at every point in the exterior boundary, even though when this would not be the natural configuration for the block. This is, if the exterior boundary is not within a region were the stress state is similar to that of a simple compression case, reaction forces are required at that boundary in order to keep it horizontal. This can be verified by computing the magnitude of the resultant reaction at the exterior boundary, plotted in Figure 5.20 for different substrate heights. The reaction forces for $H_{\text{sub}} = 20\sigma_z$ and $H_{\text{sub}} = 40\sigma_z$ are extremely large compared with the other two cases. With increasing RCE height, this quantity converges to zero. This is also illustrated by the auxiliary plot in Figure 5.20, where the arrows (nodal reaction forces) are larger for the case $H_{\text{sub}} = 20\sigma_z$, and the gradients of the Cauchy stress σ_{yy} (field variable) extend up to the exterior boundary.

Finally, since in the previous test the height of the finer mesh was reduced from the initial value $80\sigma_z$ to $40\sigma_z$, without affecting the contact area results, it proves prudent to confirm if the field variables are well capture by the discretization, near the contact interface. To this end, the nodal stress values at each element are plotted in Figure 5.21, for the high resolution mesh. The nodal values are computed for each element, and no FE averaging is applied, which results in a discontinuous plot of the variables defined within the element. As it can be seen from this figure, the nodal values of stress for neighboring elements near the transition layers are similar, and the respective gradients are small. Therefore, the stress field is well captured by the numerical mesh, and even half this value ($H_{\text{ref}} = 40\sigma_z$) can be used without any harm.

5.4.5 Influence of the number of realizations

In all previous studies, the average of 10 different realizations of the rough topography were considered, in order to compute the average contact area evolution curve—which was assumed to be representative of all simulated realizations. In fact, the simulation of 10 different topographies was based on a trade-off between representativeness and computation time, and also on the results of a similar study, yet relative to the statistical determination of the coefficient of friction (Wagner, 2018). Whilst by considering the

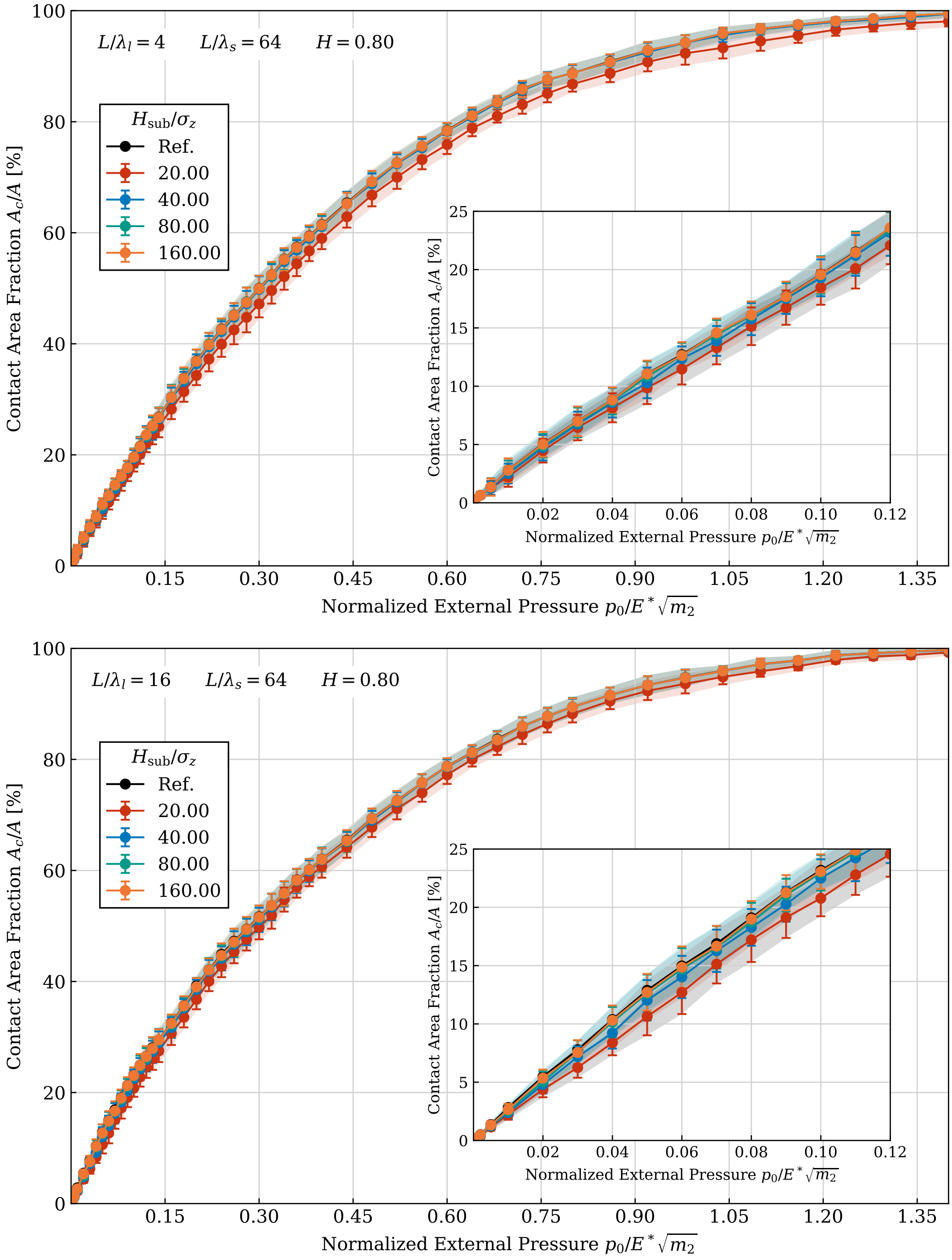


Figure 5.19: Influence of the RCE height in the contact area of 2D self-affine profiles, for different topography parameters. The reference solution is obtained by using the profile length as the block height. The ratio $H_{\text{sub}}/H_{\text{ref}}$ was set equal to 2 for all values of H_{sub}, with the exception of $H_{\text{sub}}/\sigma_z = 160$, where is was chosen $H_{\text{ref}}/\sigma_z = 40$. These results are presented for $H = 0.8$ and RMS slope equal to 0.2.

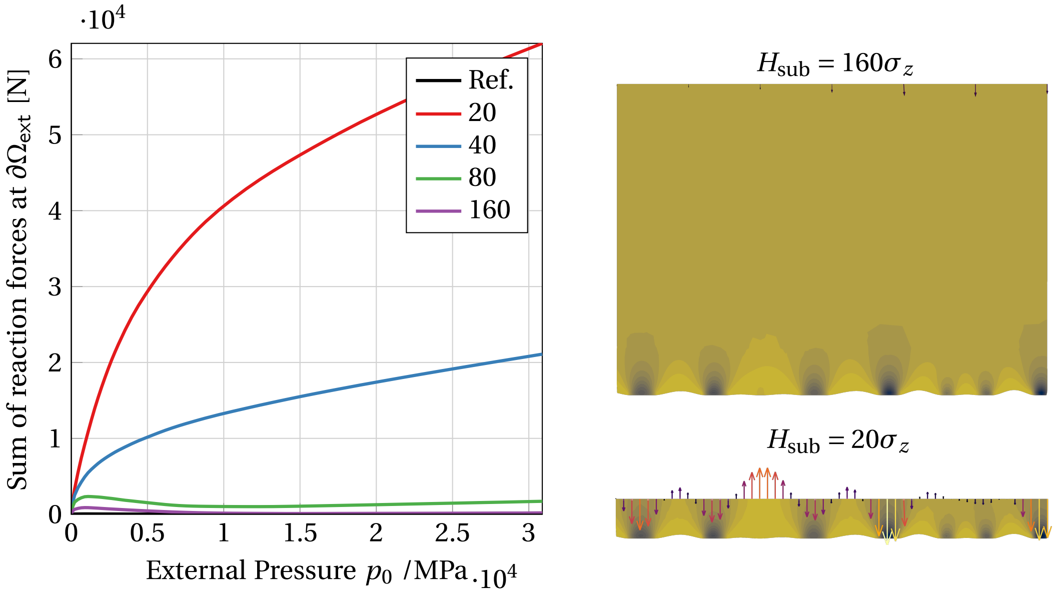


Figure 5.20: Magnitude of the sum of reaction forces at the exterior boundary $\partial\Omega_{\text{ext}}$. These reactions are responsible for keeping the vertical displacements equal at all nodes of that boundary. When they are large, it means that the exterior boundary is still within the region of effect of the contact interface. In the side illustration, the arrows represent the reaction forces, and the filed variable is the Cauchy stress σ_{yy}.

average of more than 10 different realizations the results would be more representative of the whole ensemble of possible realizations, the total time required to obtain all the results would be substantially increased.

Nonetheless, the impact of considering more and less realizations must be addressed, with the purpose of verifying whether the size considered for the sample is sufficient. This can be accomplished first, by considering a virtually impracticable number of realizations. Second a number samples with less realization are randomly selected from the ensemble and the average contact area is computed for each sample. Third, the standard deviation of the average contact area curves is computed across all samples, and this process is repeated for different numbers of realizations of these smaller samples. In theory, if the mechanical response of the RCE is to be representative, the ensemble standard deviation of the average contact area must go to zero, meaning that the average response is the same for any random set of realization chosen from the ensemble.

In this study, a total of 200 realization were generated with $H = 0.8$ and $L/\lambda_l = 8$, for each of the following high cut-offs $L/\lambda_s = 32$ and $L/\lambda_s = 64$. The height of the RCE was set to the length of the block, and for the height of the high resolution region it was considered $H_{\text{ref}} = 40\sigma_z$—the height of the block was unnecessarily large, but it was simpler at the time this study was performed. Only two nodes *per* asperity were generated in the FE mesh, such that the variability of the contact area is larger, and so the conclusions here extracted imply a safety factor. The ensemble standard deviation of the average contact area is computed in 10 sets with the testing number of realizations.

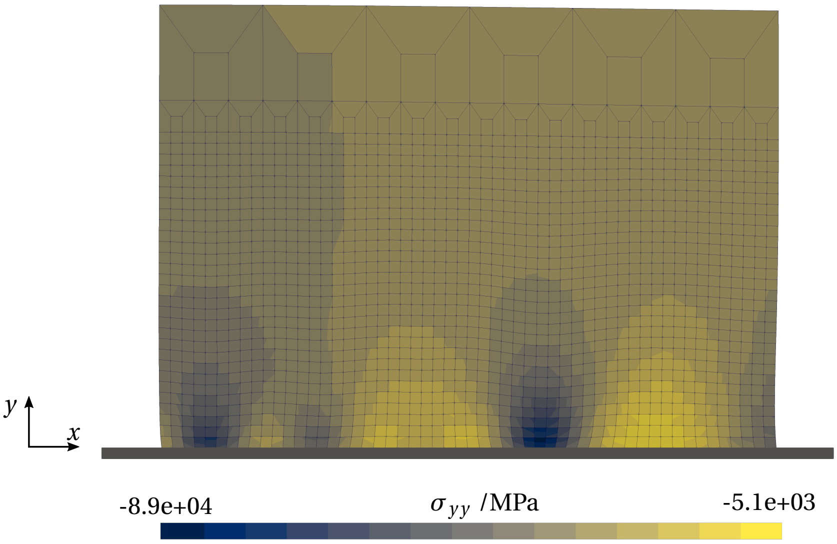


Figure 5.21: Plot of the element-based nodal Cauchy stresses, before any FE averaging operation, for the case $H_{\text{sub}} = 160\sigma_z$ and $H_{\text{ref}} = 40\sigma_z$. This figure evidences that the gradients inside the elements are small for this configuration for the RCE height and mesh, near full contact condition. This proportion established for the high resolution region is thus adequate and assures mechanical response representativeness.

Figure 5.22 shows the results of this numerical investigation, covering all the relevant load range. This plot emphasizes that the standard deviation is not uniformly distributed along all loads—which could already be verified from the previous results. Higher variance exist for nominal external pressure between 0.2 and 0.6. As expected, by increasing the number of realizations, the standard deviation reduces significantly. This effect is much more pronounced for very small numbers of realizations, and with increasing size of the RCE set, the consequent outcome improvements are very shallow. Additionally, by simply increasing the number of elements in the profile, even if it concerns a different bandwidth, the overall standard deviation decreases. From these results, it can be concluded that by considering 10 realizations for computing the representative response, the region of high standard deviation is avoided. From this value further, no significant improvements occur, yet the computation time increases—there are more simulations to run. Recall that, if the ensemble averages are normally distributed, there is 99.7% of finding an average curve within an interval of width 6 times the standard deviation, centered at the ensemble average.

5.4.6 Fitting a contact area evolution curve

So far, the load (external pressure) steps were prescribed, essentially, based on the work of different authors, such as Yastrebov, Anciaux, *et al.* (2015), and by numerical experience, i.e., by manually adjusting the increments until a nicely behaved convergence is achieved. This strategy is quite unsatisfactory, since it cannot be applied to a general case, and prohibits the full automation of the framework. One needs to determine a general

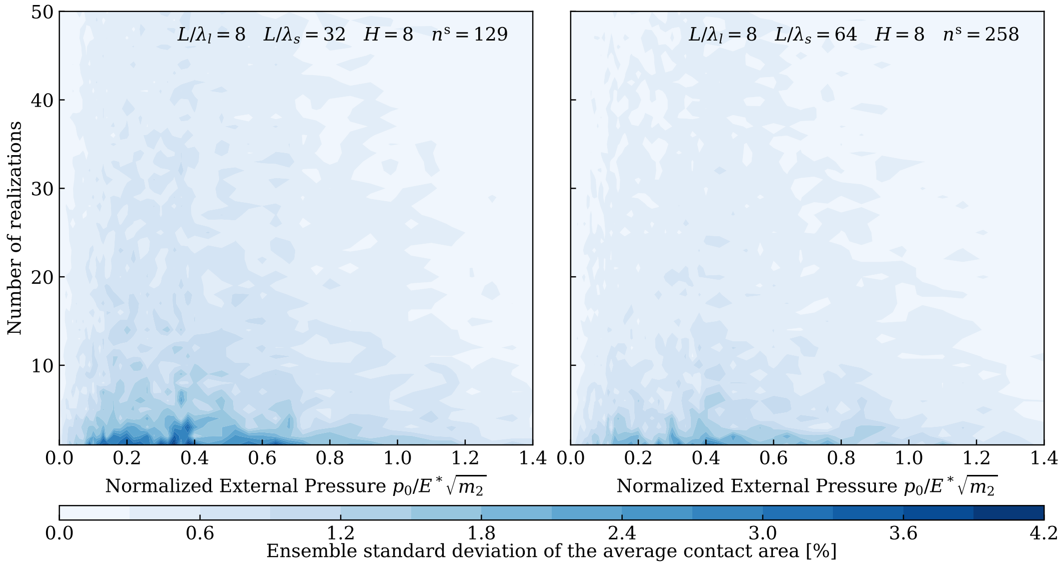


Figure 5.22: Ensemble standard deviation of the average contact area response, computed across 10 different groups of realization, for several number of realizations in each group.

evolution law for the numerically computed real contact area, such that increments can be automatically determined. For example, by specifying increments of approximately constant contact area, one induces similar changes in the deformable body configuration, hence it is hypothesized that it will stabilize the numerical convergence, as long as the increments are sufficiently small.

Preceding the fitting procedure, the average contact area curves relative to 10 RCE realizations comprising the combinations of $H \in [0.2,\ 0.5,\ 0.8]$ and $\lambda_l/\lambda_s \in [4,\ 8,\ 16]$, and verifying $L/\lambda_l = 8$ with 4 nodes *per* asperity are computed and plotted. For simplicity, the fitting function is chosen to be a third-order polynomial passing through the origin—so that for zero load, the real contact area is zero, as well. Note that here a precise numerically derived contact evolution law is not pursued, but a rather rough fit, for practical applications. For an example of such numerical contact evolution laws, see Yastrebov, Anciaux, *et al.* (2012).

The numerical results are plotted in Figure 5.23, together with the fitted function and the Persson's model for 2D contact. The result of the numerical fitting writes

$$\left\{\frac{A_c}{A}\right\}_{\text{fit}} = 2.2565\left(\frac{p_0}{E^*\sqrt{m_2}}\right) - 1.8433\left(\frac{p_0}{E^*\sqrt{m_2}}\right)^2 + 0.5333\left(\frac{p_0}{E^*\sqrt{m_2}}\right)^3 . \tag{5.67}$$

The numerical fit approximates all curves with small error for all the different topography characteristics examined. The increments based on the real contact area can be computed by simply inverting Equation (5.67), using a root finding numerical technique. The analytical solution from Persson's model is also plotted in this figure, providing an initial assessment of the quality of the numerical results. It can readily be observed that the numerical results fall close to the theoretical curve, and relatively large deviations are mostly verified for light contact. There is no abundant discussion in the literature on this topic, and the only source where 2D numerical results were compared with analytical models, namely, the Persson's model, is the publication of Carbone, Scaraggi, *et al.* (2009). These authors opted to normalize the profile height by the RMS height, and no reference to the RMS slope is provided, therefore it is not practical to plot such results in Equation (5.67). However, from a qualitative perspective, the results by these authors seem to overestimate Persson's model, almost doubling the analytical result in light contact. In the results obtained in the present dissertation, the difference is not so significant. Nonetheless, such comparison should be carefully interpreted, as different numerical strategies are employed (FEM in this contribution, and GFMD in Carbone, Scaraggi, *et al.* (2009)), which can justify the differences in the results.

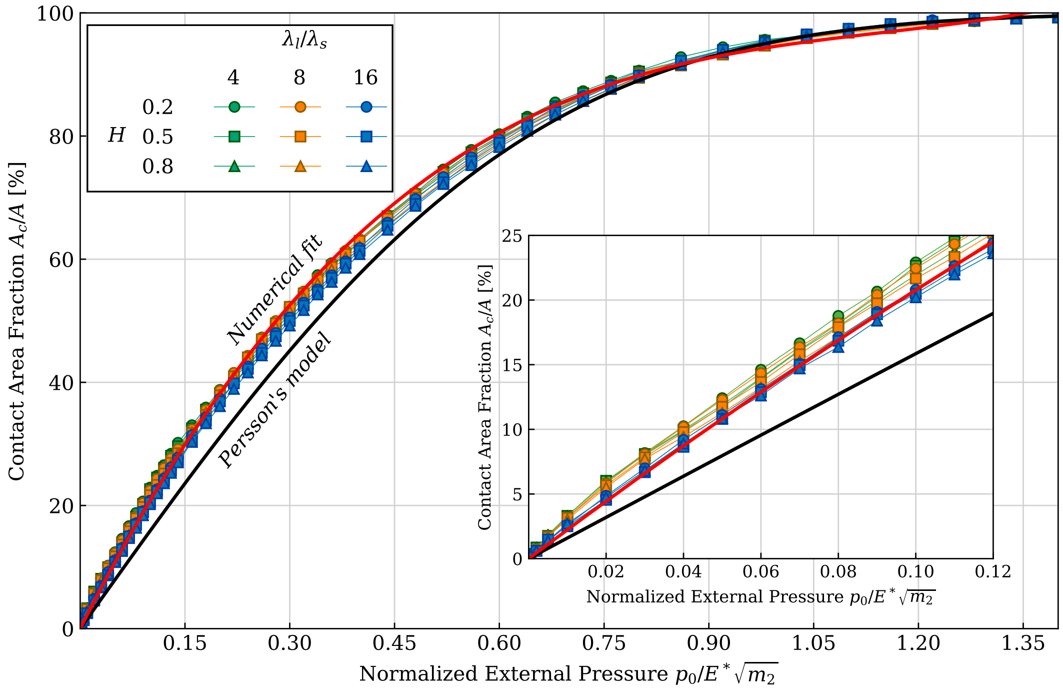


Figure 5.23: Numerical fit of the contact area-pressure curves for different topography parameters, for the definition of the incrementation rule.

5.5 Rules of thumb for the definition of a 2D RCE

The general conclusions of the numerical investigations regarding the definition of a RCE for a 2D self-affine profile, discussed in throughout this chapter, can be summarized in simple rules of thumb. These shall consider some tolerance regarding the required representativeness, such as concerning the RCE length, as discussed in Section 5.4.3, such that the numerical models do not become excessively large and, therefore, limiting. The following criteria will be required for the numerical multiscale analysis of rough contact, inasmuch that RCEs at different scales shall be required to model different scales of roughness, and from the results presented in this chapter, one can readily establish the dimensions and mesh for the micromechanical problem.

Rules of thumb for the definition of 2D RCE for self-affine rough profiles

$$\text{Mesh:} \qquad \Delta x \le \lambda_s/8\,; \tag{5.68}$$

$$\text{Length:} \qquad L \ge 8\lambda_l\,; \tag{5.69}$$

$$\text{Substrate height:} \qquad H_{\text{sub}} \ge 160\sigma_z\,; \tag{5.70}$$

$$\text{Fine mesh height:} \qquad H_{\text{sub}} \ge 20\sigma_z\,; \tag{5.71}$$

$$\text{Number of realizations} \quad \ge 10\,. \tag{5.72}$$

Page intentionally left blank

Chapter 6

Multiscale finite element modeling of rough contact by contact homogenization

Roughness features are known to extend throughout several length scales in real rough surfaces. In order to incorporate large bandwidth roughness spectra in numerical models, very fine discretizations are required to correctly assemble the smallest scales. At the same time, the model must be large enough to encompass the largest scales. The dangerous combination of large models with fine discretization conveys to an inconveniently fast growth of the computational cost of numerical models with increasing roughness bandwidth. Several multiscale approaches have been proposed in the last decades, in order to circumvent the prohibitively expensive computational resources requirements from rough contact modeling—the vast majority within the framework of finite element analysis.

Most multiscale strategies rely on the principle of separation of scales, which assumes that the characteristic length l of microscopic features is very small compared with the microscale characteristic length L, viz.

$$l \ll L\,. \tag{6.1}$$

Figure 6.1 provides an illustration of the scale separation applied to the contact of two rough bodies. The macroscopic shape of the bodies defines the largest characteristic length, while a zoom in the contact interface reveals that the apparently smooth boundaries are, in fact, rough—leading to the definition of a microscale. The basic idea of multiscale analysis, relying on the separation of scales, is to establish a *Representative Volume Element* (RVE), which is assumed to be statistically representative of the microscale features (R. Hill, 1963). The results of the equilibrium solution of the RVE are then incorporated in the larger scale by an *averaging* or *homogenization* step—the larger scale does need to model the microscopic features, and so can be modeled as homogeneous. This approach is typically applied to the bulk of the bodies, in order to obtain complex constitutive laws on the fly. Similar procedures can also be adopted to interfaces and to mechanical contact, see Stupkiewicz (2007). Thus, the fundamental ideas driving contact homogenization strategies are the replacement of an highly complex contact interface with a *smoothed* or *homogenized* interface, with averaged properties computed from microscale analysis.

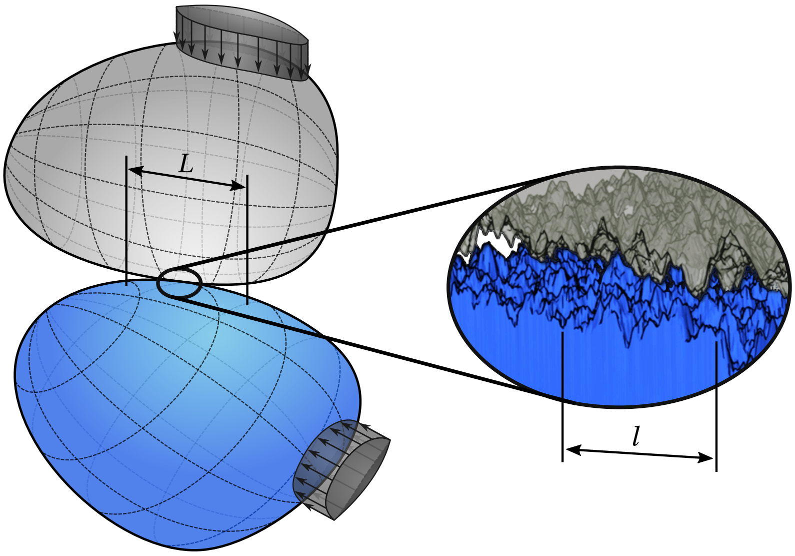


Figure 6.1: Scale separation in the contact of rough bodies. Adapted from Pinto Carvalho (2018).

Based on the principle of separation of scales, the division between macroscale geometry and microscale roughness features is a valid preposition, as long as the contact area is small compared with the roughness characteristic length, as suggested by Figure 6.1. However, often one wants to investigate the influence of the diverse roughness length scales, i.e., different ranges of the surface power spectrum, on the contact properties, rather than isolate the effect of roughness from geometrical shape. Roughness length scales cover a continuous spectrum, as observed from a typical PSD, cf. Figure 2.17. This poses a major complication for the multiscale analysis of rough contact, because it violates Equation (6.1). The smallest length scale at a given PSD range is the largest scale at the following one—scales are not naturally separated. By excluding intermediate length scales, scale separation could be artificially incorporated, allowing the application of classical hierarchical multiscale homogenization methods. However, intermediate length scales have been verified to contribute to contact properties and, thus, cannot be omitted.

Multiscale modeling of rough contact is a growing research field, on which diverse attempts have been made to cope with the latter difficulties. In the last decade, major advances have been motivated by the study of the frictional contact of rubber materials on rough road surfaces, and several works and approaches within the FEM framework have been proposed.

Recently, a different type of multiscale analysis has been performed, consisting in coupling different numerical methods at different scales. For example, molecular dynamics can be used to model contact down to the nanoscale, and a concurrent strategy is used to couple it with a continuum based numerical method, such as the FEM (Anciaux *et al.*, 2012). Nonetheless, this kind of multiscale analysis falls out of the scope of the present work, and is referred here for completeness. For a general overview of numerical methods applied at modeling rough contact at several scales, and more insights into multiscale approaches under this topic, the reader is referred to the recent review by Vakis *et al.* (2018).

From a physical point of view, the principal caveat of multiscale modeling is the definition of which scales are important for a certain physical phenomena and, thus, worth of being modeled. Conversely, from a computational perspective, it is important to understand how the number of scales considered to model rough contact problems translates into a computational advantage over Direct Numerical Simulations (DNS) counterparts.

6.1 Review of multiscale approaches to rough contact

Multiscale algorithms can be classified, generally, as coupled or information-passing. In coupled multiscale strategies, also commonly designated by FE^2, a micromechanical numerical calculation is started, typically, at every contact integration point at the macroscale. The microscale results are then incorporated—in a fully integrated fashion—in the macroscale simulation, from which the loading at the microscale is established in the next integration step, in the analysis at different scales. This procedure is usually accomplished by using standardized geometries at the different scales. Information-passing algorithms, on the other hand, are based on the simulation of the contact problem at different scales independently, which are then combined to give a multiscale result. Coupled algorithms are computational expensive, but ensure maximum quality of the information passed between the scales. In contrast, information-passing algorithms are faster, but since only simple average quantities are passed between the scales, some information might be lost in the scale transitions.

Remark 6.1 on the following presentation of the multiscale approaches.
As referred in the beginning of this chapter, most advances in numerical strategies in modeling rough contact have been made in the context of rubber friction research. Hence, most of the multiscale algorithms are formulated for frictional contact, which is not addressed in this work. However, most of these strategies can be readily reformulated for frictionless contact, regarding the computation of the real contact area, specially because most of the approaches also account for adhesion and, therefore, the issue of the real contact area is necessarily addressed. Moreover, all following multiscale methods are formulated within the FEM framework, hence this shall be tacitly assumed in the ongoing presentation.

Early application of multiscale approaches based on contact homogenization report to Tworzydlo *et al.* (1998). The macroscopic normal and frictional contact response is computed through statistical homogenization of the results from FE simulations on individual asperities with different geometries. The method was restricted to the realm of small deformations.

Temizer and Wriggers (2008) developed a coupled contact homogenization multiscale strategy, regarding the contact between a flat rubber block and moving particles. In accordance to the previous introduction on the types of multiscale algorithms, at each integration point at the interface a Representative Contact Element (RCE) is analyzed, which removes the need to model individual particles at the macroscale. This is, the contact properties are computed on-the-fly. Several aspects related to the RCE definition for the particular case in study were studied, as well. The typically high computational cost of FE^2 approaches is the major drawback of this method.

In Reinelt (2009) and Wriggers and Reinelt (2009), frictional contact between a flat rubber block and a rigid rough surface is approached with a formulation based on the Height Difference Correlation (HDC) function. This is defined in a similar fashion to the ACF, yet the correlation of the height difference between two points is regarded. At each scale, the rough surface is approximated by a sinusoidal function, whose amplitude is selected from the spectrum of the HDC function. Starting with a frictionless simulation at the smallest scale, a micromechanical friction law is defined at different values of pressure and velocity. This law is, then, incorporated in the larger scales by means of a frictional contact formulation. Since the results from each scale are not coupled, it allows for a reduction in computation time, compared with the approach by Temizer and Wriggers (2008). In the work of De Lorenzis and Wriggers (2013), single scale tests were performed to investigate the influence of several parameters in the quality and convergence of the results, which revealed to be difficult in some cases. Moreover, the selection of the number and frequency of the sinusoidal contributions is not well-established in the literature.

A coupled multiscale algorithm was proposed by Nitsche (2011), usually referred as a projection method. It consists in starting a new micromechanical simulation at new contact spots, where the geometry and loading are determined from the local macroscale features at that region. It is noteworthy to emphasize that the geometry of the microscale problem is defined from a truncated region of the current contact region. The forces resulting from the microscale equilibrium are then projected in the macroscale problem. However, difficulties in the convergence of projected quantities and the intricate information-passing strategy renders the approach difficult to extend to more complex situations.

The recent work by Wagner (2018), also documented in Wagner, Wriggers, Klapproth, *et al.* (2015) and Wagner, Wriggers, Veltmaat, *et al.* (2017), provides a relatively simple information-passing framework for the multiscale analysis of rough surfaces. In opposition to the previous approaches, the surface roughness is modeled directly via its power spectrum (PSD). Different scales are defined by splitting the power spectrum at different frequencies, and filtering the unwanted components from the topography. The first formulation of the multiscale strategy, in Wagner, Wriggers, Klapproth, *et al.* (2015), is an extension of the work of Reinelt (2009), following a fundamentally distinct approach to model the rough road surface. Based on the contact pressure and velocity distribution at the macroscale, a set of discrete values of pressure and velocity are defined as the inputs for the next scale. With the equilibrium results from the microscale, a micromechanical friction law can be formulated and inserted at the larger scales. With grounds on the violation of the principle of separation of scales, Wagner (2018) claimed that the definition of a constant pressure loading at the smaller scale from a finite set of pressure values, selected from the macroscale distribution, is not reasonable. A non-uniform pressure loading would be physically meaningful, since the smaller scales are larger than the contact regions were normal pressure values are computed. Based on this argument, a simplified version of the multiscale approach was proposed in Wagner, Wriggers, Veltmaat, *et al.* (2017), by applying a smoothing operation to the pressure downscaling. Instead of passing a set of pressure and velocity values chosen from the respective distribution at the macroscale, the average contact pressure and the input tangential velocity are imposed on the smaller scale. Then, by downscaling the average contact pressure and upscaling the friction coefficient and real contact area, the full scale results can be obtained.

Remark 6.2 on the selection of a multiscale approach.
The work of Wagner (2018), in particular, the more recent formulation of the strategy in Wagner, Wriggers, Veltmaat, et al. *(2017), establishes the foundation for the developed multiscale framework. This selection is motivated on the coherent physical grounds of the formulation, together with its simplicity and ease of incorporation within the currently available numerical tools.*

6.2 General multiscale framework

Following the fundamental ideas on self-affine rough topography laid in previous chapters, it is verified that the multiscale roughness characteristics are held in the topography PSD. This function covers a range of frequencies defined between the low cut-off k_l and the high frequency cut-off k_s (ignoring the roll-off). Each of these frequencies is intrinsically related to the large and short cut-off wavelengths λ_l and λ_s, respectively. The starting point of any multiscale approach is the definition of the several scales involved in the the problem. A *splitting frequency* k_{split} is introduced in the rough topography spectrum, which allows its division into the macroscale and microscale spectra, see Figure 6.2. The macroscale spectrum ranges from k_l and the splitting frequency k_{split}, while the microscale spectrum goes from k_{split} to k_s. The separation in n^{s} scales can easily be generalized by introducing $\mathsf{n}^{\mathsf{s}} - 1$ splitting frequencies in the power spectrum.

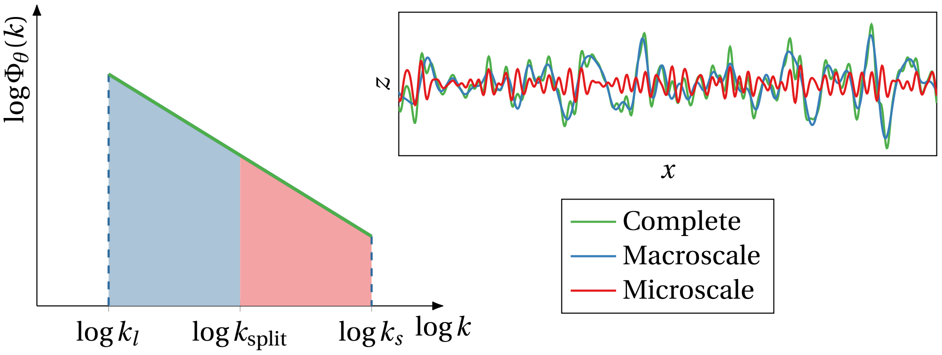


Figure 6.2: Topography splitting into a microscale and a macroscale component, by the introduction of a splitting frequency k_{split} in the PSD.

Since the reconstruction of the topography from the PSD is additive, i.e., it results from the superposition of several spatial harmonics, the complete topography can be decomposed in several scales by specifying different bandwidths, which acts as bounds in the harmonics sum—while holding the Hurst exponent H and the continuous PSD scale factor C_0 or C_0' unchanged for every scale. This is illustrated for the decomposition in two scales, in Figure 6.2.

Having specified the different roughness length scales, the contact area evolution curve at each one can be found by performing a FE simulation, independently of all the other scales. The curve shall be obtained up to full contact conditions $A_c/A \rightarrow 100\%$, so that a complete description of the contact area at all load ranges is provided. From these, a database of area-pressure pairs can be created, which can then be post-processed according to several strategies. The approach proposed by Wagner (2018) consists in passing, at each load step at the macroscale (scale 1), the mean contact pressure $\bar{p}_i^{\eta\{1\}}$ to the scale below (scale 2).[1] Thus, the load increment $p_0^{\{1\},i}$ at scale 1 will be directly related with the load increment $p_0^{\{2\},i} = \bar{p}_i^{\eta\{1\}}$ at scale 2. By mapping the nominal exterior pressure at different scales, a list of real contact area fractions at all scales, and for a given macroscale pressure can be obtained. The multiscale solution for the real contact area, denoted by the superscripts MS, can be computed in a multiplicative homogenization step, by taking the product between all contact area fractions of the said list, and repeating it for every nominal exterior pressure increment at the macroscale, viz.,

$$\left\{\frac{A_c}{A}\right\}^{\mathrm{MS}}\left(p_0^{\{1\},i}\right) = \prod_{j=1}^{\mathrm{n^s}} \left\{\frac{A_c}{A}\left(\bar{p}_i^{\eta\{j-1\}}\right)\right\}^{\{j\}}, \quad \text{with } \bar{p}_i^{\eta\{0\}} = p_0^{\{1\},i}. \tag{6.2}$$

With respect to the mathematical interpretation of Equation (6.2), it describes the multiscale solution for the real contact area fraction at the load increment $p_0^{\{1\},i}$—the pressure applied to the full scales case is applied directly at the macroscale (scale 1). Starting at scale 1, the real contact area at the pressure increment i is multiplied by the contact area fraction in scale 2, computed at the mean contact pressure at scale 1. Moving to scale 2, the product between the contact area in scale 3 at the mean contact pressure at scale of scale 2 and the previous result is determined. Repeating the sequence of mean pressure calculation, contact area fraction identification and recursive multiplication, as one moves down to the smallest scale, the multiscale solution is computed progressively.

The physical interpretation of the multiplicative homogenization step comes directly from the fact that the variable of interest is a *fraction* of some quantity. Suppose that at the macroscale the real contact area is $A_c^{\{1\}}$, which means that the real contact area fraction in this case reads $A_c^{\{1\}}/A$. Now, when looking at the contact area at scale 2, the nominal area is no longer A, but the true contact area at scale 1, i.e. for this scale the ratio is $A_c^{\{2\}}/A_c^{\{1\}}$. The real contact area accounting for the two scales thus comes

$$\frac{A^{\{2\}}}{A} = \frac{A_c^{\{1\}}}{A}\,\frac{A_c^{\{2\}}}{A_c^{\{1\}}} \Leftrightarrow \left\{\frac{A_c}{A}\right\}^{\mathrm{MS}} = \left\{\frac{A_c}{A}\right\}^{\{1\}}\left\{\frac{A_c}{A}\right\}^{\{2\}}. \tag{6.3}$$

In sum, the general multiscale framework for predicting the real contact area fraction consists in, first, splitting the power spectrum in several ranges, by introducing a number

[1] The notation $p_0^{\{j\},i}$ represents the nominal external pressure at the load increment i in scale j, and $\bar{p}_i^{\eta\{j\}}$ stands for the mean contact pressure at the load increment i in scale j.

of splitting frequencies k_{split}. Second, the mean contact pressure is downscaled at each load increment, starting from the macroscale, whose load is the same as the applied in a single scale model. Third, the real contact area fraction is upscaled, from the microscale up to the macroscale, at the downscaled pressure values. Last, all contact area fractions are homogenized by taking the product of all upscaled ratios (at each load increment), resulting in the multiscale solution for the contact area ratio. A graphical interpretation of the multiscale approach is provided in Figure 6.3.

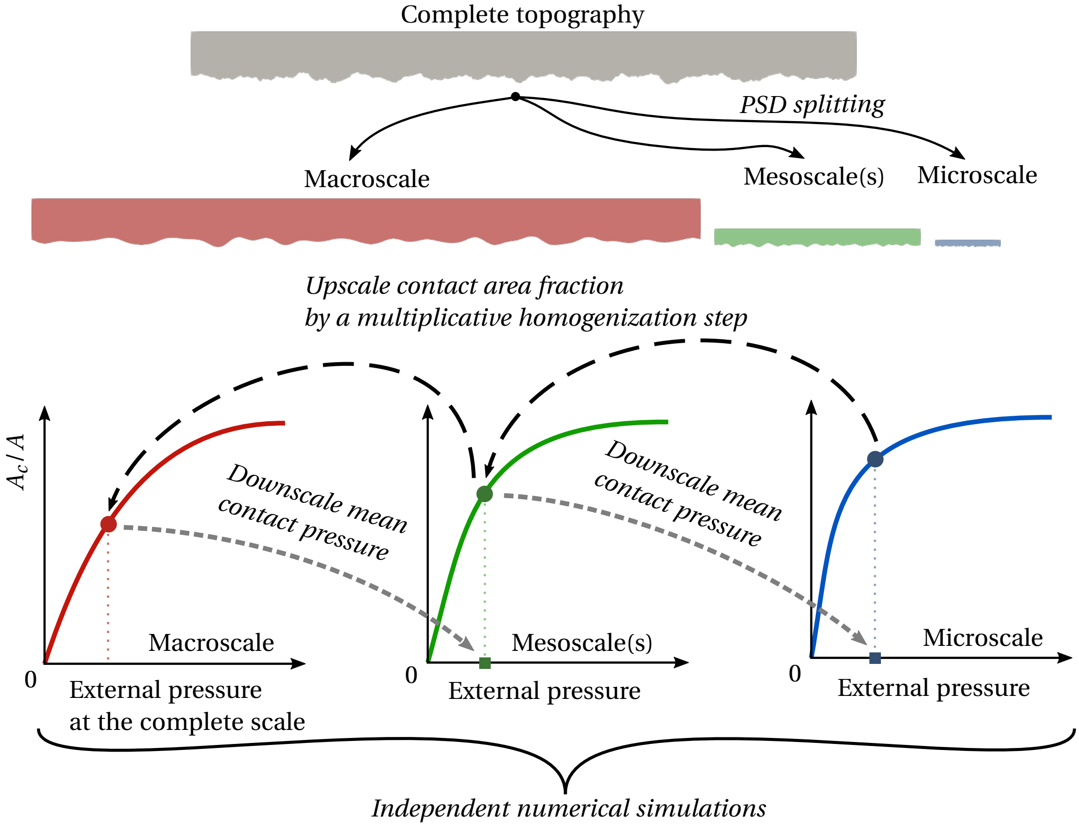


Figure 6.3: Schematics of the multiscale strategy to predict the real contact area fraction in rough contact. It consists in performing several independent FE simulations, whose results can be connected in a post-processing step. First, the surface is decomposed into several scales, by defining split frequencies of the spectrum. Second, for each scale, the FE solution is computed for the load range of the original problem. Third, for each value of pressure at the macroscale, the contact area at the mean contact pressure is listed for each succession of scales. Last, a point-wise multiplication of the list elements, treated as a multiplicative homogenization step, gives the multiscale homogenized solution.

6.2.1 Definition of the splitting frequencies

In the original work of Wagner (2018), the definition of the splitting frequencies is not explored, and no other sources referring similar issues have been found. It is paramount to have a criterion for setting the splitting frequencies, in order to simplify the automation of the multiscale framework, and also for the natural extension of the strategy to more than two scales. With this in mind, a splitting criterion is proposed here, entirely based on a computational profit point of view.

From the single scale simulations presented previously, several rules for the RCE are already established, so that if one wants to analyze a given set of topography characteristics, namely, the long and short cut-off wavelengths, the RCE length and mesh can be uniquely determined. Since the ultimate goal of a multiscale approach is the reduction of computation time relative to the respective DNS, the maximum computational efficiency is achieved when the computational load is evenly distributed by all scales. Hence, knowing beforehand that for each scale the ratio $L^{\{j\}}/\lambda_l^{\{j\}}$ and $\lambda_s^{\{j\}}/\Delta x^{\{j\}}$ is fixed by the aforementioned rules of thumb for the RCE definition, similar meshes and, thus, uniformly distributed computational loads, can be achieved if the bandwidth $\zeta = \lambda_l^{\{j\}}/\lambda_s^{\{j\}}$ is constant across scales. This criterion writes

$$\frac{\lambda_l}{\lambda_{\text{split}}^{\{1\}}} = \frac{\lambda_{\text{split}}^{\{1\}}}{\lambda_{\text{split}}^{\{2\}}} = \ldots = \frac{\lambda_{\text{split}}^{\{\mathrm{n^s}-1\}}}{\lambda_s} = \zeta^{\text{MS}} \,. \tag{6.4}$$

From Equation (6.4), the optimal scale bandwidth ratio is

$$\zeta^{\text{MS}} = \sqrt[\mathrm{n^s}]{\frac{\lambda_l}{\lambda_s}} \,. \tag{6.5}$$

All splitting frequencies can now readily be obtained from

$$\lambda_{\text{split}}^{\{j\}} = \frac{\lambda_l}{\left(\zeta^{\text{MS}}\right)^j}, \quad \text{for } j = 1, \ldots, \mathrm{n^s} - 1 \,. \tag{6.6}$$

6.2.2 Generation of the microscale topography

The power spectrum splitting is followed by the generation of each scale's topography. While for the macroscale this procedure may seem straightforward, following the usual topography synthesis with a longer short cut-off wavelength, some conceptual complications arise regarding the generation of the meso and microscales. For example, in Figure 6.2 the separation of a macro and microscale with the same length can be observed. These topographies are numerically generated from the single scale characteristics, but with different, yet consistent, cut-offs. Nevertheless, if the rough profiles represented in that figure were to be used directly in the multiscale strategy, the overall computational cost would increase, comparatively with the DNS. Even though a coarse mesh could be used at the macroscale (since high frequencies are removed), the microscale would still require a mesh as fine as in the single scale scenario, since its the physical length is the same—and, thus, notably excessive. Nonetheless, one can cut the length of the microscale by employing the minimum RCE length established before, for single scale analyses, without compromising the accuracy and representativeness of the results. In practice, this means that only part of the microscale profile Figure 6.2 shall be used, thus reducing the overall computational cost.

One way to produce the rough topography at the microscale is by generating a full length microscale, as in Figure 6.2, and then truncating the profile, such that only a minimum length is preserved. However, such approach is not advisable, because it distorts the microscale PSD and ruins the topography periodicity—which must be assured by any other method. Within the current numerical framework, the simplest way of generating topography at the microscale is by using the random roughness generator directly, which synthesizes the rough topography with prescribed characteristics and length, accordingly to the representativeness rules.

The frequency resolution of the power spectrum, i.e., the spacing between the discrete PSD points in the frequency axis, is inversely proportional to the topography length (see Appendix A). Conversely, the maximum frequency available in the PSD is inversely proportional to the sampling length. Using the full scale numerical model as a reference, the macroscale topography can reproduce exactly the discrete PSD points, with the exact same spacing. Differently, even though the microscale can reproduce all the required frequency range (equal sampling length), the spacing between the discrete PSD points at the microscale are considerably different from the reference points, owing to the shorter microscale length. This can be observed in the plot of all previously discussed discrete power spectra in Figure 6.4. It is self-evident that the microscale spectrum is considerably less populated than the complete spectrum, for the same frequency range. In the extreme case, the reduction of the number of discrete frequencies points with increasing number of splits may lead to the approximation of a complex sum of harmonics by a purely sinusoidal topography, or even to the complete absence of frequencies in certain ranges.

In order to assess whether the direct microscale topography generation verifies other topography features apart from the power spectrum, a very brief test can be performed, comparing the RMS slope of a generated microscale profile with a full length microscale profile (as in Figure 6.2) and a truncated profile from the full microscale. The single scale is defined by $H = 0.8$, $L = 1$, $\lambda_l = L/8$, $\lambda_s = L/128$, and $\hat{C}_0'$ is computed such that the theoretical single scale RMS slope is 0.2. The largest wavelength in the microscale is $\lambda_l^{\{2\}} = L/32$. The full scale microscale is generated with 1024 points, and the directly generated microscale with 256. The results are plotted in Figure 6.5.

At a qualitative level of analysis, by visual inspection of the two profiles, there are not generally obvious differences and, in fact, the peaks and relative amplitudes of some features are similar. For a quantitative analysis of the results, the RMS slope was computed with a forward finite differences scheme for all cases. The RMS slope of the generated and full microscale are uniquely determined by the PSD, and share very similar values. As for the truncated microscale, it is observed that it can assume a wide range of values, and its probability distribution closely resembles a Gaussian curve, centered around the values from both the full and directly generated microscale. Note that the RMS of the microscale

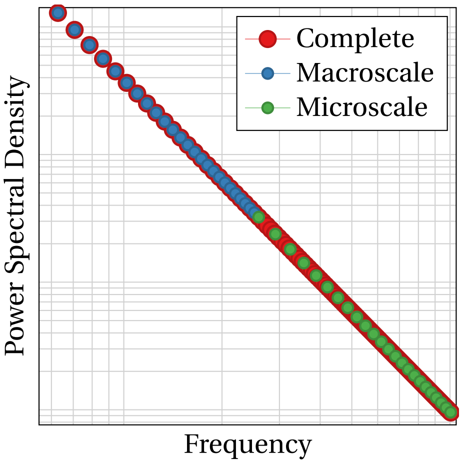


Figure 6.4: Power spectral density of the complete scale topography and the macro and microscale contributions. It is paramount to note that the micro and macroscale spectra are plotted on the top of each other for mere convenience. In fact, recall that the discrete spectrum is scaled by the sampling length, therefore the spectrum at different scales does not necessarily match the values of the full scales spectrum.

is necessarily smaller than 0.2, since this is the value prescribed for the complete, single scale topography.

In sum, it can be concluded that, despite the smaller frequency resolution in the power spectrum, direct generation of the microscale can reproduce other topography characteristics with reasonable accuracy, and can safely be used within the multiscale framework—as long as the number of point frequencies in the topography is not dangerously small.[2] As a closing comment, it must be mentioned that, since only Gaussian topography is addressed in the single scale models (see Section 5.3.1), it seems reasonable to generate Gaussian topographies for the microscale models, as well—actually, even for the macroscale.

Remark 6.3 on the treatment of the discrete splitting frequencies.
The rough topography at the smaller scales can be generated with the numerical topography generator, by specifying combinations between the cut-off wavelengths of the single scale topography and the splitting wavelengths. While this methodology is established based on a continuous roughness spectrum, where each single frequency has a null contribution for the topography height, only discrete quantities are concerned within the numerical framework. In contrast with the continuous case, the discrete frequencies have a finite contribution to the topography. Then, this raises the question whether to include the splitting frequency in both macro and microscale (as short and long cut-offs, respectively), or to exclude it in one of the scales. In the following, it was chosen to set the long cut-off wavelength at each microscale at the next single scale discrete frequency,

[2] Further investigations shall be done in the future, in order to establish the impact of the number of frequencies incorporated in the topography to the quality of the generated rough topography at smaller scales.

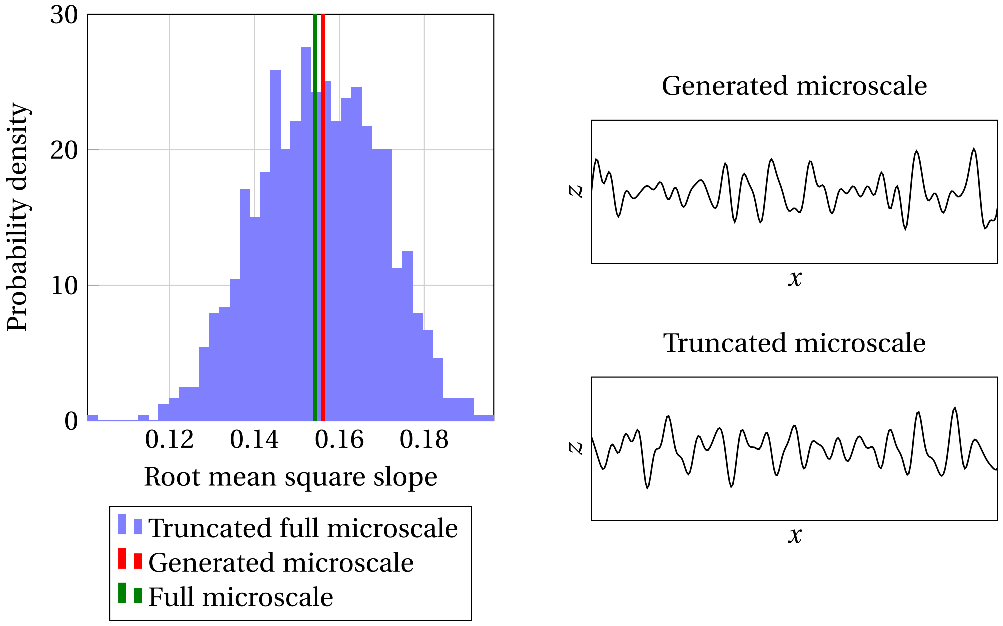


Figure 6.5: Comparison between the profile statistics of a directly generated topography and a truncated full microscale. The complete topography is characterized by $H = 0.8$, $L = 1$, $\lambda_l = L/8$, $\lambda_s = L/128$, and the RMS slope is set to 0.2. The largest wavelength in the microscale is $wl_l^{\{2\}} = L/32$. 1024 points are used to discretize the full microscale, and 256 for the directly generated microscale. A forward finite differences scheme is used to compute the RMS slope.

thus avoiding the repetition of the splitting frequencies in two consecutive scales. This way, the contribution of each frequency is included only once in the analysis, and the consistency with the complete spectrum is ensured.

6.2.3 Update of the power spectrum scaling factor

For the single scale numerical modeling of rough contact, the PSD scaling factor was computed as a function of the prescribed RMS slope. When the same numerical framework is embedded in a multiscale strategy, in order to keep the PSD continuous after assembling all scales, the scaling factor C_0, or C_0', shall remain unchanged across scales. If the continuous scale factors are fixed, the discrete scaling factor $\hat{C}_0$ and $\hat{C}_0'$ must be adapted in every scale accordingly to the sampling length. Recalling Equation (2.74), it comes, for two different scales j and k

$$C_0' = \hat{C}_0'\Big|^{\{j\}} l_s^{\{j\}} = \hat{C}_0'\Big|^{\{k\}} l_s^{\{k\}} . \tag{6.7}$$

Thus, for each scale, the discrete PSD scaling factor shall be updated by

$$\hat{C}_0'\Big|^{\{j\}} = \frac{l_s^{\{k\}}}{l_s^{\{j\}}} \hat{C}_0'\Big|^{\{k\}} . \tag{6.8}$$

An identical expression can be derived for the case of rough surfaces,

$$\hat{C}_0\Big|^{\{j\}} = \frac{l_{s_x}^{\{k\}}\, l_{s_y}^{\{k\}}}{l_{s_x}^{\{j\}}\, l_{s_y}^{\{j\}}}\, \hat{C}_0\Big|^{\{k\}} . \tag{6.9}$$

Summarizing, the single scale discrete spectrum is initially set up to verify a given RMS slope. At each scale, the current discrete scaling factor needs to be updated via Equations (6.8) and (6.9), in order to guarantee the continuity of the complete input PSD. As referred before, the RMS at each scale is necessarily smaller than the relative property of the complete topography. Additionally, the sum of the RMS slope squared at each scale must be equal to the single scale RMS slope squared—recall the definition of RMS slope and its relation with the spectral moments. The aforementioned conditions can be used to check if the topography at each scale is being generated correctly.

6.3 Multiscale numerical analysis with two scales

Having established the conditions for the definition of a RCE from its topography characteristics in Chapter 5, the numerical application of the multiscale approach can be performed, knowing beforehand how to set up each scale. Initially, the contact homogenization procedure is applied to the division of the original topography into only two scales, following the splitting rule introduced in Section 6.2.1. At this point, focus is placed into the influence of the topography characteristics λ_l, λ_s and H, inasmuch that the length, mesh and height are implicitly defined from the aforementioned representativeness assessment.

For the following numerical examples, only topographies with $H = 0.8$ were considered, since no significant effects of this parameter have been observed in the single scale results. Also, the single scale RMS slope is, again, fixed at 0.2, and the long cut-off wavelength is $\lambda_l = 5\,\text{mm}$. For each bandwidth $\lambda_l/\lambda_s \in [4,\ 8,\ 16,\ 32,\ 64]$, a single scale *Direct Numerical Simulation* (DNS) is performed, together with the multiscale (MS) approach for 2 scales. In Table 6.1, the number of elements in the non-mortar interface required to model the RCE in the two scenarios is shown. Despite that the bandwidth increases by a factor of 16 from the lowest bandwidth $\lambda_l/\lambda_s = 4$ to the widest spectrum $\lambda_l/\lambda_s = 64$, the size of the meshes used in the multiscale approach increases only by a factor of $\approx \sqrt{16} = 4$. Although, the multiscale solution comprises two solutions of a finite element problem with these smaller meshes. Ten realizations were generated for each scale of the multiscale approach and also for the DNS.

Before presenting and discussing the main results of this section, it is paramount to resume a previous topic on the numerical determination of the real contact area, first mentioned in Section 5.3.3. In sum, it has been remarked that the real contact area of the finite element model can be computed either based on a geometrical argument, on which only elements whose all nodes belong to the active set would count for the real contact area, or from the Lagrange multipliers. The latter strategy departed from equilibrium considerations in the direction of the applied external pressure and from the

Table 6.1: Number of elements in the non-mortar boundary for the single scale *Direct Numerical Simulations* (DNS) and the respective *Multiscale* (MS) multiscale approach.. For the MS case, the total number of elements is twice the shown in this table, yet they are considered in two separate simulations.

	λ_l/λ_s				
Case	4	8	16	32	64
DNS	264	528	1032	2064	4104
MS (×2)	144	192	264	360	528

physical interpretation of the Lagrange multipliers as the symmetric of the contact traction vector at the non-mortar interface—in particular, in frictionless contact, it relates to the contact normal pressure. The real contact area can then be computed as the product of the nominal contact area and the ratio between the nominal external pressure and the arithmetic mean of the Lagrange multipliers.

This issue is even more important in the context of the current multiscale approach. Not only the way of computing the real contact area at each scale is not unique, but also the downscaled mean contact pressure can be computed following identically alternative approaches. In particular, the mean contact pressure can be determined as the arithmetic mean of all contact pressures (Lagrange multipliers), or by dividing the external pressure by the real contact area fraction—which in turn can be computed following the already mentioned approachers. Although all methods are fundamentally equivalent, concerning a continuous description of contact, the numerical discretization inherently introduces deviations which remain unknown so far, in this dissertation.

In order to investigate the effect of the methodology for computing the individual scale results and transition variables on the homogenized solution, the multiscale approach has been tested, within the set of parameters referred in the begging of this section, in three distinct situations:

(i) Both the real contact area and the mean contact pressure are computed uniquely on the basis of the geometric definition of the contact area;

(ii) Both the real contact area and the mean contact pressure and computed uniquely on the basis of the Lagrange multipliers, i.e, relying on equilibrium considerations;

(iii) The real contact area is computed based on the geometrical argument, but the transition mean contact pressure is computed from the arithmetic mean of the Lagrange multipliers.

The DNS and multiscale results for all relevant scenarios are plotted in Figure 6.6, for the full load range, and with special attention to the light contact region. The case $\lambda_l/\lambda_s = 16$ is not plotted, for simplicity, but without loss of relevant detail. The differences between all methodologies are evident, even for the single scale DNS results. While this impacts the quantitative accuracy of the results, as different alternatives predict different values for the real contact area at the same external load, the qualitative behavior of the curves does not seem to be affected, if some consistency is preserved within the post-processing calculation. This is, if ether the geometrical contact area or the Lagrange multipliers are used individually to compute both the contact area and the scale the transition pressure. Under such conditions, it can be observed that this two pairs of curves behave nearly in the same fashion—e.g., by observing the response at light contact, the DNS and MS curves for each method are almost shifted versions of each other. In contrast, the curve relative to the mixed method, where the area is computed from the geometrical area and the transition pressure from the Lagrange multipliers, appears quite apart from the other cases, while converging to the area-based curve.

In Section 5.4, the discussion of the results relied, in fact, on the geometric argument for the contact area. Without any reference results, such as coming from experimental research, it is not feasible to select one methodology based on true accuracy of the outcome. Furthermore, the choice shall be between the methodologies solely based on either the geometrical area or on the contact pressure schemes, as suggested by Figure 6.6. Henceforth, the alternative based on Lagrange multipliers is adopted, for future convenience. In particular, this will prove interesting regarding the incorporation of more information from the contact pressure distribution in the transition variable—the contact pressures shall be used directly in such application.

Focusing now on the discussion of the numerical results, considering only the pressure-based computation, the multiscale results for two scales and the respective single scale DNS analogous are plotted in Figure 6.7, for several spectrum bandwidth. These curves are shown for all the contact area fraction range, i.e., from infinitesimal to full contact, with emphasis on the small load range—as in Figure 6.6. Roughly speaking, and looking only at the full load range, the multiscale solution curve is practically coincident with the single scale solution, until around $p_0 = 2 \times 10^4$ MPa, from where on slight divergence is observed. The two curves join again at full contact. From this broad point of view, the spectrum bandwidth does not seem to impact considerably the accuracy of the multiscale approach, which may be regarded as fairly acceptable, so far.

Restricting the attention to the light contact region, the apparently coincident solution are, actually, quite distant from each other. In addition, the bandwidth shows a clear effect on these results. With increasing bandwidth, the multiscale solution is observed to converge to the single scale DNS solution. This behavior can be interpreted with grounds on the principle of separation of scales. In profiles with short bandwidths ($\lambda_l/\lambda_s = 4,\ 8$), the longest length scale of the macroscale is not large compared with the largest length scale at the microscale. Therefore, the hypothesis of the application of a constant nominal pressure at the microscale associated with a certain load at the macroscale is not tightly valid. By increasing the bandwidth of the single scale topography and, consequently, the bandwidth of each scale, the largest wavelength at the macroscale is progressively more separated from the largest scale of the microscale. This way, the mechanical coupling between the extrema of each scale's spectrum are reduced, allowing the individual results for each scale to be homogenized in a single one.

The physical interpretation of the convergence of the results to the DNS with increasing bandwidth can be clarified by considering a rough topography (characterized by some

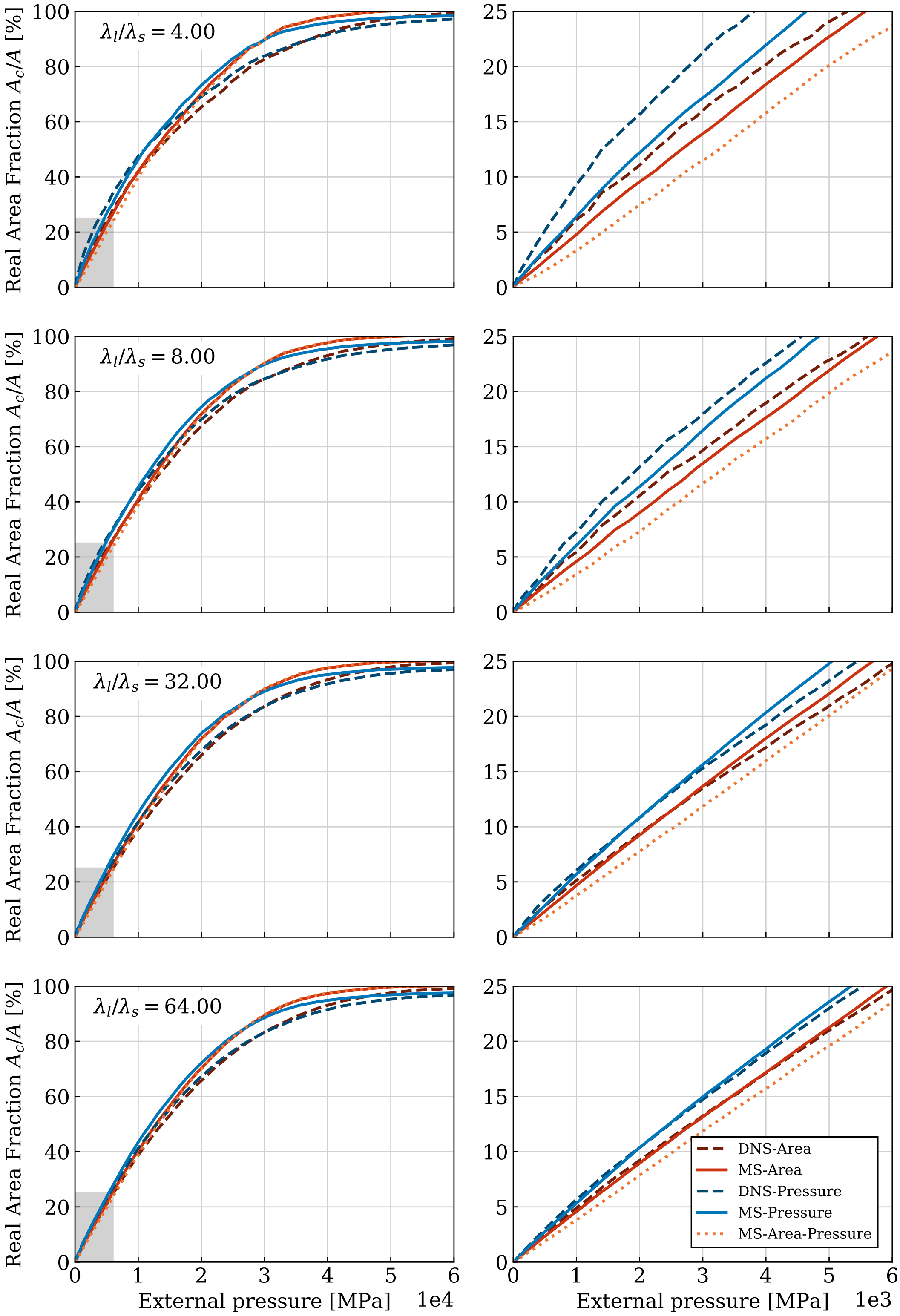


Figure 6.6: Comparison between different methods for computing the real contact area and the mean contact pressure in the scale transition.

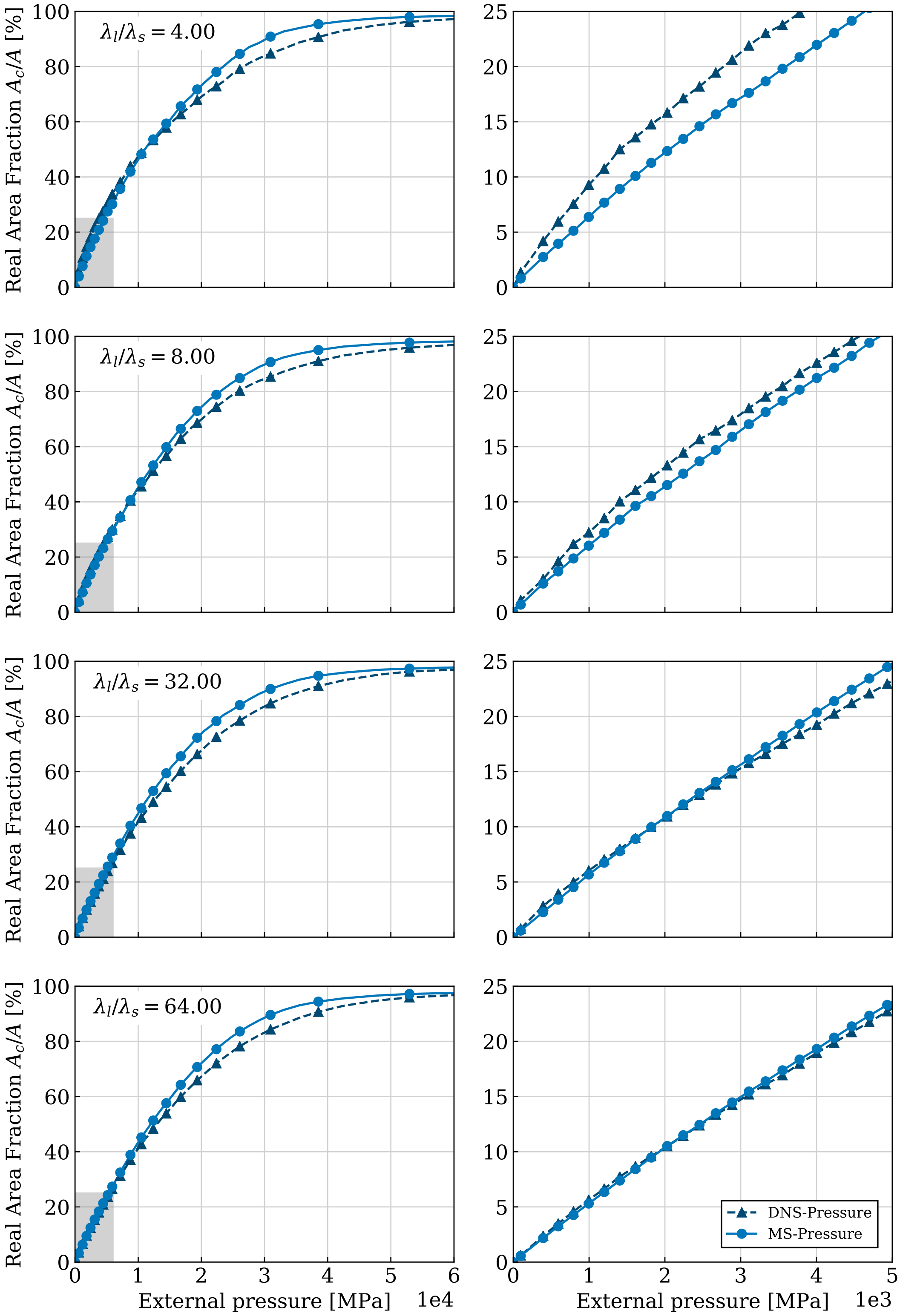


Figure 6.7: Comparison of the real contact area evolution (computed based on the Lagrange multipliers) computed with the multiscale approach by contact homogenization with two scales and the direct numerical simulation, for different spectrum bandwidths.

cut-offs k_l and k_s), from which the highest frequency k_s has been removed. The *filtered* spectrum is defined from the intact low cut-off k_l to the frequency $k_s^{(-)}$, which is slightly lower than the original cut-off k_s. In light contact, the load is mostly supported by the longest wavelengths around λ_l, due to their typically high amplitude. The shorter wavelengths are rapidly flatten at these initial contact spots. However, even though these short wavelengths cannot withstand significant loads, they do contribute considerably for the real contact area, since there are small gaps between the microasperities that are very difficult to close completely. Thus, if the *filtered* profile is loaded in the light contact region, by virtually including the extracted frequency k_s, no changes in the supported load are verified, but the contact area changes accordingly to the contact pressure verified at the contact spots. Such logic, however, assumes that $\lambda_s \ll \lambda_l$, otherwise the load supported by λ_s would be comparable with the amount supported by λ_l. This *mechanical coupling*, coming from the similitude in wavelength and, consequently, in the amplitude, is not accounted in the contact homogenization procedure, and explains, at some extent, why increasingly more accurate values are predicted for large bandwidths. For larger loads, towards the full contact, the DNS and multiscale approaches predict different solutions, possibly because the information passing strategy does not convey sufficient information of the current state of the contact to the next scales.

The error of the multiscale solution relative to the reference DNS is plotted in Figure 6.8. The reduction of the relative error with increasing bandwidth at load loads is substantiated by this figure. By approaching the full contact conditions, the large bandwidth is no longer associated with smaller errors—in fact, the relative errors are of the same order of magnitude of the lower bandwidths. The relative error decreases to zero with increasing loads, since the real contact area fraction is capped at 100%, which is naturally satisfied by both methods. It can be observed that the error goes to zero at a specific load, dependent on the bandwidth, where the MS and DNS curves intersect. However, this does not seem to be associated with any noteworthy feature.

The main goal of the multiscale approach is to reduce the computational resources required to solve the problem of rough contact. Thus, it is paramount to measure both the total simulation time required to complete the analysis, but also the maximum *Random Access Memory* (RAM) that needs to be allocated to the simulation. For this purpose, both quantities have been traced during the simulations, under the same circumstances. Namely, for every bandwidth, ten analysis have been run in parallel at each scale and also for the DNS, with the exception of the ratio $\lambda_l/\lambda_s = 64$. For this bandwidth, only five simultaneous parallel jobs have been executed, due to RAM constraints.

Figure 6.9 shows the computational resources required by the DNS and MS analysis, namely, the total simulation time and the peak RAM usage. The differences between the multiscale approach and the DNS are striking, regarding both the total simulation time and the maximum RAM requirements. The multiscale solution proves increasingly advantageous with increasing roughness bandwidth. Some bars on the RAM plot are not even visible, since they require less RAM than the minimum value resolved by the tracing strategy (about 160 MB). While for $\lambda_l/\lambda_s = 4$ the total simulation time is identical for both approaches, when the bandwidth is increased up to $\lambda_l/\lambda_s = 64$, the multiscale approach reduces the total simulation time from roughly 1 day to about 50 minutes. The same

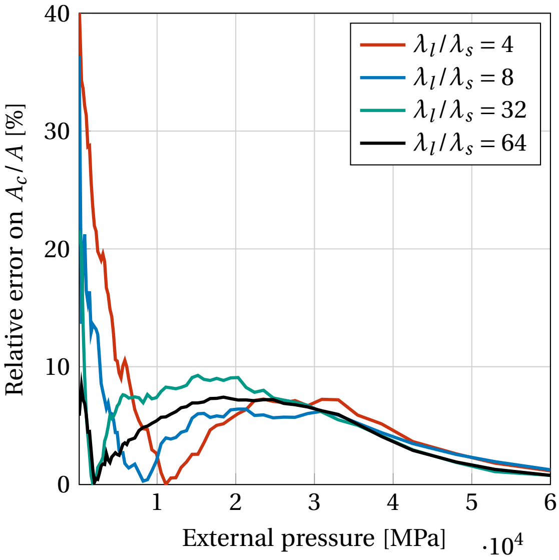


Figure 6.8: Relative error on the real contact area fraction computed with the multiscale approach, relative to the respective DNS solution, for different bandwidths.

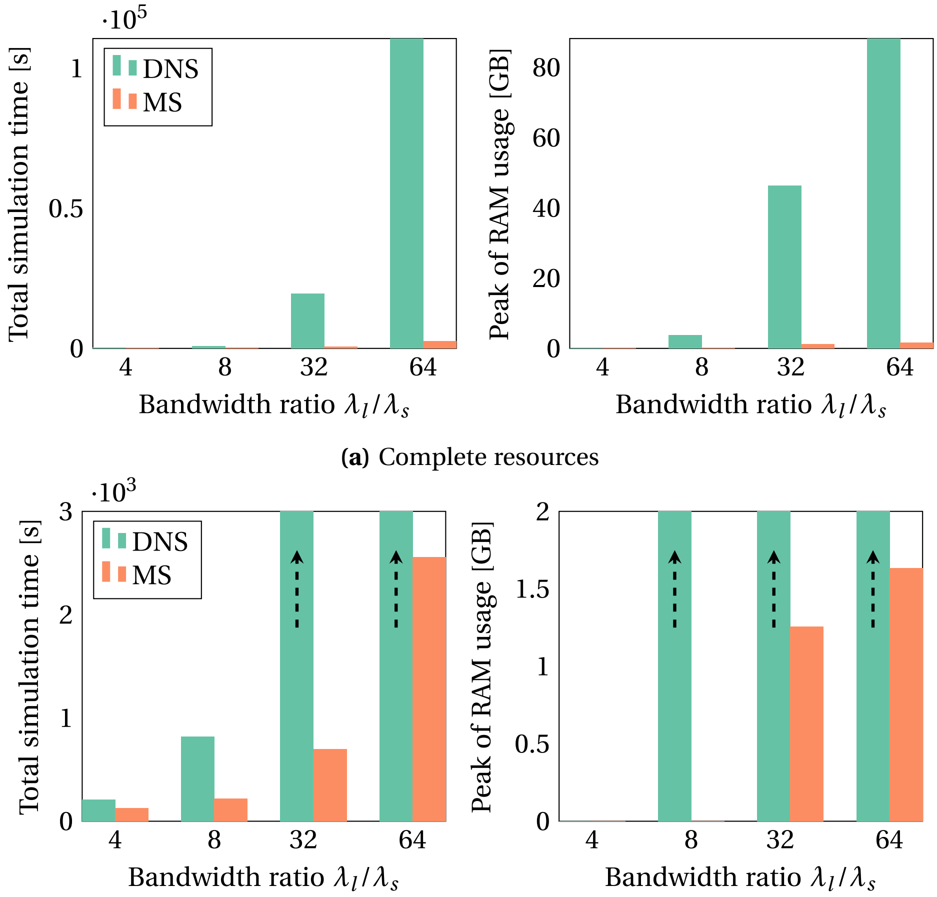
happens with the RAM requirements, which are reduced from 80 GB to simply 1.5 GB, for the same bandwidth.

Summarizing, the multiscale approach provides tremendous computational advantages over the classical DNS approaches, specially for the range of applications where the DNS methodology is the most questionable, i.e., wide roughness spectra. While the currently discussed multiscale contact homogenization scheme converges to the DNS solution at light contact with increasing bandwidth, an improved multiscale strategy which assures convergence for the DNS across all load range is to be proposed. Other great advantage of the general formulation of the multiscale scheme here treated is that it relies completely on post-processing operations. Hence, one can improve the multiscale algorithm, namely, the information passing scheme, and apply it to the already computed numerical results, bypassing any repetition of RCE analysis.

6.4 Enhancing the information passing scheme

It has been observed from the previous results that the information passing strategy adopted initially proves adequate for moderately low bandwidths at each scale level. Namely, at small, yet physically reasonable values of the contact area fraction, similar results can be obtained from the multiscale solutions and the respective DNS. The scale transition was performed following a *zeroth-order* approximation, where only the average contact pressure was passed to the inferior scales. In order to improve the multiscale solution for the contact area evolution, specially at large contact areas, more information shall be conveyed to the next scales.

In fact, if the contact pressures at each load step were normally distributed, the mean



(a) Complete resources

(b) Zoom in to the multiscale resources

Figure 6.9: Comparison of the total simulation time and maximum RAM usage of the DNS and multiscale approach. In the peak RAM usage plot, the boxes for the multiscale solution are not visible, because they require less then 160 MB of RAM, which is the resolution of the methods adopted to trace the RAM consumption.

contact pressure would contain almost all the information on the contact conditions at each scale. However, it is verified from the single scale results that the contact pressure distribution is strongly non-Gaussian throughout almost the entire load range. In Figure 6.10, the contact pressure distribution at each load step is plotted for two different single scale results with different bandwidths.[3] Both topographies are generated with $H = 0.8$, $\lambda_l = 5\,\text{mm}$ and the RMS slope is set to 0.2. It can readily be seen that the contact pressure distribution is not symmetric, and thus, as a consequence, it cannot be Gaussian. The mean contact pressure, computed either from the arithmetic mean or from equilibrium considerations, is slightly higher than the mode of the distribution—the spatially

[3] This figure is obtained by first dividing the contact pressure values in 20 evenly spaced ranges between 0 and the maximum contact normal pressure at each load increment. Then, probability density is computed, and plotted at the center of the contact pressure bin. This justifies the blank regions at very low and very high contact pressures.

most frequent contact pressure. The major impact of the bandwidth is on the overall noise of the distribution. This comes from the different discretization resolution considered for the two cases. The larger λ_l/λ_s requires a denser mesh and, hence more points enter the calculation, resulting in a smoothing effect. Note that at the end of the spectrum range, the contact pressure distribution approaches a Gaussian curve, and the external pressure is almost equal to the mean contact pressure—the contact area fraction is nearly 100%.

At each load increment, the contact pressure field is generally distributed through different disconnected regions at the contact interface. The microscale setup can be thought as being loaded by the local pressure distribution, at every region. The previously discussed information passing strategy, basically, homogenized all possible load distributions into a single uniform pressure loading case. Yet, from Figure 6.10 it is concluded that some values of pressure are spatially more prevailing that others—in particular, the mean contact pressure is not the most frequent value.

In order to include more data in the scale transition, a new information passing strategy is proposed. It must be emphasized that the following methodology is uniquely based on post-processing, hence, all numerical results obtained for the application of the former multiscale approach are still valid—a great advantage over FE^2 strategies.

Starting at the macroscale roughness level, the contact area is computed from the FE models. At each point of the curve, i.e., at each value of the external pressure and real contact area of scale 1, the distribution of contact pressures is compacted in n^p equally spaced bins, between 0 and the maximum contact pressure. Each bin will be represented by its center value. This step corresponds to a discretization of the contact pressure spectrum. At each discrete contact pressure value, the *multiscale solution up to scale 2* for the real contact area is evaluated. The multiscale solution is updated with the results from scale 1, by performing the multiplicative homogenization step, yet with an additional weighted average of the contact areas at the smaller scales,

$$\left\{\frac{A_c}{A}\right\}^{\mathrm{MS},1}\left(p_0^{\{1\},i}\right)=\left\{\frac{A_c}{A}\right\}^{\{1\}}\left(p_0^{\{1\},i}\right)\cdot\left(\sum_{n=1}^{\mathsf{n}^\mathsf{p}} f_{i,n}^{\{1\}}\Delta p_{i,n}^{\eta\{1\}}\left\{\frac{A_c}{A}\right\}^{\mathrm{MS},2}\left(p_{i,n}^{\eta\{1\}}\right)\right). \tag{6.10}$$

The newly introduced notation follows:

- $\left\{\frac{A_c}{A}\right\}^{\mathrm{MS},j}$ Multiscale solution for the real contact area fraction for scales smaller and including j;
- $f_{i,n}^{\{j\}}$ Probability density of the n-th discrete contact pressure value at scale j and load increment i;
- $\Delta p_{i,n}^{\eta\{j\}}$ Width of the n-th contact pressure bin at scale j and load i;
- $p_{i,n}^{\eta\{j\}}$ n-th discrete contact pressure bin at scale j and load i.

The multiscale solution up to scale 2 can be computed by applying Equation (6.10) at scale 2, which would then be expressed in terms of the multiscale solution up to scale 3.

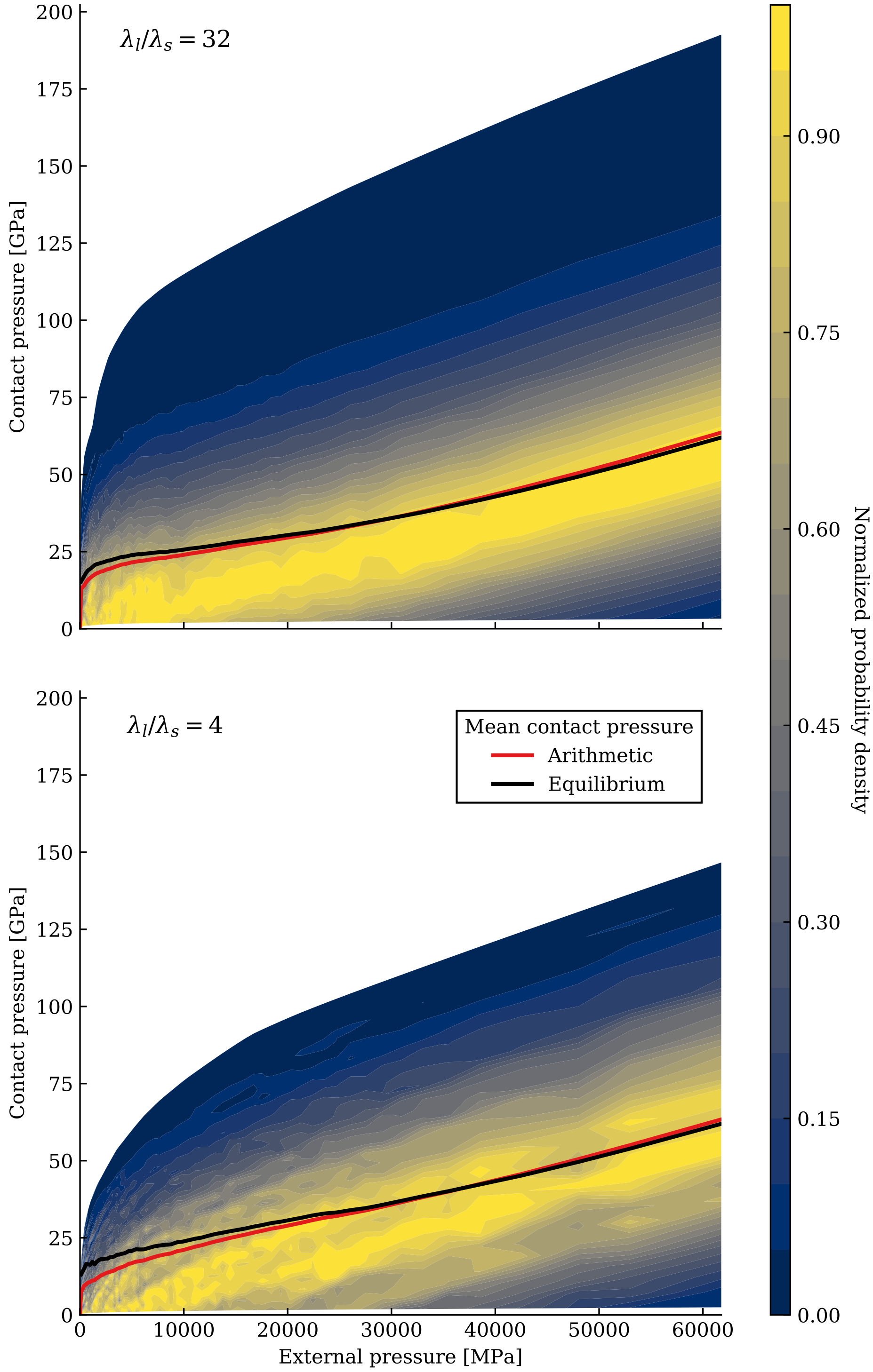


Figure 6.10: Contact pressure distribution for several values of the nominal external pressure, and two different bandwidths. The results come from single scale FE simulations with $H = 0.8$, $\lambda_l = 5\,\text{mm}$ and RMS slope equal to 0.2. For each load increment, the contact pressure distribution is normalized to verify a maximum unit value. The mean contact pressure computed both from the arithmetic mean and from equilibrium consideration are plotted, as well.

It can be conclude that Equation (6.10) must be applied recursively at all scales, down to the second smallest scale ($\mathsf{n^s} - 1$), since the multiscale solution at scale $\mathsf{n^s}$ is simply the single scale result. This leads to definition of the following recursive expression for the multiplicative weighted homogenization scheme:

$$\begin{aligned}
\left\{\frac{A_c}{A}\right\}^{\mathrm{MS},j}\left(p_0^{\{j\},i}\right) &= \left\{\frac{A_c}{A}\right\}^{\{j\}}\left(p_0^{\{j\},i}\right)\cdot\left(\sum_{n=1}^{\mathsf{n^p}} f_{i,n}^{\{j\}}\Delta p_{i,n}^{\eta\{j\}}\left\{\frac{A_c}{A}\right\}^{\mathrm{MS},j+1}\left(p_{i,n}^{\eta\{j\}}\right)\right), \\
&\qquad \text{for } j = 1,\dots,\mathsf{n^s}-1\,, \\
\left\{\frac{A_c}{A}\right\}^{\mathrm{MS},j}\left(p_0^{\{j\},i}\right) &= \left\{\frac{A_c}{A}\right\}^{j}\left(p_0^{\{j\},i}\right), \quad \text{for } j = \mathsf{n^s}\,.
\end{aligned} \tag{6.11}$$

Figure 6.11 shows the schematics of the enhanced information passing strategy. Only one transition is displayed, due to its inherent recursive character. Note that the index i refers to all quantities related with the macroscale load p_0^i, and not to the i-th load increment. Another key point to mention is that the factors $f_{i,n}^{\{\mathsf{n^s}-1\}}\Delta p_{i,n}^{\eta\{\mathsf{n^s}-1\}}$ represent the weights of each contact area fraction value, which must verify

$$\sum_{n=1}^{\mathsf{n^s}} f_{i,n}^{\{\mathsf{n^s}-1\}}\Delta p_{i,n}^{\eta\{\mathsf{n^s}-1\}} = 1\,. \tag{6.12}$$

From a practical point of view, and establishing the link with the computer implementation of this strategy, the procedure starts at the microscale, where the multiscale solution up the to that scale is known. Moving to the next upper scale, the discrete contact pressure values are identified at each load step, jointly with the weights $f_{i,n}^{\{\mathsf{n^s}-1\}}\Delta p_{i,n}^{\eta\{\mathsf{n^s}-1\}}$. Then the real contact area fraction at the multiscale solution in the smaller scale is interpolated at the discrete contact pressure values of the current scale, and the homogenization step is applied, updating the multiscale solution. This procedure is repeated for all scales until scale 1, ending with the homogenization step at the macroscale.

6.5 Multiscale numerical analysis with the improved transition scheme

The assessment of the improved contact homogenization procedure follows the same the structure adopted in Section 6.2, with an additional degree of freedom in the analysis, namely, the number of discrete values of contact pressure $\mathsf{n^p}$. The results for all the examined bandwidths and different levels of discretization of the contact pressure distribution $\mathsf{n^p}$ are compared with the DNS solution and the previous MS result (MS-pressure) in Figure 6.12.

When only one value is chosen to represent the full pressure spectrum, i.e., $\mathsf{n^p} = 1$, the contact area is overestimated, for all cases. In fact, in such scenario, the value which is being used in the scale transition equals half the maximum contact pressure at each load step, whose spatial frequency is very low—cf. Figure 6.10. Thus, in the homogenization step, it is being considered that the microscale is loaded uniquely with an excessively high nominal pressure, therefore, overpredicting the real contact area fraction.

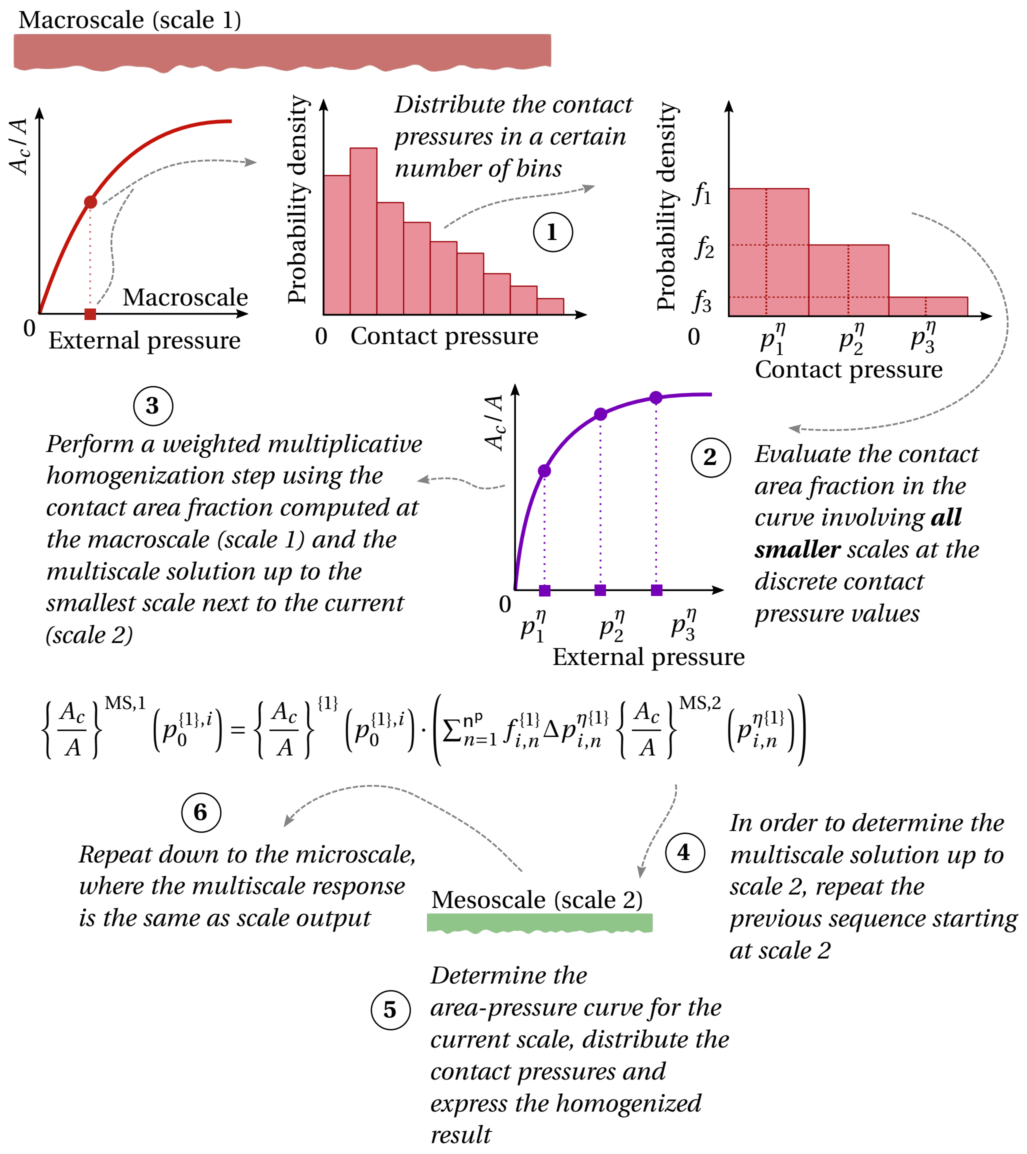

Macroscale (scale 1)
A_c/A
Macroscale
0
External pressure
Probability density
Distribute the contact pressures in a certain number of bins
1
0
Contact pressure
Probability density
f_1
f_2
f_3
0
p_1^η
p_2^η
p_3^η
Contact pressure
3
Perform a weighted multiplicative homogenization step using the contact area fraction computed at the macroscale (scale 1) and the multiscale solution up to the smallest scale next to the current (scale 2)
A_c/A
0
p_1^η
p_2^η
p_3^η
External pressure
2
Evaluate the contact area fraction in the curve involving **all smaller** scales at the discrete contact pressure values
$\left\{\frac{A_c}{A}\right\}^{\mathrm{MS},1}\left(p_0^{\{1\},i}\right) = \left\{\frac{A_c}{A}\right\}^{\{1\}}\left(p_0^{\{1\},i}\right)\cdot\left(\sum_{n=1}^{\mathsf{n}^{\mathsf{p}}} f_{i,n}^{\{1\}}\Delta p_{i,n}^{\eta\{1\}}\left\{\frac{A_c}{A}\right\}^{\mathrm{MS},2}\left(p_{i,n}^{\eta\{1\}}\right)\right)$
6
Repeat down to the microscale, where the multiscale response is the same as scale output
4
In order to determine the multiscale solution up to scale 2, repeat the previous sequence starting at scale 2
Mesoscale (scale 2)
5
Determine the area-pressure curve for the current scale, distribute the contact pressures and express the homogenized result

Figure 6.11: Improved information passing strategy for the multiscale solution of the real contact area fraction. At each scale, the contact pressure spectrum is discretized, and the discrete values are passed to the multiscale solution involving all inferior scales. The multiplicative homogenization step is performed in a weighted average sense.

It is interesting to note that the homogenized solution converges with increasing number of points in the pressure spectrum—the curves for $\mathsf{n}^{\mathsf{p}} = 10$ and $\mathsf{n}^{\mathsf{p}} = 20$ are practically indistinguishable. The converged multiscale solution, however, deviates slightly from the DNS at light contact, but for progressively larger loads, it gives a good approximation of the single scale solution—at least, with smaller error than the mean pressure-based multiscale approach. Despite that the converged solution does not provide an exact estimate for the area evolution at low loads, when the homogenization is performed with $\mathsf{n}^{\mathsf{p}} = 3$,

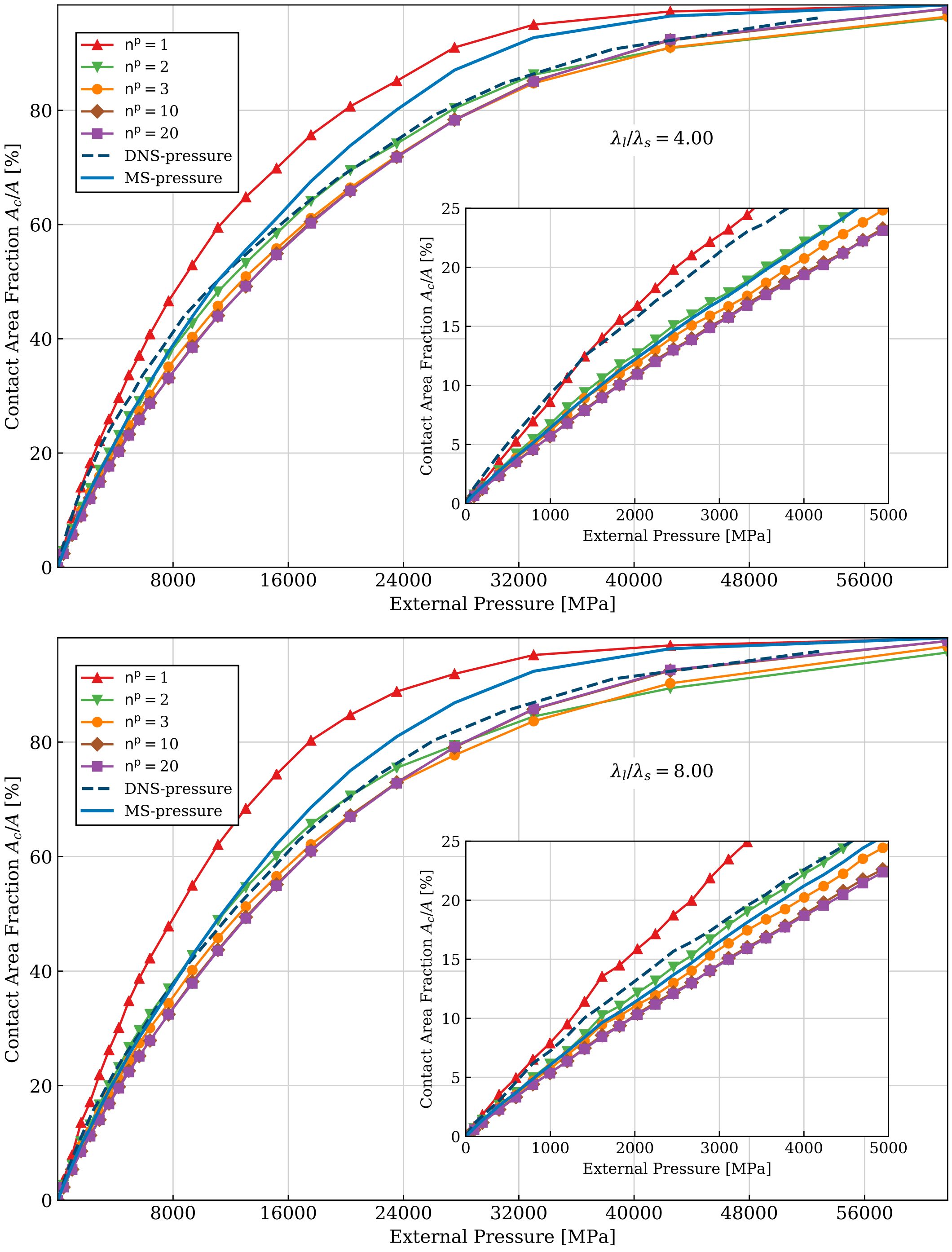


Figure 6.12: Contact area evolution with pressure, computed with the improved contact homogenization procedure, for different bandwidths and levels of discretization of the contact pressure distribution n^p.

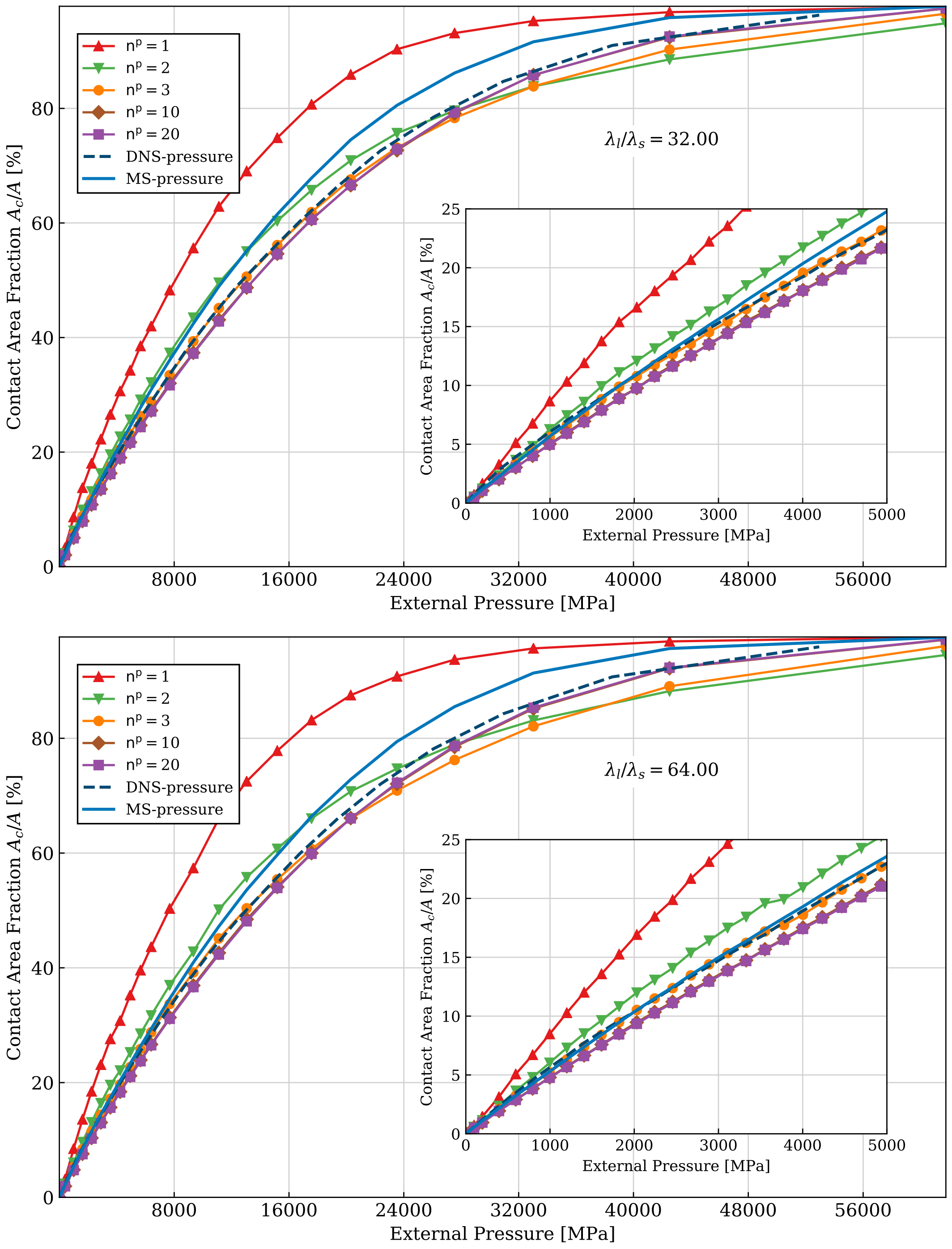


Figure 6.12: Contact area evolution with pressure, computed with the improved contact homogenization procedure, for different bandwidths and levels of discretization of the contact pressure distribution n^p (continued).

the DNS solution is recovered almost exactly, for the two largest bandwidths $\lambda_l/\lambda_s = 32$ and $\lambda_l/\lambda_s = 64$. Unfortunately, the same improvements are not verified if the spectrum is narrow, and, overall, the new transition scheme does not prove to be a better solution comparatively with the previous multiscale approach.

All in all, the weighted average multiplicative homogenization strategy gives accurate solutions for the rough contact problem by tuning the discretization of the contact pressure spectrum, as long as the roughness spectrum bandwidth is sufficiently large. The homogenized results converges with the number of discrete points in the contact pressure distribution, and such solution approximates the DNS curve with high accuracy for relatively high loads, towards the full contact conditions. The converged area curve does not capture the single scale results exactly, in the region of light contact, and the mean pressure-based strategy (Section 6.2) provides more reliable results—for large ratios λ_l/λ_s. Nonetheless, if the contact pressure distribution is represented by 3 bins, i.e., $\mathsf{n}^{\mathsf{p}} = 3$, the improved scheme reproduces almost exactly the DNS solution, at the low pressure regime, for large bandwidths. These conclusions are substantiated by Figure 6.13, where the error relative to the DNS solution for different multiscale solutions computed with the improved strategy are plotted against the relative error of the mean pressure-based scheme.

As the bandwidth is crucial in assuring the quality of the multiscale solution, for both scale transition schemes, one can establish the minimum bandwidth required for each spitted scale, based on the results in Figure 6.12. Since the bandwidth of each individual scale is constant, according to the splitting rule in Section 6.2.1, and observing that the accurate results from the homogenization step are obtained starting at a minimum of $\lambda_l/\lambda_s = 32$, it can then be stated:

Minimum bandwidth for each individual scale
Each scale j resulting from the PSD splitting procedure must obey a minimum bandwidth of

$$\left\{\frac{\lambda_l}{\lambda_s}\right\}^{\{j\}} \geq \sqrt{32} \approx 5.6\,, \tag{6.13}$$

so that the homogenization step produces accurate values, relative to the numerical solution of the single scale problem.

6.6 Application to wide spectra and extension to several scales

It has been concluded that the multiscale approach, employed either with the original or improved information-passing formulation, can provide accurate results for the single scale solution, if the spectrum at each scale is wide enough. Furthermore, the computational advantages of the multiscale approach are self-evident (cf. Figure 6.9). All these conclusions have been extracted from a setup with only two scales. At this stage, having validated the multiscale approach, one shall aim at, first, applying this technique to more than two scales, and second, at solving problems out of the range of the typical direct numerical strategies—such as the BEM.

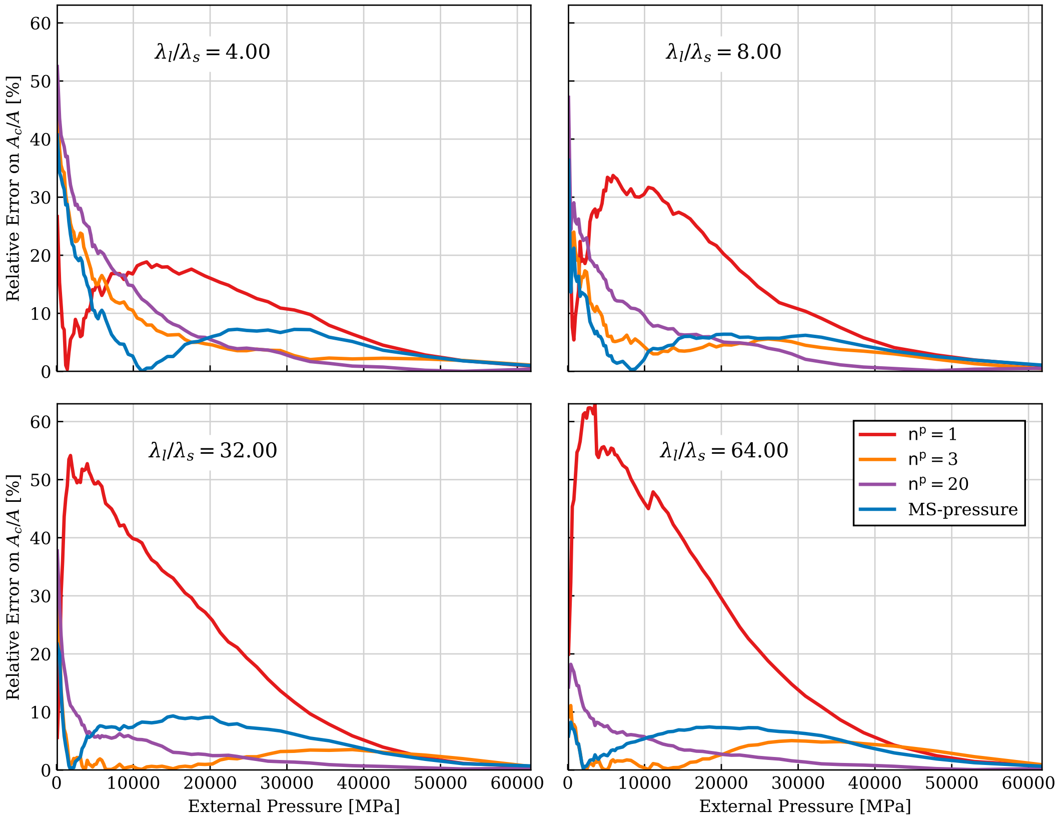


Figure 6.13: Error of the improved homogenization scheme relative to the DNS solution, for different levels of discretization of the contact pressure distribution.

So far in this dissertation, the maximum spectrum width considered was $\lambda_l/\lambda_s = 64$, which is also the cap commonly employed by other authors. Roughly speaking, this indicates that the topography contains roughness features which extend through approximately two orders of magnitude, say between 1 mm and 10 µm. In line with the roughness description in Chapter 2, roughness details can be as small as the size of atoms, meaning that real spectrum may cover much more than two orders of magnitude. It should not be overlooked that the continuum hypothesis, and the idealization of an homogeneous material cannot be conveyed to all the smaller scales. As soon as the roughness length scale reaches the size of metal grains, the homogeneous description of matter must be dropped, and the continuous hypothesis is violated at the atomic scale. The present section proceeds without accounting for any of the previous complexities, since it is only intended to analyze the performance of the multiscale strategies under more severe conditions.

For this purpose, a rough profile with $H = 0.8$, $\sqrt{m_2} = 0.2$, $\lambda = 5$mm and having an extremely large bandwidth $\lambda_l/\lambda_s = 4096$, covering length scales across almost four orders of magnitude, and containing roughness features from about 1 mm down to 1 µm is examined. The rough topography in question is shown in Figure 6.14, together with the successive magnifications of one order of magnitude each, emphasizing its level of detail. The PSD of this profile can be split into 4 scales, while assuring a minimum bandwidth of $\lambda_l/\lambda_s = 8$ in each one. The multiscale strategy is applied by splitting the topography into 2, 3 and 4 scales, separately, and for each of the previous number of scales n^{s}, the homogenized result is computed with both transition schemes. The largest FE sub-problems used within the current application of the multiscale approach to this problem are, naturally, the ones concerning the splitting into two scales, and hold about 4100 elements in the non-mortar interface. The number of elements is reduced to approximately 1000 and 500 for the splitting into 3 and 4 scales, respectively. If the single scale solution were to be found from DNS, it would require a mesh of nearly with 200 000 elements in the non-mortar boundary, rendering this problem extremely difficult to address. The results for the multiscale solution following the initial and improved transition schemes are plotted in Figure 6.14.

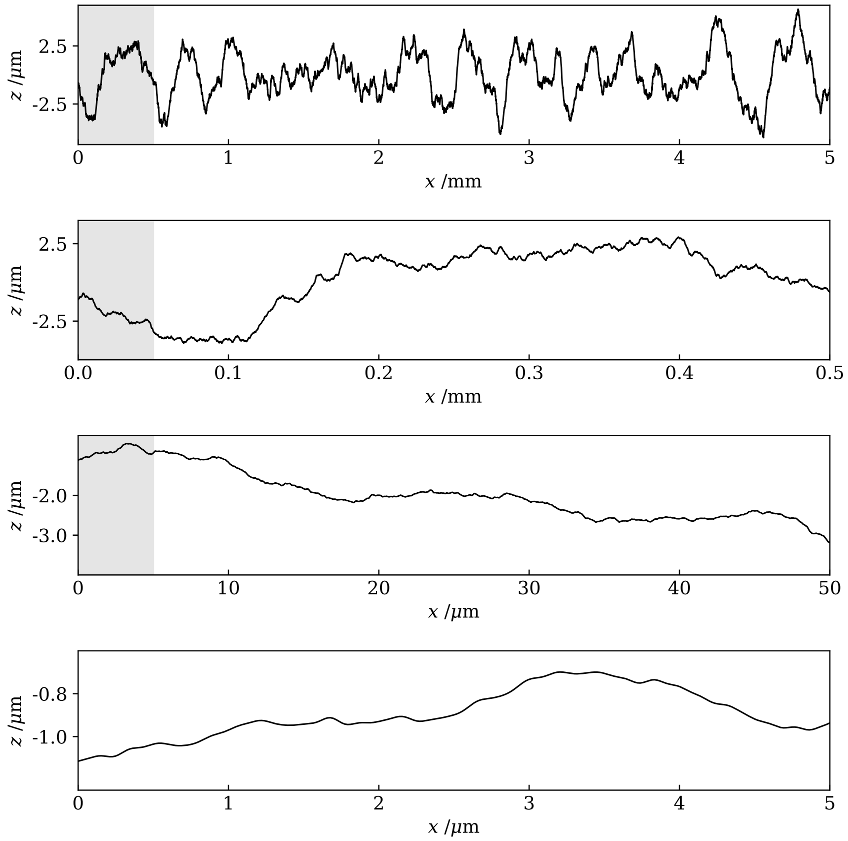
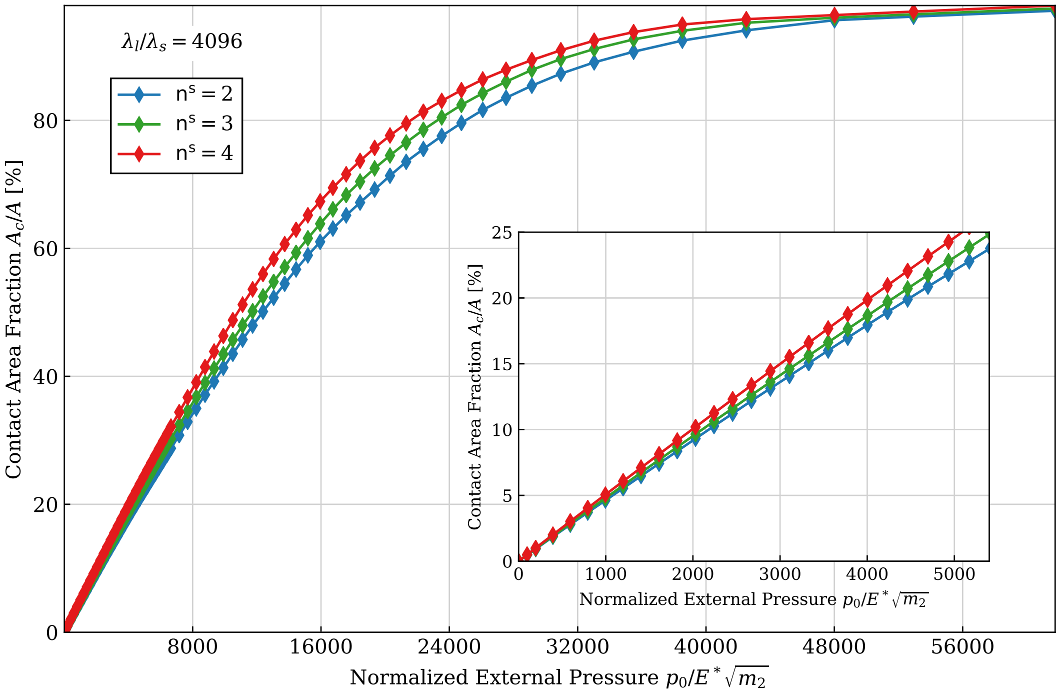
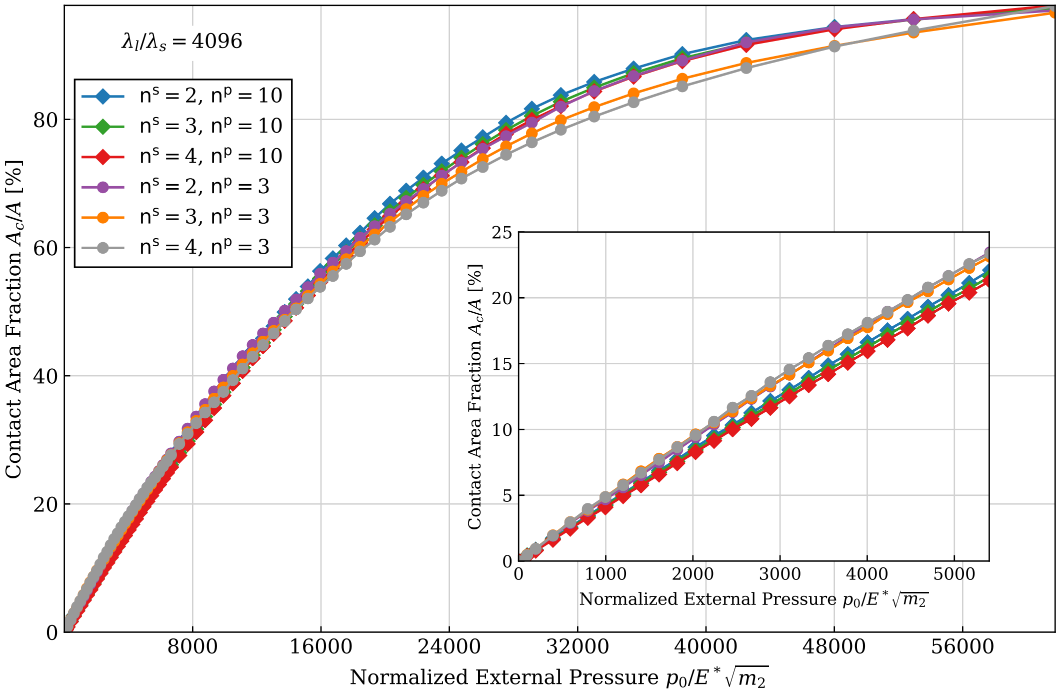


Figure 6.14: Self-affine profile with $L = 5\,\mathrm{mm}$, $H = 0.8$, $\sqrt{m_2} = 0.2$ and $\lambda_l/\lambda_s = 4096$, at different magnifications. The numerical model for the RCE of this topography would require about 200 000 elements at the non-mortar interface. Each subplot shows the rough profile at the previous (above) shaded region.

Intuitively, as long as the scales are well separated (cf. Equation (6.13)), one could expect the results to be mostly independent of the number of scales considered for the



(a) Scale transition with mean pressure



(b) Improved scale transition

Figure 6.15: Contact pressure evolution with pressure of a wide spectrum topography ($\lambda_l/\lambda_s = 4096$), computed with the mean pressure-based and improved scale transition strategies, for different numbers of scales n^s.

problem. Starting the discussion with the initial transition scheme, in Figure 6.15a, it can be visualized that homogenized contact area evolution depends on the number of scales considered, in particular, for high nominal pressure. In fact, in the region of infinitesimal contact, represented in the inset plot, all the curves are very similar, in particular, for two and three scales. However, the distance between the curves seems to increase monotonically with the load, extending until near the full contact.

The improved homogenization is considerably more insensitive to the number of scales, than the pressure-based alternative. In fact, the curves are almost coincident throughout all the load range, and for all number of scales examined. Two different levels of discretization of the pressure distribution were considered, namely $\mathsf{n}^{\mathsf{s}} = 10$, which is verified to represent the converged solution, and also $\mathsf{n}^{\mathsf{p}} = 3$, based on previous observations that it represents an optimum value for obtaining accurate solutions. Both results with $\mathsf{n}^{\mathsf{p}} = 3$ and $\mathsf{n}^{\mathsf{p}} = 10$ are essentially independent of the number of scales. Comparing these results with the ones computed with the initial homogenization sequence, it can be noted that by considering $\mathsf{n}^{\mathsf{p}} = 3$, the curves at the low pressure region are moved towards the solution for $\mathsf{n}^{\mathsf{s}} = 2$ of the pressure-based transition results, which is arguably the best estimate of the single scale solution (see Figure 6.12). In addition, it is paramount to observe that the results for all n^{s} are coincident for $\mathsf{n}^{\mathsf{p}} = 3$, in the infinitesimal contact regime.

In sum, in this section, interest has been placed on the application of the multiscale approaches to modeling large problems, practically out of range of DNS strategies. The improved contact homogenization method proves more satisfactory than the initial pressure-based strategy, inasmuch that it is not sensible to the number of scales considered, and predicts tightly accurate results, specially when using the optimal value n^{p}=3.

6.7 Influence of roughness bandwidth in the real contact area

It has been shown in the previous section that the multiscale strategies can be used to model rough contact involving wide roughness power spectra. In particular, the improved scheme seems very convenient to perform such studies, since it has been verified that it is practically independent of the number of scales involved in multiscale setup, as long as the bandwidth at each scale is sufficiently large. An interesting topic that can now be addressed is how the real contact area changes as the width of the power spectrum increases—or the short cut-off wavelength λ_s decreases. Note that in this scenario, the RMS slope is not constant for each topography, since only the high cut-off is increasing, while the rest of the PSD remains unchanged. This is equivalent at measuring the roughness profile with increasing resolution.

In order to assess numerically the effect of the roughness bandwidth in the real contact area, using the improved multiscale scheme with $\mathsf{n}^{\mathsf{p}} = 3$, a single topography case was considered, with $H = 0.8$, $\lambda_l = 5\,\text{mm}$, and five different bandwidth ratios between 512 and 8192, inclusively. The RMS slope for the largest ratio is fixed at 0.2 and for the remaining cases it must computed according to the respective ratio and the scale factor set by the largest bandwidth. Note that for large ratios, e.g., $\lambda_l/\lambda_s = 8192$ and $\lambda_l = 5\,\text{mm}$, it follows that $\lambda_s \approx 0.61\,\mu\text{m}$. At this length scale (actually, even for smaller bandwidth ratios), the hypothesis of an homogeneous media can naturally be questioned, as the mesh size would be smaller than the size of microconstituents. Nevertheless, this shall not be considered here, since simple qualitative results are to be obtained.

Figure 6.16 shows the results for the contact area evolution with nominal pressure for all examined cases, up to moderately high loads. While it could be hypothesized that the contact area response would saturate for increasingly wider roughness power spectrum, following the idea that from a certain short cut-off wavelength on, all smaller wavelengths would be flattened out. However, this is not verified, and it can be observed in Figure 6.16 that the real contact area decreases monotonically with increasing bandwidth, for the same applied nominal pressure. In other words, all roughness wavelengths contribute for real contact area, no matter how short they are.

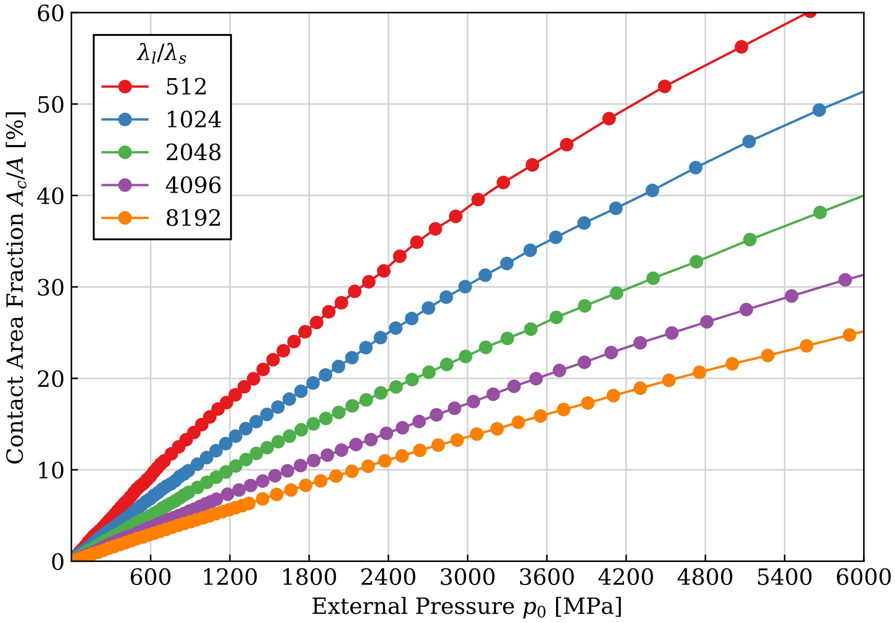


Figure 6.16: Evolution of the real contact area with nominal pressure for an increasingly wide roughness power spectrum.

One can also interpret theses results by plotting the normalized external pressure in the horizontal axis, as illustrated in Figure 6.17. All results follow a very similar linear trend, after the introduction of the non-dimensionalization of external pressure. Thus, even though the RMS slope is different across all cases, since the short cut-off changes, the non-dimensionalized results are mostly independent of the bandwidth, in the light contact region (low pressure). The deviations between the results for each bandwidth may come from inherent errors in the multiscale strategy, and can also be related to the dependency of the real contact area evolution with the Nayak's parameter α, as suggested in Yastrebov, Anciaux, *et al.* (2017).

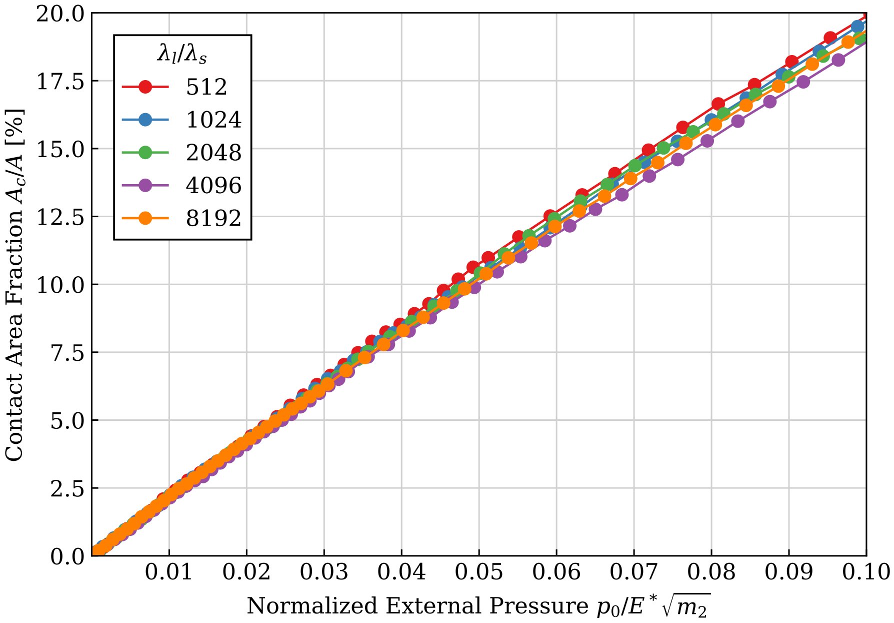


Figure 6.17: Evolution of the real contact area with normalized nominal pressure for an increasingly wide roughness power spectrum, in the low pressure (light contact) region.

6.8 Three dimensional analysis

From early on in this dissertation, it has been emphasized that rough contact is inherently a three dimensional problem. Due to their computational attractiveness, a great effort has been put in performing 2D rough contact simulations, in order to extract general qualitative and quantitative information. In fact, the main downside of the 3D micromechanical contact models is the high computational complexity, which motivates the application of multiscale strategies. To complete the numerical framework for modeling rough contact developed during the present work, the aforementioned multiscale techniques are applied to 3D problems. A key point to mention is that the formulation of the multiscale approaches does not depend on the number of dimensions of the problem, hence they can readily be applied to 3D cases. Computational resources scale very rapidly with the size of 3D problems, thus, they pose serious constraints on the range of topography parameters which can be tested. In fact, in the following, only multiscale solutions are to be found, and these are to be assessed via the comparison with reference numerical results, namely from Yastrebov, Anciaux, *et al.* (2015).

The contact of a self-affine elastic rough surface characterized by $H = 0.8$, $\sqrt{m_{02} + m_{20}} = 0.2$, $\lambda_l = 5\,\text{mm}$ and bandwidth $\lambda_l/\lambda_s = 64$ with a rigid flat base is the problem considered throughout this section. A division into 3 scales is adopted here, each with bandwidth equal to 4, and the RCE length set to 4 times the low cut-off at each scale. Note that the two previous specifications violate both the criterion regarding the scale splitting and the minimum RCE length. On the one hand, it should not be overlooked that those rules have been established for 2D contact, and are only means of estimating 3D RCE characteristics. On the other hand, some of the conditions established in the course of this thesis are conveniently relaxed here, in order to allow the numerical treatment of the problem.

Also the mesh size is required to be much smaller than the ideal. The non-mortar interface of the macroscale is meshed with 60 × 60 elements, while the second and third scales are meshed with 33 × 33 elements. Observe that as $L/\lambda_l = 4$, and $\lambda_l/\lambda_s = \sqrt[3]{64} = 4$, then $L/\lambda_s = 16$. Thus, the macroscale is meshed with a minimum of approximately 4 nodes *per* asperity (2 in each direction), and the scales 2 and 3 are guaranteed a minimum of one node *per* asperity—as in Hyun, Pei, *et al.* (2004), Pei *et al.* (2005), and Hyun and Robbins (2007). Fewer elements are used for scale 2 and 3 since this has been verified to stabilize the convergence of the algorithm. Owing to such coarse discretization on the non-mortar boundary, the high resolution region is made up by a single layer of elements. Only one RCE realization is used for scale 1 (the most nicely behaved), while 3 topography realizations are considered for the scale 2 and 3. The finite element meshes are illustrated in Figure 6.18. The homogenized results are presented in Figure 6.19, side by side with several analytical theories and the results from Yastrebov, Anciaux, *et al.* (2015). For simplicity, and also based on the physical relevance of this contact area regime, only the light contact results are presented.

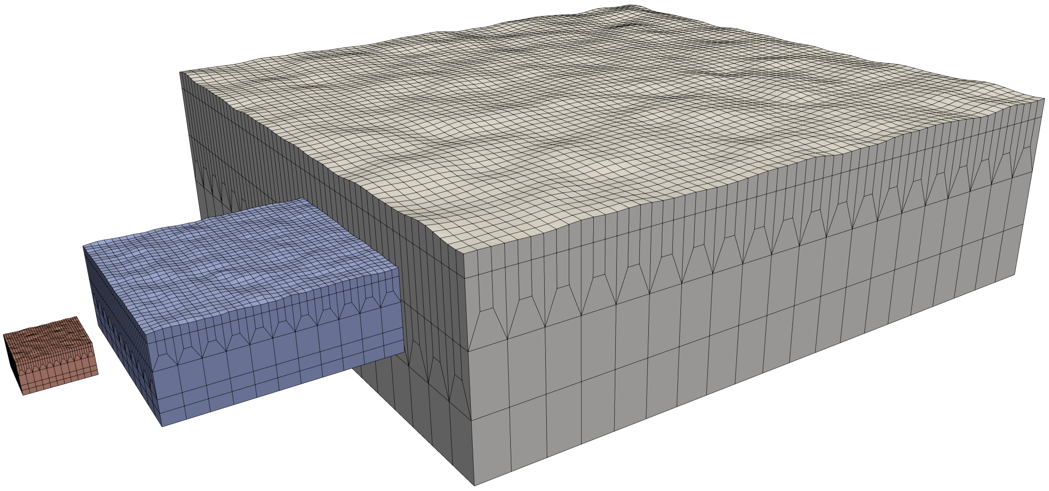


Figure 6.18: Three dimensional finite element meshes for each scale used in the multiscale numerical simulations.

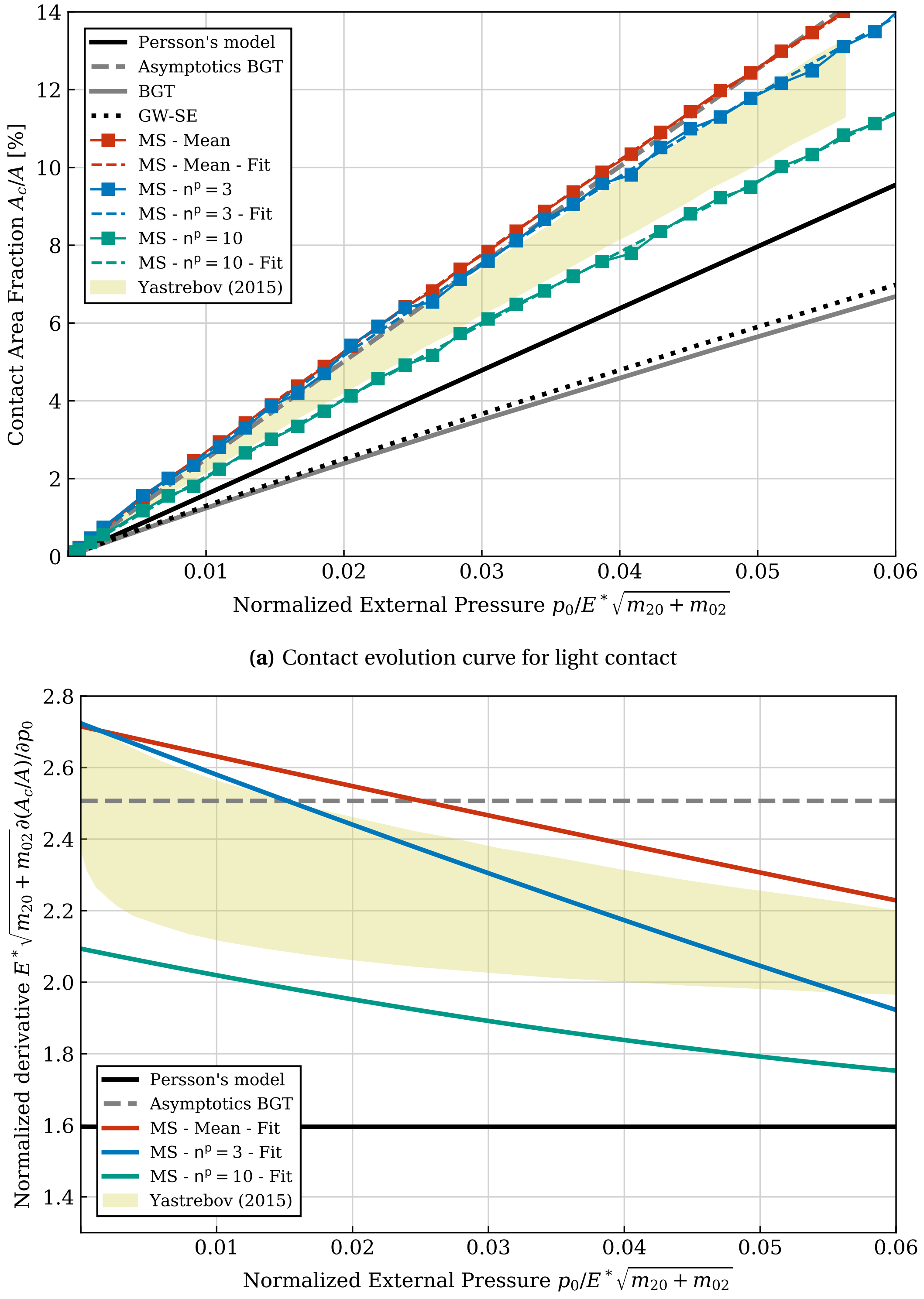
Taking into consideration all the restrictions which had to be imposed on the numerical model, the multiscale results fall fairly near the reference results from Yastrebov, Anciaux, *et al.* (2015). The mean pressure transition scheme and the improved strategy with $\mathsf{n}^\mathsf{p} = 3$ predict similar values, which are larger than the results from the converged improved scheme ($\mathsf{n}^\mathsf{p} = 10$)—has likewise realized for 2D contact. Additionally, recall that poorly discretized RCEs lead, in general, to the overestimation of the real contact area (cf. Figure 5.17). From this result, it can be argued that the numerical solutions determined by the multiscale strategy could provide a better approximation of the reference values, if the RCE dimensions were set to the calibrated standards.

The results published in Yastrebov, Anciaux, *et al.* (2015) are presented following a philosophy slightly different from the one adopted here, defining an array of low and high cut-offs wavelength, thus repeating some ratios λ_l/λ_s. For this reason, the graphical representation of the their results, in Figure 6.19, corresponds to a shaded region, which embeds all the points in the original publication. As a word of caution, in the aforementioned work, the authors introduce the same set of curves twice. In their second iteration of the plot, the contact area is corrected with some factor. The region plotted in Figure 6.19 concerns the first, uncorrected, set of contact area curves,

Regarding the analytical model, it can readily be seen that the original BGT and the GW-SE are clearly not fit to describe the contact area evolution in this case. In fact, for the considered topography, the spectral breadth is $\alpha = 73.7$, and for such large values of this parameters, it has already been discussed that the results diverge significantly from well-known asymptotic limits (see Chapter 4). Both the asymptotics of the BGT model and Persson's theory provide approximate upper and lower bounds, respectively, for the real contact area curves, and the homogenized results presented in this section fall inside such region. Actually, the curve for the mean pressure homogenization (MS - Mean) and for the improved scheme with $\mathsf{n}^{\mathsf{p}} = 3$ end up almost hiding the the Asymptotic BGT curve, due to the superposition of numerical and analytical results.

In addition to the contact area evolution curve, it also common to analyze its derivative, since typically a linear evolution of the contact area with the load is sought for, and the linearity coefficient is the key parameter for such characterization. In Chapter 4, it has been shown that the linearity coefficient, or in other words, the normalized derivative of the contact area curve, is $\sqrt{2\pi}$ for the Asymptotics BGT and $\sqrt{8/\pi}$ for Persson's model. These results are plotted in Figure 6.19b, together with the reference values and the numerical derivatives of the homogenized results, computed with centered finite differences. Since few realizations were considered for each scale, the numerical derivatives of the contact area curves appears with an irregular behavior. Therefore, for the sake of analyzing and improving the readability of the results, a cubic function with 3 degrees of freedom is fit to each of the homogenized curves—it is forced that the fitted curve is zero for zero pressure. The derivative of the fitting function is plotted in Figure 6.19b. Despite the fact that the reference numerical values (Yastrebov, Anciaux, *et al.*, 2015) of the derivatives are not quantitatively captured by the homogenized solutions, the overall qualitative tendency, i.e., decreasing derivative with increasing load, is well predicted.

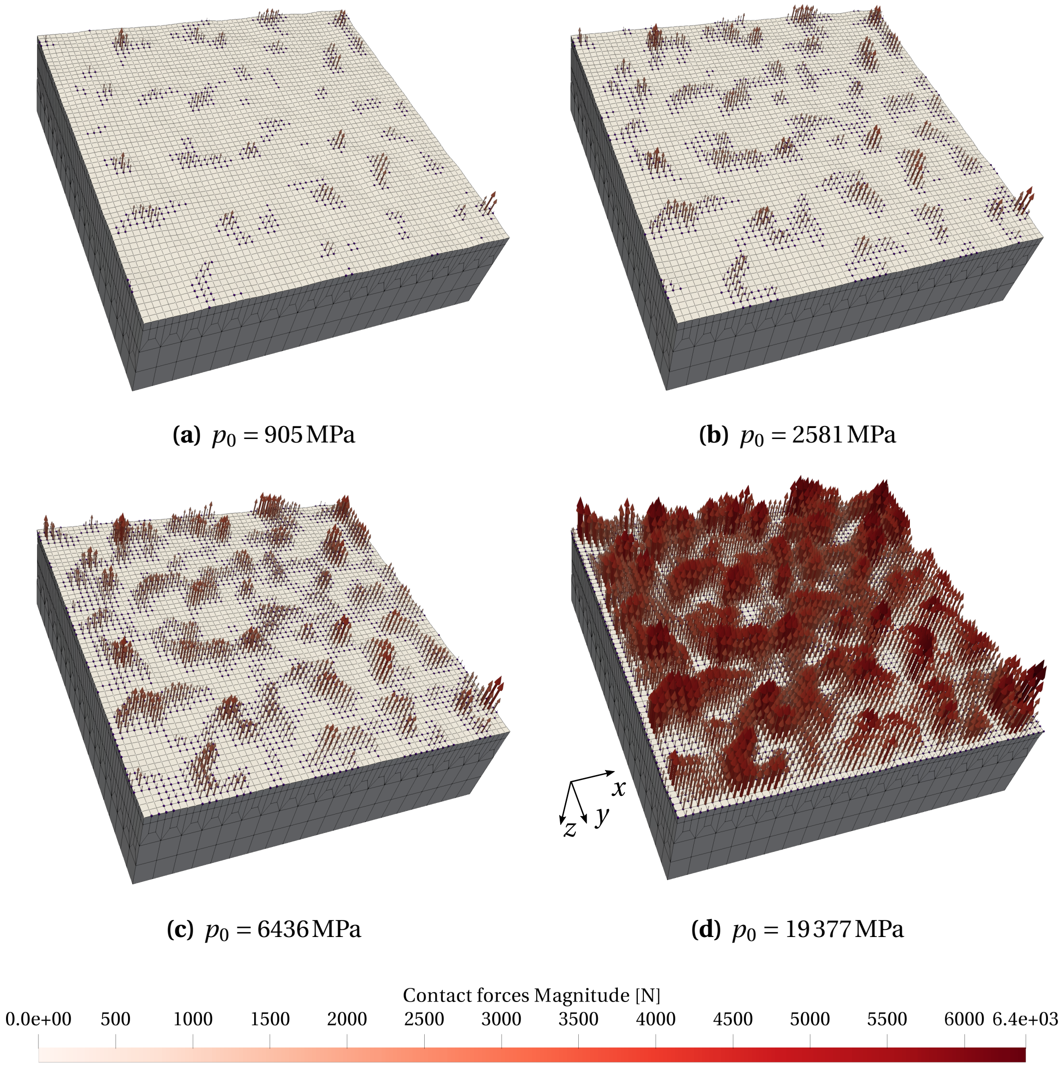
Before closing this chapter, it is interesting to analyze the distribution of contact forces and contact spots throughout the loading, in a 3D numerical model. Figures 6.20 and 6.21 provides a graphical representation for both evolutions, regarding the macroscale model. At light loads, contact starts at several disconnected small sized spots, which grow and coalesce with neighboring contact regions, with increasing external load. As the full contact condition is approached, the contact forces distribution resembles a scaled copy of the original rough topography.



(a) Contact evolution curve for light contact

(b) Normalized derivative of the contact area curve

Figure 6.19: Results for the real contact area and respective derivative, computed with the multiscale strategy, and comparison with analytical solutions and results from other authors.



(a) $p_0 = 905\,\text{MPa}$

(b) $p_0 = 2581\,\text{MPa}$

(c) $p_0 = 6436\,\text{MPa}$

(d) $p_0 = 19\,377\,\text{MPa}$

Figure 6.20: Distribution of the contact forces on the non-mortar interface of the macroscale topography considered for the 3D analysis. The dark dots denote active nodes.

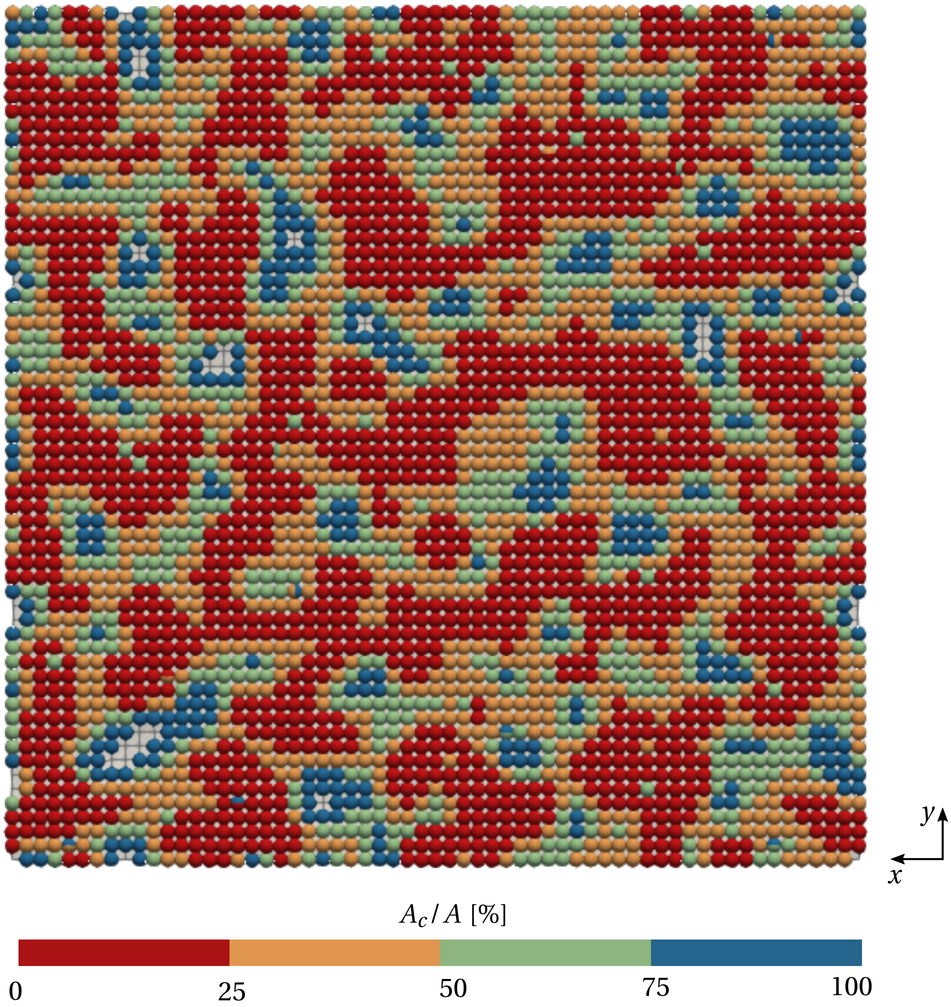


Figure 6.21: Active nodes and contact area fraction at different stages of contact.

Page intentionally left blank

Chapter 7

Concluding remarks and future work

In this dissertation, the elastic, non-adhesive and frictionless contact between self-affine rough topographies and a rigid and flat surface is modeled within a single scale and multiscale finite element method framework coupled with a dual mortar contact formulation. The multiscale approach, based on contact homogenization, aims at reducing the computational resources required to model rough contact across a wide range of scales, while preserving the accuracy of the solution relative to the single scale model. The majority of the numerical simulations have been carried out for Signorini problems in two dimensions, concerning the contact between an elastic rough profile and a rigid flat base. The multiscale strategy is then extended to the contact of numerical rough profiles with a remarkably wide roughness spectrum and also to three dimensional contact of rough surfaces.

The main conclusions of present dissertation are presented in the following, including an explicit reference to the original contributions introduced in this work. Finally, an outlook for future works is provided at the end of this chapter.

7.1 General conclusions

In the numerical modeling of rough contact, the ability to generate discrete rough topographies verifying well defined characteristics is of high practical interest. Thus a thorough review on roughness characterization techniques is presented, focusing on the spectral description via the Power Spectral Density, and its connection with self-affine roughness. Then, the numerical generation of randomly rough topography is addressed. Two FFT-based generation algorithms have been implemented from scratch, one for the generation of Gaussian topography (J.-J. Wu, 2000b) and other focused on non-Gaussian topography (J.-J. Wu, 2004). Both are capable of generating periodic profiles and surfaces from any given input Power Spectral Density or Autocorrelation Function. Additionally, the non-Gaussian generator can guarantee, in an approximate sense, prescribed values for the skewness and kurtosis on the synthesized topography. While the prescribed ACF is only approximately recovered at the output, with larger deviations for increasingly longer autocorrelation lengths, the prescribed PSD is preserved throughout the numerical procedure, for both algorithms. Furthermore, it has been verified that the heights distribution of the artificially generated topography, produced with the Gaussian generator, are normally distributed if the low cut-off wavelength is small, in comparison with the topography length (approximately for $L > 4\lambda_l$). Surfaces and profiles synthesized with the non-Gaussian algorithm can verify accurately the prescribed skewness and kurtosis. While the kurtosis does not influence considerably the convergence of the algorithm, skewness is verified to cause a major impact. In particular, by prescribing large values for the skewness (more than 2, in magnitude), the non-Gaussian algorithm struggles to converge. In truth, the effect of skewness and kurtosis is combined, since varying the required kurtosis helps some cases with large skewness to converge. Nonetheless, the influence of skewness is much more striking. These numerical methods have been applied to real cases, by using experimental measurements on the washer of a roller bearing and on the flank of a gear tooth, before and after power loss tests. The autocorrelation function, skewness and kurtosis of the output topographies match closely the input counterparts, and the spatial distribution of heights is verified to resemble accurately the original topography.

The framework for single scale modeling of rough contact is established by assembling the random topography generator, the mesh generator and the finite element code (`LINKS`), within a set of `Python` scripts. Single scale finite element analyses on 2D self-affine rough contact have been performed, with the purpose of establishing the conditions under which the micromechanical contact problem is representative—the so called Representative Contact Element (RCE). This issue has not been addressed frequently in the literature, in particular, in the context of finite element modeling. The RCE parameters which need to be characterized for some topography (described by the long and short cut-off wavelengths, Hurst roughness exponent and RMS slope) are the mesh size, rough block length and height. From the mesh convergence test, it has been concluded that, in order to obtain a representative solution for the contact area-pressure curve, it must be guaranteed that a minimum of 4 nodes exist in each asperity. This is equivalent to stating that the mesh step must be 8 times smaller than the shortest wavelength contained within the profile. With regard to the RCE length, the numerical results suggest that it shall be at least 8 times the longest wavelength of the profile, such that the mechanical response is representative. The RCE height does not impact the contact area results, as long as it is not excessively small. However, in order to resolve properly the stress field within a thin region near the rough boundary, the height must be kept at least at 160 times the RMS height, and the height of the high resolution mesh shall measure at least 40 times the RMS height. The number of different topography realizations used to determine the average response of the micromechanical problem is also paramount. By comparing the standard deviation of the mean contact area response for different samples with the same number of realizations, it has been concluded that at least 10 different RCE's are required to reduce standard deviation of the contact area curve down to some acceptable tolerance and, thus, assure a representative mechanical response.

The single scale simulations also confirmed how rapidly the computational resources required to solve the rough contact problem grow with increasing spectrum bandwidth. Therefore, a multiscale approach for the prediction of the real contact area, based on the contact homogenization strategy proposed in Wagner (2018), has been implemented in the numerical framework. It consists in dividing the topography into several scales, by introducing several splitting frequencies in the PSD. This multiscale strategy is classified as an information-passing type, since the simulation of all roughness scales are uncoupled, and the multiscale solution is computed via an homogenization procedure, in a post-processing operation. Several fundamental aspects of the multiscale approach are discussed for the first time in this dissertation, namely, a systematic rule for the definition of the splitting frequencies, based on the uniform distribution of computational load across scales. Additionally, the numerical scheme for the evaluation of the real contact area and the consistency of topography generation at the microscale are addressed, as well. By applying the multiscale procedure to roughness profiles with spectra of various widths (in a two-scales setup) and inspecting the quality of the results and their computational cost, it is verified that the homogenized solution matches the DNS solution closely at light contact, for ratios $\lambda_l/\lambda_s \geq 32$. For larger contact area fractions, the DNS and the multiscale solution do not give the same area response. Nonetheless, the differences in resources usage starts to become very pronounced, specially for wide spectra. For example, for $\lambda_l/\lambda_s = 64$, the single scale results require about 80 GB of RAM, and take approximately one day to finish, while the multiscale approach with 2 scales completes within nearly 50 minutes, and requires less than 2 GB of RAM.

In order to improve the quality of the contact area predictions of the multiscale approach, a new contact homogenization procedure is proposed. The basic idea consists in incorporating several discrete pressure values, extracted from the pressure distribution at each scale, in the scale transition. Then, the multiscale solution is computed via a weighted average multiplicative homogenization step, so that the averaging procedure is based on the multiplication of the real contact area fraction at some scale, with a weighted average of the contact areas of the smaller scales, computed at the discrete contact pressure at the larger scale. This introduces a new adjustable parameter, being the number of pressure values discretized from the contact pressure distribution. Again, for large bandwidths, this strategy can reproduce almost exactly the DNS solution for small loads, even if only 3 discrete contact pressure values are considered for the homogenization step. Additionally, the homogenized results are observed to converge for a solution close to the DNS results, at moderately high loads, with increasing number of discrete pressures. The initial and improved homogenization strategies have been applied to rough profiles with considerably wide spectra, $\lambda_l/\lambda_s = 4096$, which, to the author's knowledge, have not been analyzed in the literature. For such scenario, numerical tests have been carried with different number of scales. The results computed with the improved homogenization technique are practically independent of the number of scales considered, as long as the spectrum width at each scale is sufficiently wide. The initial multiscale approach, however, shows quite sensitive results to the number of scales considered. By employing the improved multiscale strategy, a qualitative study on the influence of the roughness power spectrum bandwidth in the real contact area was performed. It was verified that every roughness scale contributes significantly for the real contact area, in particular, by increasing the roughness bandwidth, the real contact area decreases in a monotonic fashion, for the same applied load, and no signs of saturation in the response have been observed. Nevertheless, the non-dimensionalized results are mostly insensitive to the width of the power spectrum.

Finally, the multiscale approach is applied to a 3D problem, which is, in fact, the ultimate goal of such strategies. The homogenized contact area results, for light contact, lie very close to numerical results obtained by other authors, and also within the acceptable range, relative to analytical micromechanical contact models.

7.2 Original contributions

In the course of this dissertation, two novel computational aspects were introduced in the rough contact modeling framework, namely, a modification of the original non-Gaussian generation algorithm by J.-J. Wu (2004), and an improved contact homogenization scheme. While the modification of the non-Gaussian generation ended up having virtually no influence in the global context of this thesis, and is mentioned here only for future reference, the proposed contact homogenization strategy is one of the milestones of the present work.

In the original non-Gaussian generation algorithm, the author refers that the optimization procedure incorporated in the global algorithm can be solved by applying *"some numerical method, such as the bisections method"*. However, by numerical experience, it has been observed that the convergence of such iterative procedure is highly dependent on the initial guess. In order to improve the robustness of the generation algorithm, the iterative procedure is preceded by a **brute force trial-and-error strategy**, in order to select the best initial guess for the optimization procedure. With this modified generation algorithm, the overall precision of the method in the output skewness and kurtosis is significantly improved.

The multiscale approach for contact area prediction proposed by Wagner (2018), based on a multiplicative homogenization step, has been verified to provide an accurate solution for light contact, as long as the spectrum width is sufficiently large. However the accuracy for moderately to high loads deteriorates, even for such spectra. In order to enhance the multiscale solution, a new scale transition step consisting in a **weighted average multiplicative homogenization step** is proposed. This procedure can predict accurate solutions for both low and high external pressure, by adjusting the number of discrete points in the pressure distribution. Additionally, it is also verified to be insensitive to the number of scales, for wide bandwidths.

7.3 Future work

Throughout this work, some aspects were referred to be amenable for treatment in future developments. In fact, multiscale modeling of rough contact by means of the finite element method is a relatively recent research field, and there is a wide margin for progression. In addition, several aspects, as those mentioned in the main chapters of this document, require further clarification. Future developments on the topic may address the following points.

- **Extension to frictional contact.** In truth, the multiscale approach to frictionless contact is an essential development towards the formulation of a multiscale framework to model frictional contact, as in the original work of Wagner (2018). Therefore, the extension of the multiscale strategy developed in this dissertation to model friction is certainly of primary interest, due to the practical and multidisciplinary importance of friction.
- **Implementation of an improved topography generation algorithm.** Despite that only Gaussian topography has been considered for the contact analysis, the numerical tests on the non-Gaussian generator suggest that the range of skewness and kurtosis on which the algorithm's accuracy is assured is quite limited. The hybrid method proposed by Francisco and Brunetière (2016) seems to be less restrictive, and shall be considered if the limitations of the currently implemented algorithm become unacceptable.
- **Assessment of the microscale topography generation.** Within the multiscale framework employed in this contribution, the microscale topography was generated directly with the required length. As has been remarked, by doing so, the discrete spectrum of the microscale topography will contain less points than the full scale analogous. Since the discrete frequencies of the full scale spectrum are not completely represented at the lower scales, the microscale topography may verify different properties in comparison with respective component at the full scale topography. This issue shall be studied thoroughly, in order to determine the minimum frequency representativeness required at the microscale.
- **Incorporation of the global body shape.** The present work concerns only micromechanical contact. Naturally, these micromechanical models are part of the interface of a real macroscale contact situation, for example, regarding the contact between a sphere and a flat base. While only micromechanical variables have been analyzed, the incorporation of the real body shape as the largest scale in the multiscale analysis is paramount for predicting interface properties in real applications.
- **Definition of a 3D Representative Contact Element.** The representativeness assessment for the micromechanical problem concerned only 2D rough contact. Naturally, these investigations shall also be validated more extensively for 3D cases.
- **Validation of the numerical contact area evaluation.** It has been remarked that the real contact area, within the dual mortar finite element method, can be computed either following a geometrical argument, or via the Lagrange multipliers. In the course of the present thesis, one of the two formulations has been adopted, yet knowing beforehand that they predict different contact area ratios for the same load. This issue shall be substantiated by experimental evidence, which will, most likely, favor one of the alternatives as the physically more correct scheme.
- **Modeling of more complex cases in frictionless contact.** The multiscale approach can be extended to model more complex situations, such as anisotropic roughness and the contact of non-Gaussian surfaces. The latter topic is very appealing, since there is no clear understanding on how classical non-Gaussian parameters, such as skewness and kurtosis, change within different scales. Additionally, non-linear material laws, namely, elasto-plasticity, are also prone to investigation.

Page intentionally left blank

Appendix A

Notes on Fourier transforms

While the spatial characterization of random rough surfaces is useful in identifying periodicity of random surfaces and evaluating the correlation between its points, the spectral characterization in the frequency domain provides powerful ways to describe the surfaces, regarding the importance of each frequency in the surface shape. Furthermore, the fractal behavior of rough surfaces is closely related to their spectral properties, since increasing the magnification of a fractal leads to the revelation of higher frequencies—the topography is visualized with more details.

The classical approach the the computation of surfaces' spectral properties is based on concepts of Fourier Analysis. In particular, the *Discrete Fourier Transform* (DFT), mostly known due to the efficient *Fast Fourier Transform* (FFT) algorithms, is extremely relevant in the numerical evaluation of Fourier Transforms, since it describes both frequency and time (space) at discrete points and frequencies.

In the following sections, some basic features of Fourier Analysis are presented, as a basis for their usage in rough surface description and on the numerical generator of random rough surfaces. Firstly, the Fourier series and the continuous Fourier Transform are presented, since they establish the fundamental concepts of Fourier Analysis (even though they are not explicitly used through the text). Following the basic concepts stated through the continuous Fourier Transform, its discrete versions, namely, the *Discrete-Time Fourier Transform* (DTFT) and the Discrete Fourier Transform are revisited, with special focus on the DFT, the only one relevant for computer implementation. The next paragraphs provide a compact but by no means exhaustive introduction to the principal concepts of Fourier transforms. Nevertheless, for a comprehensive treatment on the topic, the interested reader is referred to Kreyszig (2010), Chaparro (2010), Newland (1984), Orfanidis (1996), and Rao *et al.* (2011).

A.1 Fourier series

A function $f(x)$ is called periodic of period λ (wavelength), if λ is the smallest value that verifies

$$f(x+n\lambda) = f(x), \quad \forall\, n \in \mathbb{Z}\,, \tag{A.1}$$

where $\mathbb{Z}$ denote the set of integers. Under these conditions, λ is also commonly termed as the fundamental period of the function. The Fourier analysis is based on the fundamental fact that it is possible to represent a periodic function of period λ as the superposition of a infinite number of sinusoidal functions, whose frequency is a multiple of the fundamental frequency k_f. Note that this frequency is defined as the angular frequency (expressed in $\mathrm{rad\,s^{-1}}$) in the time domain and as the wavenumber in the spatial domain (expressed in $\mathrm{rad\,m^{-1}}$). Mathematically, this can be written as

$$f(x) = a_0 + \sum_{n=1}^{\infty} \left[a_n \cos\left(nk_f x\right) + b_n \left(\sin nk_f x\right) \right], \quad \text{where } k_f = \frac{2\pi}{\lambda}\,. \tag{A.2}$$

The coefficients a_0, a_n and b_n are called the Fourier coefficients, and are given by the following relations:

$$a_0 = \frac{1}{\lambda} \int_{-\frac{\lambda}{2}}^{\frac{\lambda}{2}} f(x)\,\mathrm{d}x\,; \tag{A.3a}$$

$$a_n = \frac{2}{\lambda} \int_{-\frac{\lambda}{2}}^{\frac{\lambda}{2}} f(x) \cos\left(nk_f x\right)\,\mathrm{d}x\,; \tag{A.3b}$$

$$b_n = \frac{2}{\lambda} \int_{-\frac{\lambda}{2}}^{\frac{\lambda}{2}} f(x) \sin\left(nk_f x\right)\,\mathrm{d}x\,. \tag{A.3c}$$

As a brief example, we shall consider a sawtooth-wave function of period 2π,

$$f(x) = x + \pi \quad \text{for} - \pi < x \leq \pi, \quad \text{with } f(x + 2\pi) = f(x)\,. \tag{A.4}$$

The Fourier coefficients of this function are:

$$\begin{aligned} a_0 &= \pi\,; \\ a_n &= 0; \\ b_n &= -\frac{2}{n}\cos(n\pi)\,. \end{aligned} \tag{A.5}$$

A graphical representation is provided in Figure A.1, showing the partial sums of the Fourier series S_N, i.e., the truncated summation until $n = N$. It can be observed that the inclusion more harmonics in the summation leads to a more precise result.

Alternatively, the Fourier series in Equation (A.2) can be rewritten using complex exponentials. Denoting by $\mathrm{i} = \sqrt{-1}$ the imaginary number, the complex Fourier series can be stated as

$$f(x) = \sum_{n=-\infty}^{\infty} c_n e^{\mathrm{i}nk_f x}\,. \tag{A.6}$$

The coefficients c_n are to be obtained in a similar fashion as a_n and b_n, namely

$$c_n = \frac{1}{\lambda} \int_{-\frac{\lambda}{2}}^{\frac{\lambda}{2}} f(x) e^{-\mathrm{i}nk_f x}\,\mathrm{d}x\,. \tag{A.7}$$

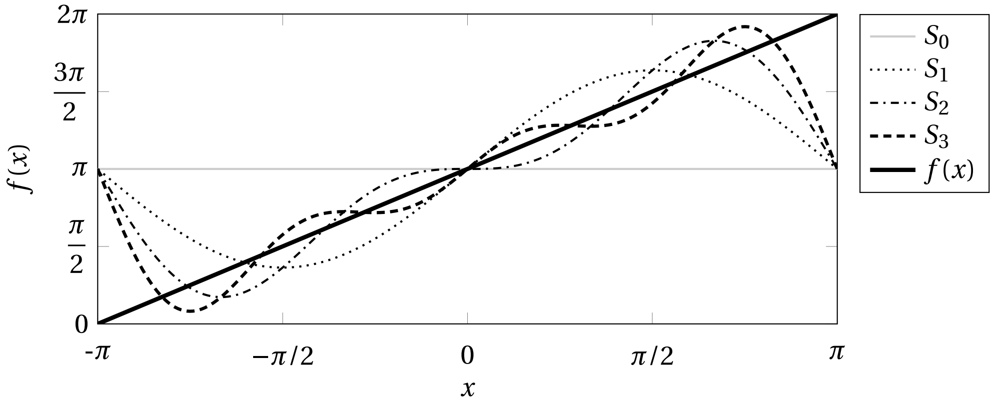


Figure A.1: Partial sums of the Fourier series of a sawtooth-wave.

It should be highlighted that when the Fourier series are written in a complex form rather than with trigonometric functions, one needs to sum both positive and negative harmonics (waves with negative frequency). This comes as a consequence of writing a real valued function using complex numbers. The contribution of a *real* harmonic of frequency nk_0 to the function $f(x)$, is thought as the contribution of a *real* sinusoidal. Thus, using the mathematical abstraction of complex numbers to represent the real variable, the real sinusoidal wave is to be related to the complex number. Using Euler's formula, it follows that

$$e^{\mathrm{i}nk_f x} = \cos\left(nk_f x\right) + \mathrm{i}\sin\left(nk_f x\right) \ , \tag{A.8a}$$

and

$$e^{-\mathrm{i}nk_f x} = \cos\left(-nk_f\right) + \mathrm{i}\sin\left(-nk_f x\right) = \cos\left(nk_f x\right) - \mathrm{i}\sin\left(nk_f x\right) \ . \tag{A.8b}$$

By taking the difference of these relations, it results

$$\sin\left(nk_f x\right) = \frac{1}{2}\left(e^{\mathrm{i}nk_f x} - e^{-\mathrm{i}nk_f x}\right) \ . \tag{A.9}$$

This means that in order to know the contribution of a real valued sinusoidal wave with a specific frequency, one needs to compute the contribution of two complex exponentials with symmetric frequencies. In fact, one shall expect that the contribution of frequencies nk_f and $-nk_f$ are the same. Additionally, for a real valued function, one can write that $f(x) = f^*(x)$, where the superscript $(\cdot)^*$ stands for the complex conjugate, which means that

$$c_n = \frac{1}{\lambda}\int_{-\frac{\lambda}{2}}^{\frac{\lambda}{2}} f(x)e^{-\mathrm{i}nk_f x}\,\mathrm{d}x \,, \tag{A.10a}$$

and

$$c_n^* = \frac{1}{\lambda}\int_{-\frac{\lambda}{2}}^{\frac{\lambda}{2}} f(x)e^{\mathrm{i}nk_f x}\,\mathrm{d}x \,. \tag{A.10b}$$

Inspecting Equations A.10 in more detail, it can be concluded that

$$c_n = c_{-n}^* \,. \tag{A.11}$$

This property is valid not only for Fourier series, but for Fourier transforms as well, when applied to real valued functions—this is called the *conjugate symmetry property*, an important concept when dealing with discrete functions. In fact, it can be readily interpreted: since for real valued functions the sum of imaginary parts must be zero, for symmetric frequencies the real part of each pair of symmetric frequencies (involving the cosine) will add up, and the imaginary part (involving the sine) will cancel out.

In sum, when using complex exponentials to express the contribution of a real valued sinusoidal wave, one considers the contribution of two complex exponentials of symmetric frequencies. However, due to the conjugate symmetry property, all the information is contained in a single complex exponential.

A.2 Fourier transform

While Fourier series are typically used when dealing with periodic functions only, they can still be applied to non-periodic signals in order to investigate its frequency content. Such analysis can be carried out by still assuming a periodic function of period λ, such that its Fourier series can be computed by setting the period tend to infinity and, thus, setting the function as non-periodic. In doing so, the summation in Equation (A.6) will yield an integration, because as the smaller the fundamental frequency is, the closer the harmonics will be between each other. The Fourier Transform of an integrable function $f(x)$, this is, whose integral in all x-domain is finite, can be stated as

$$\mathcal{F}\left\{f(x)\right\} = F(k) = \int_{-\infty}^{\infty} f(x) e^{-\mathrm{i}kx}\,\mathrm{d}x\,. \tag{A.12}$$

Its Inverse Fourier Transform follows[1]

$$\mathcal{F}^{-1}\left\{F(k)\right\} = f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(k) e^{\mathrm{i}kx}\,\mathrm{d}k\,. \tag{A.13}$$

The Fourier Transform of $f(x)$, since it is expressed as an integral over the frequency domain, can be interpreted as the *intensity density* of sinusoidal waves of frequency k contained in the function. This is, a specific frequency frequency k has zero contribution to the function, since $f(x)$ results from the superposition of infinite frequencies around k. To get the contribution of frequencies around k, one needs to integrate over k, such that $F(k)$ acts like a measure of intensity *per* frequency interval. The role of $F(k)$ is basically the same as c_n from the Fourier series: it is a complex number (in general) that contains the contribution, both in magnitude and phase, of sinusoidal waves of frequency k in the function $f(x)$—its frequency content. The conjugate symmetry property can be applied to the Fourier Transform as well. In doing so, one can conclude that the magnitude of the Fourier Transform is symmetric along the frequency domain $|F(k)| = |F(-k)|$, if the function is real valued..

[1] The presence of 2π in the denominator of Equation (A.13) is not consistent in the literature. For example, Kreyszig (2010) uses $\sqrt{2\pi}$ in the denominator of both the Fourier Transform and on the Inverse Fourier Transform. Here the definitions in Equations (A.12) and (A.13) is employed.

A.2.1 Convolution theorem

The Fourier Transform encompasses a very useful property, stated by the convolution theorem. This property is widely used in its DFT (or FFT) version, to compute the convolution of two signals very efficiently. The linear convolution of two continuous functions f and g is defined as

$$\left[f * g\right](\tau) = \int_{-\infty}^{\infty} f(x)g(\tau - x)\,\mathrm{d}x\,. \tag{A.14}$$

Inspecting Equation (A.14), a linear convolution operation is performed by taking the symmetry along the x-axis of one of the functions, and then sliding it along the x-axis, such that it is displaced relative to the other. These two functions are then multiplied, and the result is integrated over all domain. The convolution theorem states that

$$\mathcal{F}\left\{f * g\right\} = \mathcal{F}\left\{f\right\}\mathcal{F}\left\{g\right\}\,. \tag{A.15}$$

Thus, it is possible to compute the linear convolution of two functions via the product of their Fourier Transforms, making use of Equation (A.15) and taking its inverse Fourier Transform.

A.3 Discrete-Time Fourier Transform

In real applications, the possibility to handle with continuous functions or signals is very uncommon. Instead, one typically has to work with a finite number of points, either sampled from a real measurement or discretized using a computer. While the tools presented previously are suited to work with continuous functions/signals, they can still be adopted to deal with discrete-time or discrete-point signals.

First of all, a sampled signal needs to be extracted from a continuous one. If the sampling of the signal is done with a period l_s at points $x = nl_s$, to get a *finite-energy* signal, i.e. to keep the signal integrable and with non-zero integral, the sampling can be described by the application of several impulse functions at every sampling points. Let $\delta(x - nl_s)$ be the Dirac Delta function applied at point $x = nl_s$. Assuming the summation is done for n between $-\infty$ and ∞, the sampling function is

$$\delta_s = \sum_n \delta(x - nl_s)\,. \tag{A.16}$$

Multiplying the sampling function by the original continuous signal, the sampled signal can be generated by

$$f_s(x) = \sum_n f(nl_s)\delta(t - nl_s)\,. \tag{A.17}$$

At this stage, the computation of the Fourier transform of the sampled signal, using the linearity properties of the Fourier Transform, and the transform of the Dirac Delta function, yields the result

$$F_s(k) = \sum_n f(nl_s)e^{-\mathrm{i}kl_s n}\,. \tag{A.18}$$

Notice that the function f is now written with square brackets $f\,[\cdot]$, denoting a discrete function. Finally, defining the discrete frequency $\omega = l_s k = 2\pi k/\Omega_s$, and suppressing the sample period l_s from the argument of the function, we obtain the Discrete-Time Fourier Transform of a discrete signal $f[n]$

$$F_s(\omega) = \sum_n f[n]e^{-\mathrm{i}\omega n}, \quad \text{for } -\pi < \omega < \pi \,. \tag{A.19}$$

The respective inverse transform comes

$$f[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} F(\omega)e^{\mathrm{i}\omega n}\,\mathrm{d}\omega \,. \tag{A.20}$$

Note that, even though the signal is a discrete in the x-domain, its DTFT of is continuous and periodic of period 2π in the discrete frequency ω. In other words, recovering the definition of discrete frequency ω, the period of the DTFT is equal to the sampling period l_s in the physical/real frequency. The periodicity can be readily demonstrated as

$$\begin{aligned} F_s(\omega + 2\pi) &= \sum_n f[n]e^{-\mathrm{i}(\omega+2\pi)n} \\ &= \sum_n f[n]e^{-\mathrm{i}\omega n}e^{-\mathrm{i}2\pi n} \\ &= \sum_n f[n]e^{-\mathrm{i}\omega n} \\ &= F_s(\omega) \,. \end{aligned} \tag{A.21}$$

Finally, it is important to mention that Equation (A.19) defines the DTFT of $f[n]$ for every frequency ω. Since it is periodic (with period 2π), one only needs to define $F(\omega)$ within a period, which is taken to be between $-\pi$ and π. It is noteworthy to refer that in all the definitions above, $f(x)$ is assumed a continuous function defined for every x, and that the discrete signal $f[n]$ is sampled from $f(x)$ at a infinite number of points through all the domain.

A.3.1 Nyquist-Shannon sampling theorem

As already discussed, the DTFT of a discrete signal $f[n]$ is a continuous and periodic function of the discrete frequency ω. Its period is equal to 2π, which is equivalent to stating that it is a continuous and periodic function of the spatial frequency k with period equal to T_s. Alternatively, the DTFT can be rewritten as

$$F_s(k) = \frac{1}{l_s}\sum_{r=-\infty}^{\infty} F(k - r\Omega_s), \quad \text{with } \Omega_s = \frac{2\pi}{l_s} \,. \tag{A.22}$$

This means that the Fourier Transform of the sampled signal $f[n]$ can equally be written as the superposition of the Fourier Transform of the real function $f(x)$, shifted in the frequency domain by a multiple of the sampling frequency, and scaled by a constant equal to the inverse of the sampling period. Additionally, the periodicity of the DTFT can be verified.

In order to analyze the consequences regarding the frequency representation of taking a infinite sampled signal as a representation of a continuous one, consider a continuous real function $f(x)$ whose Fourier Transform is limited by a frequency k_{max}. Due to the conjugate symmetry property, it follows that $F(k)$ is also limited by $-k_{\text{max}}$ (existing symmetry along the frequency axis). Figure A.3a shows the Fourier Transform of the continuous signal. In Figures A.3b and A.3c, it is presented the contribution for the DTFT of each shifted version of the transform, which matches the DTFT in Figure A.3b. The DTFT of the sampled signal corresponding to the contributions in Figure A.3c is shown in Figure A.3d. On one hand, if the continuous signal is sampled with a sufficiently high frequency, the DTFT will be composed of several scaled Fourier Transforms of the continuous signal that do not overlap. On the other hand, if signal is sampled with a low sampling frequency, there will occur overlapping between the different contributions. Real world signals, or ones generated in the computer, are sampled signals, not continuous ones. Consequently, the frequency content of those signals are related to their DTFT. By not choosing a sampling frequency high enough, the frequency content of those discrete signals will not match the continuous signal frequency content, due to the overlapping, as shown in Figure A.3c. This leads to the distortion of the frequency content, cf. Figure A.3d, where the contributions of two overlapping frequencies cannot be distinguished, as only their sum is known, and, therefore, the continuous original signal cannot be recovered. In order to avoid the overlapping and to keep the original frequency content intact after sampling, the sampling frequency must verify the following relation

$$\Omega_s \geq 2k_{\text{max}}\,. \tag{A.23}$$

This is know as the Nyquist-Shannon sampling theorem, and k_{max} is the maximum frequency in the continuous signal (if the signal is band-limited). One can also think the other way around, saying that the maximum frequency that can be resolved by sampling with frequency Ω_s—(termed as the *Nyquist frequency*)—is

$$k_{\text{nyq}} = \frac{\Omega_s}{2}. \tag{A.24}$$

In the case of overlapping, this is, when Equation (A.23) is not verified, it lead to the phenomenon of *frequency aliasing*. Aliasing, as the name suggests, happens when two different frequencies, in a continuous signal, look similar in the discrete one. They acquire an alias, and its not possible to distinguish between them in the DTFT of the discrete signal. Figure A.2 provides a geometrical interpretation of the aliasing effect. Two continuous sinusoidal waves are plotted, one with a frequency k_1, smaller than half of the sampling frequency, and other with a higher frequency, equal to the sum of k_1 with the sampling frequency. Simultaneously, discrete points taken from each one of the waves at points $x = nT_s$ are plotted. It can readily be seen that the discrete points match in both curves. The DTFT sees these discrete points, not the continuous curves, so, the two frequencies will look alike, making it impossible to distinguish between them. If the high frequency is not present in the real signal, there is no problem, since its contribution is null. Even if the DTFT cannot distinguish the two frequencies, the overlap of both contribution will result in the original contribution of the lower frequency. Yet, if the high frequency is present in the real signal, the DTFT cannot distinguish between the contributions of the two frequencies, because they look the same, distorting the real DTFT, like in Figures A.3c and A.3d. Figure A.2 also leads to a geometrical interpretation of the

Nyquist-Shannon theorem: the maximum frequency that can be detected by the DTFT is equal to half of the sampling frequency. This is the case when just two points are sampled per oscillation, one in the top half-wave and other in the bottom half-wave, representing the minimum condition for detecting a frequency.

Finally, this discussion suggests a reason to define, in Equation (A.19), the DTFT between $-\pi$ and π, and not between any other frequency interval. Since it is periodic, one could write the function in any interval as long as its width would be equal to the period of the function. The discrete frequencies between $-\pi$ and π are minimum frequencies for which the discrete representation directly suggests the frequency of the signals. For higher frequencies, the discrete representation of the wave would graphically leads to a wrong perception of the true frequency of the signal, due to frequency aliasing.

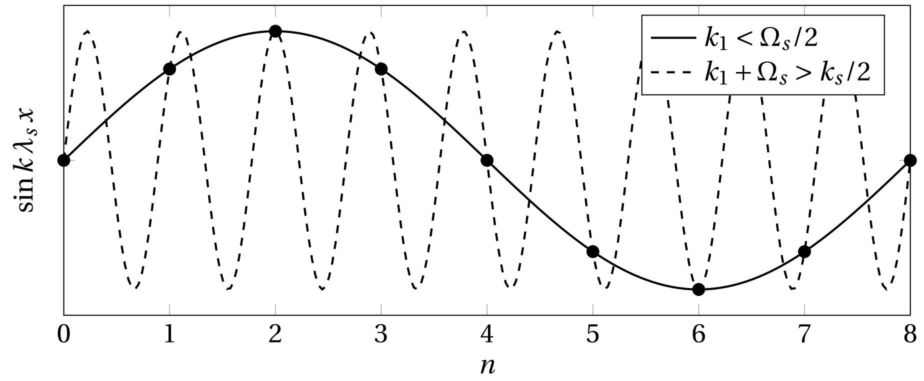
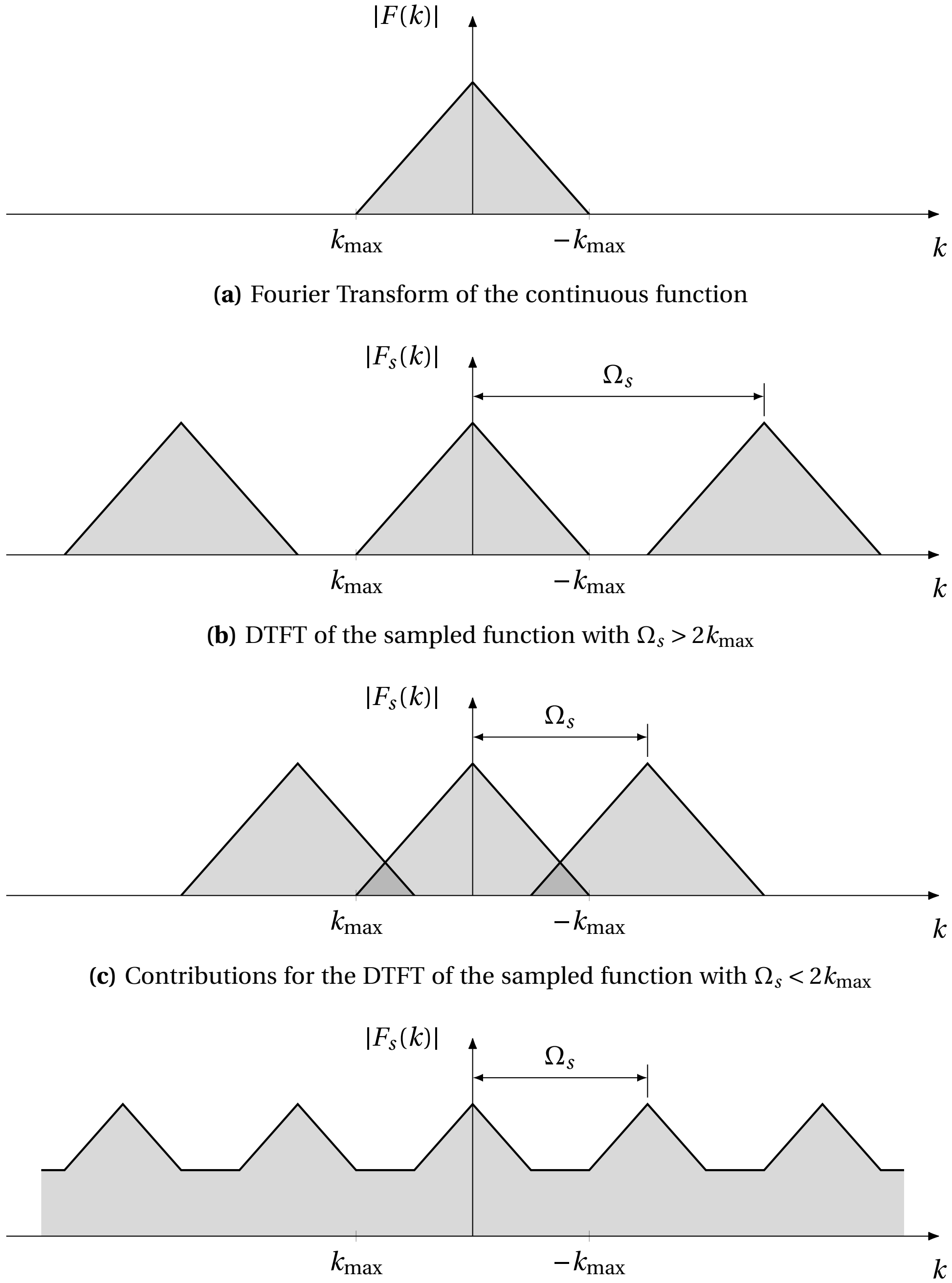


Figure A.2: Frequency aliasing effect. The sampled signal, plotted as circular dots, fit in both both frequencies, then it is impossible to distinguish between these frequencies from the sampled signal.

A.4 Discrete Fourier transform

So far, focus has been placed on signals theoretically infinite, whose frequency content, or *spectrum*, is continuous. Signals verifying these conditions cannot be analyzed numerically, since they require a continuous integral for the IDTFT and a infinite signal. Therefore, this hypothesis is relaxed and, even though it may lead to a less accurate representation of the spectrum, it enables the application of various numerical procedures.

The infinite discrete signal hypothesis is removed by truncating the signal at a finite number of points. Naturally, if the truncated signal is non-zero, there is some amount of lost information and the DTFT of the infinite signal will not be the same as the finite one. Truncating a signal is done by multiplying the signal by a window function (also called windowing the signal). The most straightforward strategy being the multiplication by a window function, which is unitary in the points one wants to keep and zero everywhere else. This induces a change in the DTFT of the original signal, which can be given by the convolution of the DTFT's of both the original signal and the window function. Frequently, windowing a signal leads to *frequency leakage* and, thus, high frequency components that



(a) Fourier Transform of the continuous function

(b) DTFT of the sampled function with $\Omega_s > 2k_{\max}$

(c) Contributions for the DTFT of the sampled function with $\Omega_s < 2k_{\max}$

(d) DTFT of the sampled function with $\Omega_s < 2k_{\max}$

Figure A.3: Fourier Transform (magnitude) of the continuous signal and discrete signal for 2 values of the sampling frequency. When the Fourier transform of the continuous functions is bounded by a maximum frequency $k_{\max}$, if the sampling frequency is greater then $2k_{\max}$, it remains unchanged after sampling, as shown in Figure A.3b. On the other hand, if the sampling frequency is less than $2k_{\max}$, theres is overlapping between the original spectrum and the shifted versions (Figure A.3c) which results in distortion of the real spectrum, depicted in Figure A.3d

would not exist in the original DTFT will show up due the discontinuity of the windowed signal on the limits of the window function (Orfanidis, 1996).

Regarding the treatment of the continuous integral in the computation of the IDTFT, there is a need to convert this integral to a discrete one (or the DTFT to a discrete-frequency version). The Discrete Fourier Transform (DFT) is obtained from the DTFT by sampling this function at N equally-spaced, discrete frequencies

$$F_s[q] = \sum_{n=-\infty}^{\infty} f[n]e^{-\mathrm{i}2\pi qn/N}, \qquad q = 0,1,...,N-1\,. \tag{A.25}$$

Note that

$$F_s\left(\omega = \frac{2\pi}{N}q\right) = F_s\left(k = \frac{q}{N}\Omega_s\right) = F_s[q], \qquad q = 0,1,...,N-1\,. \tag{A.26}$$

Here, the subscript s is kept as a reminder that the sampling is done in the DTFT, which is the Fourier transform of the sampled signal, and might not be the same as the continuous signal, based on what it was discussed earlier. The sampled spectrum can be rewritten as follows (Proakis and Manolakis, 2007)

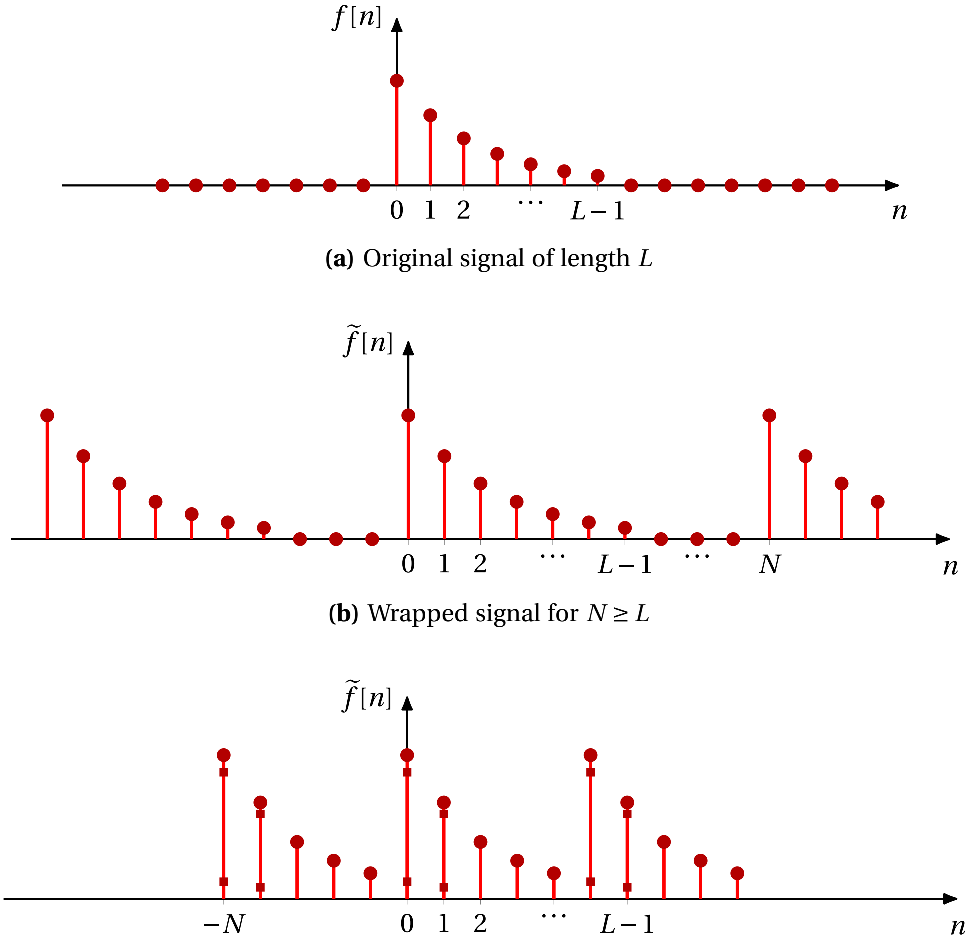
$$F_s[q] = \sum_{n=0}^{N-1}\left[\sum_{l=-\infty}^{\infty} f[n-lN]\right]e^{-\mathrm{i}2\pi qn/N} = \sum_{n=0}^{N-1}\widetilde{f}[n]e^{-\mathrm{i}2\pi qn/N}, \quad q = 0,1,...,N-1\,. \tag{A.27}$$

The new signal $\widetilde{f}[n]$ is periodic, with period N, and can be tough as a periodic, wrapped version of the original signal $f[n]$. It is called wrapped version of the original signal, because it is built by shifting the original signal by multiples of N points and summing all together, resulting in a wrapped, compact signal. On the other hand, the wrapped signal can be rebuilt as (Proakis and Manolakis, 2007)

$$\widetilde{f}[n] = \frac{1}{N}\sum_{q=0}^{N-1} F_s[q]e^{\mathrm{i}2\pi qn/N}, \qquad n = 0,1,...,N-1\,. \tag{A.28}$$

Equations (A.25) and (A.28) hold two very convenient results. First, it is possible to reconstruct the wrapped version with length N of a signal $f[n]$, using N points from the its spectrum. Conversely, using a finite N point signal, it is possible to get N equally-spaced points from the unwrapped signal's spectrum over a whole period. This correspondence from a N point signal to a N point spectrum, enables the use of this technique in a digital computer. Nevertheless, this procedure deals with the wrapped version of the discrete signal, which is different from the real signal. To work this out, the only variable parameter is the number of sampling points of the spectrum, N.

Figure A.4 illustrates the difference between the original and wrapped discrete signal for two different situations: one where the number of sampling points of the signal L is greater than the number of sampling points of the spectrum N, and other where the length is less than the number of sampling points. The periodic behavior of $\widetilde{f}[n]$ can be observed very clearly, and it resembles the periodicity of the DTFT, in Figure A.3, where the Fourier transform was shifted at multiples of Ω_s. Now the shift is done in x-domain instead, and on the original signal. Considering that the signal is of finite length L (like in Figure A.4a) N is chosen to be greater than L, the size of signal. The wrapped signal



(a) Original signal of length L

(b) Wrapped signal for $N \geq L$

(c) Wrapped signal for $N < L$

Figure A.4: Wrapped signals with different lengths, built by extending the original signal periodically, and summing all the shifted versions of the original signal within each period. It is identical to the DTFT in Figure A.3, performed on time domain. When N is greater than the length of the signal, its wrapped version, in Figure A.4b, contains the original signal padded with zeros. If N is less than L, it happens time aliasing, and the wrapped signal, plotted as circular dots in Figure A.4c, is obtained as summing the shifted versions of the original signal (small square dots).

contains the original within a period N, together with $N - L$ padded zeros, making it possible to recover the original signal. However, when N is chosen lesser than L, several shifted versions start overlapping and adding to each other. This phenomenon is similar to what happens in the frequency domain of the DTFT and is commonly termed as *time aliasing*.[2] The contribution of the original signal from different x-points add up, making the wrapped signal look different (distorted) when compared to the original. Since the difference between those x-points cannot be identified, it is impossible to recover the

[2] These concepts are typically used in signal processing, where x stands for the time variable.

original signal from this situation. Note that time aliasing always happens to infinite signals, since N is a finite integer (all this discussion is about making a discrete and finite representation of a spectrum). This is not such a problem, since in real applications signals are always finite.

Based on the previous argument, the meaning of Equations (A.25) and (A.28) can be restated, replacing the wrapped signal with the original signal. Let $f[n]$ be a finite length signal of length L. The DFT of the signal is defined as a sampling of its continuous spectrum, this is, its DTFT, and $N \geq L$ equally spaced points in a period (between 0 and 2π), can be computed as

$$F_s[q] = \sum_{n=0}^{N-1} f[n] e^{-\mathrm{i}2\pi qn/N}, \qquad q = 0, 1, 2, ..., N-1\,. \tag{A.29}$$

Since $f[n]$ is of finite length L, for the interval $L < n \leq N-1$ we set $f[n] = 0$. The signal $f[n]$, padded with $N-L$ zeros at the end, is recovered from the discrete spectrum $F_s[q]$ via the Inverse Discrete Fourier Transform (IDFT)

$$f[n] = \frac{1}{N} \sum_{q=0}^{N-1} F_s[q] e^{\mathrm{i}2\pi qn/N}, \qquad n = 0, 1, 2, ..., N-1\,. \tag{A.30}$$

Both the DFT and the IDFT provide a framework to deal with frequency analysis in a computational environment, since they work with finite length and discrete signals and transforms. The sampling of the signal sets the maximum frequency (the bandwidth) of the spectrum, related to the Nyquist frequency, via its sampling frequency. Together with the length of the signal, it also sets the computational resolution of the spectrum, this is, the largest frequency interval that can be resolved. Note that padding the length L signal with $N-L$ zeros may seem to result in a spectrum with a better resolution, with increasing values of N. Yet, this is not true, since the maximum frequency resolution is related with the non-zero length of the signal L: when we consider $N \geq L$ points for the frequency resolution we do not get more information, but interpolations in frequency domain (Rao *et al.*, 2011; Orfanidis, 1996).

A.4.1 Conjugate symmetry property

The conjugate symmetry property in Equation (A.11), proved for Fourier series in Appendix A.1, is also verified in the discrete frequency domain. Usually, it is written recalling the periodicity of the DFT as

$$F[q] = F^*[-q] = F^*[N-q], \quad \text{for } q = 1, ..., N/2\,. \tag{A.31}$$

From Equation (A.31), one concludes the discrete spectrum is conjugate symmetric relative to zero frequency, and relative to the Nyquist frequency ($N/2$). Thus, in fact, q, $-q$ and $N-q$ refer to the same frequency. For convenience, the subscript $(\cdot)_s$, adopted earlier to distinguish between the continuous and discrete Fourier transform, has now been dropped.

A.4.2 Fast Fourier transform

The Fast Fourier Transform technique is extremely popular within the scientific community, due to the highly efficient algorithms available to make the computation of both the DFT and the IDFT. These algorithms are called Fast Fourier Transform (FFT) and Inverse Fast Fourier Transform (IFFT). For a comprehensive overview on the topic, the reader is referred to Rao *et al.* (2011).

Several of these methods employ a *divide and conquer* strategy, thus splitting the initial problem in smaller parts, in order to achieve maximum efficiency. For example, the algorithm Radix-2 is optimized to deal with signals of length 2^l, where l in a positive integer, so it is frequent to find methods that pad enough zeros to get a length verifying the previous expression (Rao *et al.*, 2011; Hu and Tonder, 1992). The same is true for Radix-3 and Radix-4 algorithms, for signals of length 3^l and 4^l. A robust FFT method will most often split the length of the signal in smaller lengths where it will apply a optimized algorithm to achieve maximum efficiency, even if the length of the signal is not a power of a specific number.

A.4.3 Convolution theorem. Fast convolution

Each of the previous Fourier transforms is associated with a specific version of the convolution theorem, relating the Fourier transform of the convolution of two functions, or signals, with the product of the Fourier transform of each one. This has been exemplified in Appendix A.2.1 for the continuous Fourier transform, where the convolution is *linear*. This operation is called linear in the sense that the relative displacement of the two functions is performed without further assumptions, in particular, the assumption of periodicity. The linear convolution between two finite length discrete signals f and g, whose lengths are M and N, respectively, can be written in a similar fashion as

$$(f * g)[m] = \sum_{n=0}^{m} f[n]g[m-n], \quad \text{for } m = 0, ..., M+N-1\,, \tag{A.32}$$

with $f = 0$ for $n > N-1$ and $n < 0$, and $g = 0$ for $m-n > M-1$ and $m-n < 0$. Note that the sum in Equation (A.32) is carried out from $n = 0$ to $n = m$, and not over all domain, knowing that both signals have matching points only in this range. This comes as a consequence of the signal's finite length. Then, one can still consider the sum over all domain, by assuming that zeros are padded at the ends of both signals.

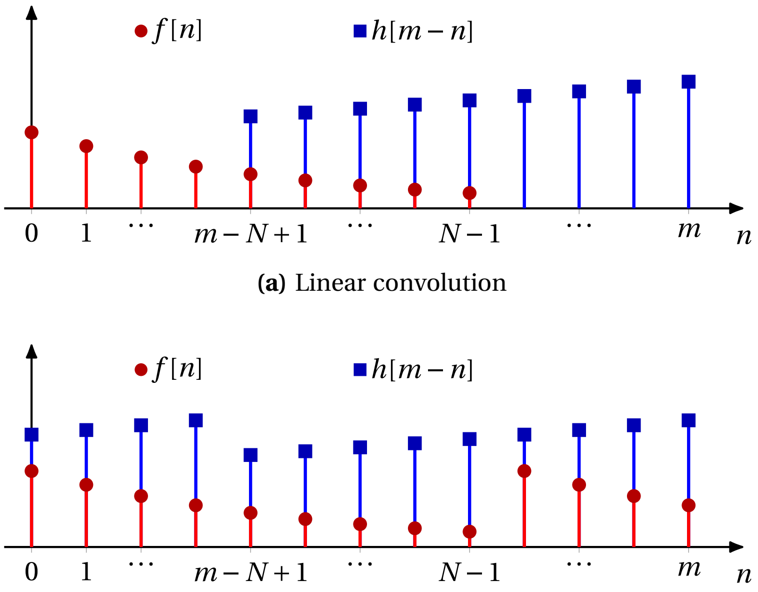
Refocusing on the convolution theorem, consider a new length-N signal h. For the DFT, the convolution theorem is stated as

$$\text{DFT}(f \circledast h) = \text{DFT}(f)\text{DFT}(h)\,, \tag{A.33}$$

where $f \circledast h$ denotes the circular convolution between $f[n]$ and $h[n]$. In turn, the circular convolution between discrete signals of same length is defined as

$$(f \circledast h)[m] = \sum_{n=0}^{N-1} f[n]h[m-n], \quad \text{for } m = 0, ...N-1\,, \tag{A.34}$$

with $f[N-n] = f[n]$ and $h[N-n] = h[n]$. This theorem is very similar to the convolution theorem presented earlier in Appendix A.2.1, the difference being the type of convolution involved—for the DFT, a circular convolution operation is used, rather than the linear convolution. The circular convolution assumes that the two functions being convoluted are periodic. Unlike the linear convolution operation, when one function is slided over the other, it must be extended periodically to fill the voids created by the relative displacement. In contrast, the extensions of the signal in a linear convolution operation are considered to be zero. Under these circumstances, the sum in a circular convolution can be taken all over the length of the signal, and is not restricted to the matching length. To clarify, Figure A.5 shows the positioning of two length-N signals being convoluted, for some value of displacement m. Starting by the linear convolution case in Figure A.5a, it can be seen that it is possible to compute the convolution of these two signals from $m = 0$ to $m = N + N - 1$—case where only one point from each signal match. For greater values of m, none of the points match, thus the linear convolution between the signals is zero. When it comes to the circular convolution, in Figure A.5b, the signal being displaced is periodically extended as it slides. As both signals have length N, their DFT will be of length N, because Equation (A.33) is a pointwise product. Furthermore, the circular convolution will be periodic of period N, differently from the length of the linear convolution between two length-N signals.



(a) Linear convolution

(b) Circular convolution

Figure A.5: Comparison between circular and linear convolution of two signals. The circular convolution, Figure A.5b, operation considers both signals as periodic, with period equal to their length, while linear convolution pads both signals with zeros at the ends.

Altogether, the convolution theorem for the DFT involves a circular convolution operation, yielding a different result than a linear convolution, in general. Yet, it suggests that, with little modification, a result equivalent to the linear convolution can be obtained.

Recall that the length of the linear convolution between f and g is equal to $N+M-1$, and that padding zeros at the end of signals does not change their frequency content. When $M-1$ zeros are padded to the end of $f[n]$ and $N-1$ zeros to the end of $g[n]$, it results in two signals of length $N+M-1$. Now, when $g[m-n]$ is slided over $f[n]$ (like in Figure A.5b), the periodic extensions will be replaced by zeros, and the result will be a circular convolution with period equal to $N+M-1$. Thus, circular convolution will be equal to the linear convolution, within each period. By using the convolution theorem from Equation (A.33), it follows the sought-after effect

$$(f * g)[m] = \mathrm{IDFT}\left(\mathrm{DFT}(f_{\mathrm{pad}}) \cdot \mathrm{DFT}(g_{\mathrm{pad}})\right) , \tag{A.35}$$

for $m = 0, \ldots, M+N+1$. Here, the signals f_{pad} and g_{pad} are the zero-padded versions of f and g. The linear convolution of two signals can be computed via DFT and IDFT using a zero-padding procedure—the *fast convolution* method. Instead of computing the convolution explicitly, it uses the highly efficient FFT algorithms to make the calculation, swapping the computational effort of the convolution to the computation of DFT's and IDFT's. This procedure is very efficient for the convolution of long signals (Orfanidis, 1996).

A.4.4 Correlation theorem

Similar to convolution theorem, the correlation theorem is a very important tool from the DFT's kit, specially regarding its application to rough surface analysis. Keeping the previous convention for the lengths of signals f, g and h, the linear correlation is defined as

$$(f \star g)[m] = \sum_{n=m}^{N-1} f[n]g[n+m], \quad \text{for } m = -M+1, \ldots, N-1 , \tag{A.36}$$

while verifying $f = 0$ for $n > N-1$ and $n < 0$, and $g = 0$ for $m-n > M-1$ and $m-n < 0$, like in linear convolution. In the correlation operation, g is displaced relative to f without taking the symmetry along x-axis, in contrast to the convolution operation. A circular correlation operation can also be defined between to signals with same length as

$$(f \circledast h)[m] = \sum_{n=m}^{N-1} f[n]h[n+m], \quad \text{for } m = 0, \ldots, N-1 , \tag{A.37}$$

which also assumes periodicity of both signals: $f[N-n] = f[n]$ and $h[N-n] = h[n]$. The correlation theorem follows

$$\mathrm{DFT}(f \circledast h) = \mathrm{DFT}(f) \cdot \mathrm{DFT}(h)^* . \tag{A.38}$$

The DFT of the correlation between two discrete signals is equal to the product of the *static* signal's DFT (f) and the conjugate of the *slidding* signal's DFT (g). A *fast correlation* method can also be derived from the fast convolution. Yet, in this case, one needs to pad $M-1$ zeros at the beginning of f and $N-1$ zeros at the end of g. Hence, the linear correlation is equal to a period of the circular correlation, which results in

$$(f \star g)[m] = \mathrm{IDFT}\left(\mathrm{DFT}(f_{-\mathrm{pad}}) \cdot \mathrm{DFT}(g_{\mathrm{pad}})^*\right) , \tag{A.39}$$

for $m = -M+1, ..., N-1$. The subscript $(\cdot)_{\text{-pad}}$ indicates the zeros are padded on the beginning of the signal, not at the end. A particular case arises when one computes the correlation of a function with itself, allowing some simplifications on its computation. First, the correlation theorem is reduced to

$$\mathrm{DFT}(f \circledast f) = \left|\mathrm{DFT}(f)\right|^2 . \tag{A.40}$$

Second, the correlation is symmetric relative to $m = 0$ and thus, in the fast correlation method, it is possible to pad $N-1$ zeros at the beginning of both copies of f, compute the correlation by Equation (A.39) and keep only the result for $m = 0$ to $N-1$.

A.5 Two-dimensional transforms

Fourier analysis and, in particular, Fourier transforms can readily be extended to higher dimensions. Here, the focus will be on the extension to two dimensions, due to its importance to rough surface analysis and numerical generation of random surfaces. The motivation behind Fourier analysis in one dimension is the decomposition of a single variable function as the superposition of sinusoidal 1D waves with varying frequency, amplitude and phase. On the two dimensional case, a function of of two variables, which can be interpreted as a surface in three dimensions, is obtained by a similar procedure. Although, the extension to two dimension brings an additional feature—the direction on which waves propagate. Thus, a function of two variables is obtained by summing sinusoidal waves which have different frequency, amplitude, phase and direction. In mathematical terms, each wave is characterized by a wave-vector $\boldsymbol{k} = (k_x, k_y)$, where its components are the frequencies on each direction, i.e., the frequency of the 1D wave generated by intercepting the 2D wave with the coordinate planes, its magnitude is the real frequency of the wave, and its direction is the direction on which the 2D wave propagates. Moreover, each wave is characterized by its amplitude and phase, which set the position of the wave on its mean plane. Figure A.6 shows three 2D waves with different wave-vectors having different frequency, amplitude and direction. The original function is recovered by spanning all wave-vector space and summing all waves with correct amplitude and phase.

The key property used in the extension of the previous concepts for two and higher dimensions is the separability of Fourier transforms. That is to say that Fourier transforms in higher dimensions are computed as 1D transforms for each direction, over all dimensions, without any dependence between dimensions. By this logic, the definition of continuous Fourier transform in two dimension is straightforward, and it comes

$$\mathcal{F}\left\{f(x,y)\right\} = F(k_x,k_y) = \iint_{-\infty}^{+\infty} f(x,y)e^{-\mathrm{i}(k_x x + k_y y)}\,\mathrm{d}x\mathrm{d}y\,, \tag{A.41}$$

and the inverse transform writes

$$f(x,y) = \mathcal{F}^{-1}\left\{F(k_x,k_y)\right\} = \frac{1}{4\pi^2}\iint_{-\infty}^{+\infty} F(k_x,k_y)e^{\mathrm{i}(k_x x + k_y y)}\,\mathrm{d}k_x\mathrm{d}k_y\,. \tag{A.42}$$

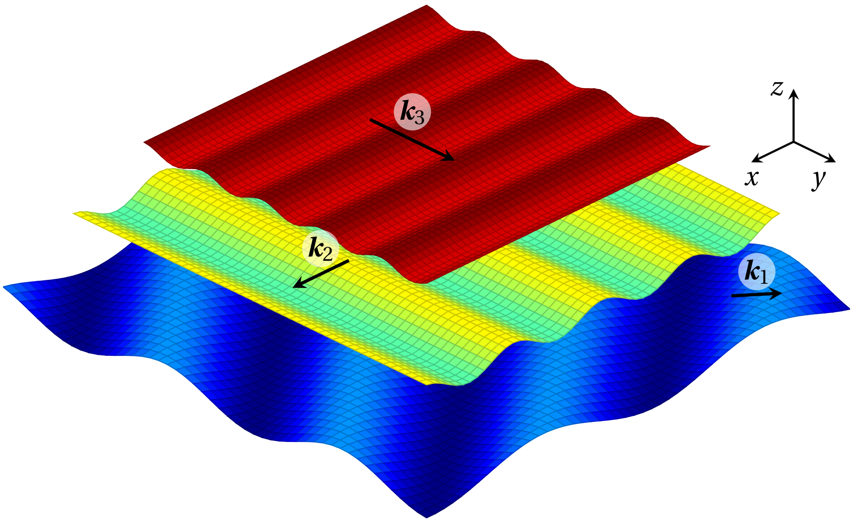


Figure A.6: Superposition of 2D waves for Fourier analysis in two dimensions. The building blocks of two dimensional Fourier analysis are 2D waves defined by some wave-vector $\boldsymbol{k}$, which contains information on the frequency and direction of the wave. Each wave is also characterized by its amplitude and phase. In the present figure, three waves with varying frequency, direction and magnitude are plotted. A change of phase corresponds to a translation in the wave's mean plane.

Regarding the DFT, it is considered that the function is sampled at a uniformly spaced grid in each direction of $M \times N$ points, where M and N is the number of points sampled in y and x direction, respectively. The sampled points are spaced by $\lambda_{s,x}$ in the x direction and by $\lambda_{s,y}$ in the y-direction, such that a general point (x, y) is written as $(q\lambda_{s,x}, p\lambda_{s,y})$. Under these circumstances, the DFT is defined as

$$F[p,q] = \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} f[m,n]e^{-\mathrm{i}2\pi(pm/M+qn/N)}, \quad \begin{cases} p = 0,1,\ldots,M-1 \\ q = 0,1,\ldots,N-1 \end{cases}, \tag{A.43}$$

and the IDFT comes

$$f[m,n] = \frac{1}{MN}\sum_{p=0}^{M-1}\sum_{q=0}^{N-1} F[p,q]e^{\mathrm{i}2\pi(pm/M+qn/N)}, \quad \begin{cases} m = 0,1,\ldots,M-1 \\ n = 0,1,\ldots,N-1 \end{cases}. \tag{A.44}$$

It should be noted that the order of the transform arguments is inverted in the discrete case. This is, in the continuous transform, the argument order is (k_x, k_y), wherein in the discrete transform is $[p, q]$, where p directs to the frequency k_y and q to frequency k_x. With this convention, the array element $[p, q]$ matches the geometrical position of its (x, y) points, where x increases from left to right and y increases from top to bottom.

Moreover, it is useful to restate some of the properties of Fourier transforms for the two dimensional case. Starting with the conjugate symmetry property, Figure A.6 leads to the geometrical argument that symmetric wave-vectors represent the same wave. Hence, the contribution of both wave-vectors to the original function must be the same, and the imaginary part associated with each one must cancel out when the transform is applied to a real valued function. Consequently, the conjugate symmetry property in two dimensions for the continuous transform follows

$$F(\boldsymbol{k}) = F^*(\boldsymbol{k})\,. \tag{A.45}$$

The formulation of this property for the discrete transform is also straightforward. Similarly to the one dimensional case, it uses the periodicity of the discrete transform in each direction, i.e.

$$F[p,q] = F[p+M,q] = F[p,q+N] = F[p+M,q+N]\,, \tag{A.46}$$

for $p = 0,...,M-1$ and $q = 0,...,N-1$. The property is then stated as

$$F[p,q] = F^*[-p,-q] = F^*[M-p,N-q]\,. \tag{A.47}$$

In two dimensions, the conjugate symmetry is relative to the origin, both in the continuous and discrete transform, In addition, due to the implicit periodicity of the discrete transform, it can also be interpreted as the conjugate symmetry relative to the Nyquist wave-vector $(M/2,\ N/2)$. The sampling of the original function is performed in two directions, hence there are two Nyquist frequencies, one for each direction:

$$k_{\text{nyq},x} = \frac{k_{s,x}}{2} \tag{A.48a}$$

and

$$k_{\text{nyq},y} = \frac{k_{s,y}}{2}. \tag{A.48b}$$

Consequently, frequency aliasing can occur in each direction on its own, meaning that p, $-p$ and $M-p$ refer to same frequency in the y direction, whilst q, $-q$ and $N-q$ refer to the same frequency in the y direction.

Finally, coming to the linear convolution and correlation in two dimensions, consider two discrete functions f and g, with sizes $M \times N$ and $P \times Q$, respectively. For two dimensional discrete functions, the linear convolution is then defined as

$$(f * g)[m,n] = \sum_{m=0}^{p}\sum_{n=0}^{q} f[m,n]g[p-m,q-n], \quad \text{for } \begin{cases} m = 0,1,...,M+P-1 \\ n = 0,1,...,N+Q-1 \end{cases}, \tag{A.49}$$

and linear correlation as

$$(f \star g)[m,n] = \sum_{m=p}^{M-p}\sum_{n=q}^{N-q} f[m,n]g[m+p,n+q]. \quad \text{for } \begin{cases} m = 1-P,...,M-1 \\ n = 1-Q,...,N-1 \end{cases}. \tag{A.50}$$

Fast computation of both operations can still be achieved via FFT in two dimension. Yet, the zero padding procedure must be performed in both directions.

Appendix B

Recipe for BGT model computation

The micromechanical contact model from Bush-Gibson-Thomas relies on the integration of a function defined through implicit relations, coming from Hertz contact theory. In order to perform the integration, the authors derived an alternative expression for the integrand function, by carrying a laborious analytical work, which started from a change of variables. Here, a brief recipe for the computation of real contact area and load is provided, based on the original work (Bush, Gibson, and Thomas, 1975) and on a remark made by Carbone and Bottiglione (2008) regarding a misprint in some expressions in the original publication. The following text aims solely at expressing the algorithmic computation sequence—the physical and mathematical significance of the results will only be mentioned for key parameters. For a complete derivation and meaning of the symbols, the reader is referred to two previously cited works.

By analytical manipulation, it results that the real contact area fraction can be computed from the following expression

$$\frac{A(t)}{A_c} = \frac{12\alpha}{\pi}\sqrt{\frac{3}{2\alpha-3}}\exp\left(-\frac{\alpha t^2}{2\alpha-3}\right)\int_{\varphi=0}^{\pi/2}\int_{\Theta=0}^{\pi/4}\cos\varphi\sin^3\varphi f(\theta,\Theta)g(\varphi,\theta)\frac{\mathrm{d}\theta}{\mathrm{d}\Theta}\,\mathrm{d}\Theta\mathrm{d}\varphi\,. \quad \text{(B.1)}$$

Numerical calculation of Equation (B.1) requires the input of the dimensionless separation $t = z_s/\sigma_z = \sqrt{m_0}$ and Nayak's parameter α. For each value of Θ in the aforementioned integration, the computation of the variable θ and its derivative $\mathrm{d}\theta/\mathrm{d}\Theta$ starts from determining the parameter k by

$$k = \sqrt{1-\tan^2\Theta}\,. \quad \text{(B.2)}$$

Then, the value of complete elliptical integral of first and second kind K and E, respectively, are computed by their definition

$$K = K(k) = \int_0^{\pi/2}\frac{1}{\sqrt{1-k^2\sin^2\psi}}\,\mathrm{d}\psi\,; \quad \text{(B.3)}$$

$$E = E(k) = \int_0^{\pi/2}\sqrt{1-k^2\sin^2\psi}\,\mathrm{d}\psi\,. \quad \text{(B.4)}$$

Next, the derivative of K in order to k can be found from the following functional relationship,

$$K' = \frac{\mathrm{d}K}{\mathrm{d}k} = \frac{E}{k(1-k^2)} - \frac{K}{k}\,. \tag{B.5}$$

Finally, θ can be computed from the previous values, and its derivative in order to Θ follows from the values of θ:

$$\theta = \arctan\sqrt{\frac{kK-(1-k^2)K'}{K'}}\,; \tag{B.6}$$

$$\frac{\mathrm{d}\theta}{\mathrm{d}\Theta} = -\frac{\tan\Theta\cos^2\theta}{2\tan\theta\cos^2\Theta}\left(\frac{3E-2K}{E-(1-k^2)K} - \frac{(1-k^2)(K-E)K}{(E-(1-k^2)K)^2}\right). \tag{B.7}$$

Having computed values of θ and its derivative for each Θ, one shall now be concerned with the function $f(\theta,\Theta)$ and $g(\varphi,\theta)$. The expression for $f(\theta,\Theta)$ writes

$$f(\theta,\Theta) = \frac{\cos2\theta\sin^3\theta\cos\theta\tan\Theta}{\tan^2\theta+\tan^2\Theta}\,. \tag{B.8}$$

As for $g(\varphi,\theta)$, its formulation is rather involved, and is conveniently expressed in a sequential manner. For each φ and θ, $g(\varphi,\theta)$ is computed by the following the next steps:

(i) Get the parameter C:

$$C = \frac{9}{2}(\alpha-1) - \frac{3}{2}(2\alpha-3)\sin^2 2\theta\,. \tag{B.9}$$

(ii) Compute γ and η from:

$$\gamma = \frac{1}{2\alpha-3}\left(\alpha\cos^2\varphi - 3\sqrt{\alpha}\cos\varphi\sin\varphi + C\sin^2\varphi\right); \tag{B.10}$$

$$\eta = \frac{t}{2\alpha-3}\left(3\sqrt{\alpha}\sin\varphi - 2\alpha\cos\varphi\right). \tag{B.11}$$

(iii) By using γ and η, find λ as

$$\lambda = \frac{\eta}{\sqrt{2\gamma}}\,. \tag{B.12}$$

(iv) Next, compute Λ from λ:

$$\Lambda = \sqrt{\frac{\pi}{2}}\exp\left(\frac{\lambda^2}{2}\right)\left(1+\mathrm{erf}\left(\frac{\lambda}{\sqrt{2}}\right)\right). \tag{B.13}$$

(v) Gathering all previous results, function $g(\varphi,\theta)$ comes

$$g(\varphi,\theta) = \frac{1}{8\gamma^3}\left(8+9\lambda^2+\lambda^4+\left(15\lambda+10\lambda^3+\lambda^5\right)\Lambda\right). \tag{B.14}$$

At this point, Equation (B.1) can be recovered, and the real contact area can be computed with numerical quadrature.

Similarly, the applied load for each dimensionless separation t also comes as a function of α. Is is determined by

$$\begin{aligned}\frac{F}{AE^*\sqrt{m_2}} = \frac{8\sqrt{3}\alpha^{5/4}}{\pi\sqrt{2\alpha-3}}\int_t^\infty\int_0^\infty\int_0^{\pi/4} & P(\theta,\Theta)(z-t)^{3/2}w^{7/2} \\ & \cdot\exp\left(\frac{-\alpha z^2+3\sqrt{\alpha}zw-Cw^2}{2\alpha-3}\right)\frac{\mathrm{d}\theta}{\mathrm{d}\Theta}\,\mathrm{d}\Theta\mathrm{d}w\mathrm{d}z\,.\end{aligned} \tag{B.15}$$

Values for θ, its derivative in order to Θ and C are computed from the same sequence described for real contact area determination. The function $P(\theta,\Theta)$ is defined as

$$P(\theta,\Theta) = \frac{\sin^3\theta\cos^2\theta\cos 2\theta}{K_2\sqrt{\tan^2\theta+\tan^2\Theta}}\,, \tag{B.16}$$

where K_2 is the complete elliptic integral of first kind, computed with $k_2=\sqrt{1-\tan^2\Theta}$,

$$K_2 = K(k_2=\sqrt{1-\tan^2\Theta}) = \int_0^{\pi/2}\frac{1}{\sqrt{1-k_2^2\sin^2\psi}}\,\mathrm{d}\psi\,. \tag{B.17}$$

Equations (B.1) and (B.15) predict real contact and load both as functions of the dimensionless separation t, which allow an indirect relation between area and load to be established.

Page intentionally left blank

Appendix C

Determination of RMS parameters from spectral properties

In the following paragraphs, the analytical expressions describing RMS parameters in terms of the spectral properties are derived, regarding both 2D and 3D topographies. Profile parameters will follow a direct differentiation of the ACF and, afterwards, the relation with spectral moments is established. As for rough surfaces, an approach for the computation of RMS parameters based on surface synthesis through inverse Fourier transform is adopted.

C.1 Profile RMS parameters from ACF derivatives

The autocorrelation function has been formally defined as

$$R(\tau) = \frac{1}{L-\tau}\int_0^{L-\tau} z(x)z(x+\tau)\ \mathrm{d}x\,. \tag{C.1}$$

The simplest result that can be derived from Equation (C.1) is the RMS height, as demonstrated previously. It comes directly from the definition of RMS height that

$$R(0) = z^2_{\mathrm{rms},x}\,. \tag{C.2}$$

Thus, the derivative of zeroth order of the ACF, computed at the origin, equals the square of the profile RMS height. Similar results can be written for RMS slope and curvature, from a more elaborate algebraic set up. For the proceeding analysis, it will be assumed that the sampling length is infinite, such that the dependency of the denominator and integration limits on τ, in Equation (C.1), can be removed, i.e.

$$R(\tau) = \lim_{L\to\infty}\frac{1}{L}\int_{-\frac{L}{2}}^{+\frac{L}{2}} z(x)z(x+\tau)\ \mathrm{d}x\,. \tag{C.3}$$

First, the ACF is differentiated in order to the shift τ. Since the integration limits do not depend on the τ, the derivative operation can be passed to the integrand function. This procedure repeats for the second derivative, as well. From the linearity properties of the derivative operator, the derivative in order to τ can be replaced by the derivative relative to x. Hence, the second derivative of the ACF rewrites

$$\frac{\mathrm{d}^2 R(\tau)}{\mathrm{d}\tau^2} = \lim_{L\to\infty} \frac{1}{L} \int_{-\frac{L}{2}}^{+\frac{L}{2}} z(x) \frac{\mathrm{d}^2 z(x+\tau)}{\mathrm{d}x^2}\, \mathrm{d}x\,. \tag{C.4}$$

At this stage, it is possible to integrate Equation (C.4) by parts and, dropping the limit to infinity for clarity, it comes

$$\frac{\mathrm{d}^2 R(\tau)}{\mathrm{d}\tau^2} = \frac{1}{L} z(x) \frac{\mathrm{d}z(x+\tau)}{\mathrm{d}\tau}\bigg|_{\partial L} - \frac{1}{L}\int_L \frac{\mathrm{d}z(x)}{\mathrm{d}x}\,\frac{\mathrm{d}z(x+\tau)}{\mathrm{d}x}\,\mathrm{d}x\,, \tag{C.5}$$

where ∂L denotes the boundary of the integration domain—in this case, it corresponds to the first and last x point in the integration line. The first term in the second member in Equation (C.5) is null (because L is infinite) while surface height and slope are always finite valued. Thus, it comes that the second derivative of the ACF at the origin is equal to the symmetric of RMS slope squared,

$$\frac{\mathrm{d}^2 R(\tau)}{\mathrm{d}\tau^2}\bigg|_{\tau=0} = -\left(z'_{\mathrm{rms},x}\right)^2\,. \tag{C.6}$$

This procedure can be repeated for all derivatives of even order. In particular, regarding the fourth order derivative of the ACF, it starts from the expression

$$\frac{\mathrm{d}^2 R(\tau)}{\mathrm{d}\tau^2} = -\frac{1}{L}\int_L \frac{\mathrm{d}z(x)}{\mathrm{d}x}\,\frac{\mathrm{d}z(x+\tau)}{\mathrm{d}x}\,\mathrm{d}x\,, \tag{C.7}$$

and by the analogy with Equation (C.1), one can compute its second derivative relative to τ by taking the previous analytical steps, which will yield the square of RMS curvature

$$\frac{\mathrm{d}^4 R(\tau)}{\mathrm{d}\tau^4}\bigg|_{\tau=0} = \left(z''_{\mathrm{rms},x}\right)^2\,. \tag{C.8}$$

C.2 Profile RMS parameters from spectral moments

An alternative definition of the autocorrelation function follows from the inverse Fourier transform of the profile PSD,

$$R(\tau) = \mathcal{F}^{-1}\{\Phi_\theta(k)\} = \frac{1}{2\pi}\int_{-\infty}^{+\infty} \Phi_\theta(k) e^{\mathrm{i}k\tau}\,\mathrm{d}k\,. \tag{C.9}$$

Additionally, recall the definition of profile spectral moment $m_{\theta n}$ as

$$m_{\theta p} = \frac{1}{2\pi}\int_{-\infty}^{+\infty} k^p \Phi_\theta(k)\,\mathrm{d}k\,. \tag{C.10}$$

In order to establish the relation between the quantities computed in the former section and the profile PSD, Equation (C.9) is differentiated in order to τ two and four times, followed by the computation of their value at the origin

$$\left.\frac{\mathrm{d}^2 R(\tau)}{\mathrm{d}\tau^2}\right|_{\tau=0} = -\left(z'_{\mathrm{rms},x}\right)^2 = -\frac{1}{2\pi}\int_{-\infty}^{+\infty} k^2 \Phi_\theta(k)\,\mathrm{d}k\,; \tag{C.11}$$

$$\left.\frac{\mathrm{d}^4 R(\tau)}{\mathrm{d}\tau^4}\right|_{\tau=0} = \left(z''_{\mathrm{rms},x}\right)^2 = \frac{1}{2\pi}\int_{-\infty}^{+\infty} k^4 \Phi_\theta(k)\mathrm{d}k\,. \tag{C.12}$$

Lastly, introducing the definition of profile spectral moments in Equations (C.9), (C.11) and (C.12), it allows to write the relations between profile RMS parameters and the respective profile spectral moments

$$z_{\mathrm{rms},x} = \sqrt{m_{\theta 0}}\,; \tag{C.13a}$$

$$z'_{\mathrm{rms},x} = \sqrt{m_{\theta 2}}\,; \tag{C.13b}$$

$$z''_{\mathrm{rms},x} = \sqrt{m_{\theta 4}}\,. \tag{C.13c}$$

C.3 Surface RMS parameters from spectral moments

The computation of surface RMS parameters from direct derivation of the ACF is considerably tougher. This is mainly due to the complexity inherent to integration by parts in two dimensions, which makes the analytical tasks laborious, and rendering this approach unattractive. An alternative methodology relies on the topography synthesis via inverse Fourier transform,

$$z(x,y) = \frac{1}{4\pi^2}\iint_{-\infty}^{+\infty} \sqrt{\Phi(k_x,k_y)L_xL_y}\; e^{\mathrm{i}(k_x x + k_y y + \phi(k_x,k_y))}\,\mathrm{d}k_x\mathrm{d}k_y\,, \tag{C.14}$$

and on the autocorrelation theorem for Fourier transforms,

$$\mathcal{F}\left\{f(x,y)\star g(x,y)\right\} = \mathcal{F}\left\{f(x,y)\right\}\mathcal{F}\left\{g(x,y)\right\}^*\,. \tag{C.15}$$

The correlation operation defined as

$$(f\star g)(\tau_x,\tau_y) = \lim_{L_x,L_y\to\infty}\int_{-\frac{L_y}{2}}^{+\frac{L_y}{2}}\int_{-\frac{L_x}{2}}^{+\frac{L_x}{2}} f(x,y)g(x+\tau_x,y+\tau_y)\,\mathrm{d}x\mathrm{d}y\,. \tag{C.16}$$

Building on these results, one can write the correlation of surface height with itself (autocorrelation) as

$$(z\star z)(\tau_x,\tau_y) = \frac{1}{4\pi^2}\iint_{-\infty}^{+\infty}\Phi(k_x,k_y)L_xL_y\; e^{-\mathrm{i}(k_x\tau_x + k_y\tau_y)}\,\mathrm{d}k_x\mathrm{d}k_y\,. \tag{C.17}$$

Dividing Equation (C.17) by the domain area L_xL_y, computing the value of this new quantity at the origin $\tau_x = \tau_y = 0$, and recalling the definition of surface spectral moment,

$$m_{pq} = \frac{1}{4\pi^2}\iint_{-\infty}^{+\infty} k_x^p k_y^q \Phi(\boldsymbol{k})\mathrm{d}k_x\mathrm{d}k_y\,, \tag{C.18}$$

the surface RMS height comes

$$z_{\mathrm{rms},xy} = \sqrt{m_{00}}\,. \tag{C.19}$$

Focusing now on the RMS slope, start by computing the gradient of surface height, in Equation (C.14), which yields

$$\nabla z(x,y) = \frac{1}{4\pi^2} \iint_{-\infty}^{+\infty} \mathrm{i}\boldsymbol{k}\sqrt{\Phi(k_x,k_y)L_xL_y}\; e^{\mathrm{i}(k_x x + k_y y + \phi(k_x,k_y))}\,\mathrm{d}k_x\mathrm{d}k_y\,. \tag{C.20}$$

Identically to RMS height calculation, one proceeds to compute the element-wise autocorrelation of the surface gradient, i.e., the correlation of each vector component with itself, and sum the results from both x and y directions. Again, dividing the result by the domain's area, and taking the value at the origin, it comes

$$\frac{1}{L_xL_y}\|\nabla z(x,y)\|^2 = \frac{1}{L_xL_y}\frac{1}{4\pi^2}\iint_{-\infty}^{+\infty}\|\boldsymbol{k}\|^2\Phi(k_x,k_y)\,\mathrm{d}k_x\mathrm{d}k_y\,, \tag{C.21}$$

which rewrites

$$z'_{\mathrm{rms},xy} = \sqrt{m_{20}+m_{02}}\,. \tag{C.22}$$

At last, in order to compute the RMS mean surface curvature, and by following the sequence of increasing derivative order from one parameter to the other, consider the Laplacian of surface height

$$\nabla^2 z(x,y) = -\frac{1}{4\pi^2}\iint_{-\infty}^{+\infty}\left(k_x^2+k_y^2\right)\sqrt{\Phi(k_x,k_y)L_xL_y}\; e^{\mathrm{i}(k_x x + k_y y + \phi(k_x,k_y))}\,\mathrm{d}k_x\mathrm{d}k_y\,. \tag{C.23}$$

Next, the the same analytical steps performed in former derivations are repeated, i.e., finding the autocorrelation of the surface height Laplacian, followed by spatial averaging and extracting the value at the origin. Additionally, the result is divided by 4, in order to obtain the average curvature between both directions. The surface RMS then reads

$$\frac{1}{4L_xL_y}\left(\nabla^2 z(x,y)\right)^2 = \frac{1}{4L_xL_y}\frac{1}{4\pi^2}\iint_{-\infty}^{+\infty}\left(k_x^2+k_y^2\right)^2\Phi(k_x,k_y)\,\mathrm{d}k_x\mathrm{d}k_y\,, \tag{C.24}$$

and introducing the surface spectral moments

$$z''_{\mathrm{rms},xy} = \sqrt{\frac{m_{40}+2m_{22}+m_{04}}{4}}\,. \tag{C.25}$$

C.4 Summary

Autocorrelation function - Profile RMS parameters

$$R(0) = \left(z_{\mathrm{rms},x}\right)^2; \tag{C.26}$$

$$\left.\frac{\mathrm{d}^2R(\tau)}{\mathrm{d}\tau^2}\right|_{\tau=0} = -\left(z'_{\mathrm{rms},x}\right)^2; \tag{C.27}$$

$$\left.\frac{\mathrm{d}^4R(\tau)}{\mathrm{d}\tau^4}\right|_{\tau=0} = \left(z''_{\mathrm{rms},x}\right)^2. \tag{C.28}$$

Profile RMS parameters - Spectral moments

$$z_{\mathrm{rms},x} = \sqrt{m_{\theta 0}}\,; \tag{C.29}$$

$$z'_{\mathrm{rms},x} = \sqrt{m_{\theta 2}}\,; \tag{C.30}$$

$$z''_{\mathrm{rms},x} = \sqrt{m_{\theta 4}}\,. \tag{C.31}$$

Surface RMS parameters - Spectral moments

$$z_{\mathrm{rms},xy} = \sqrt{m_{00}}\,; \tag{C.32}$$

$$z'_{\mathrm{rms},xy} = \sqrt{m_{20} + m_{02}}\,; \tag{C.33}$$

$$z''_{\mathrm{rms},xy} = \sqrt{\frac{m_{40} + 2m_{22} + m_{04}}{4}}\,. \tag{C.34}$$

Page intentionally left blank

References

Afferrante, L., Carbone, G., and Demelio, G. (2012). “Interacting and coalescing Hertzian asperities: A new multiasperity contact model”. In: *Wear* 278-279, pp. 28–33.

Alart, P. and Curnier, A. (1991). “A mixed formulation for frictional contact problems prone to Newton like solution methods”. In: *Computer Methods in Applied Mechanics and Engineering* 92 (3), pp. 353–375.

Anciaux, G., Ramisetti, S. B., and Molinari, J. F. (2012). “A finite temperature bridging domain method for MD-FE coupling and application to a contact problem”. In: *Computer Methods in Applied Mechanics and Engineering*. Special Issue on Advances in Computational Methods in Contact Mechanics 205-208, pp. 204–212.

Ao, Y., Wang, Q. J., and Chen, P. (2002). “Simulating the worn surface in a wear process”. In: *Wear* 252 (1), pp. 37–47.

Archard, J. F. (1957). “Elastic deformation and the laws of friction”. In: *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* 243 (1233), pp. 190–205.

ASME B46 (2009). *Surface Texture (Surface Roughness, Waviness, and Lay)*.

Bakolas, V. (2003). “Numerical generation of arbitrarily oriented non-Gaussian three-dimensional rough surfaces”. In: *Wear* 254 (5), pp. 546–554.

Bandeira, A. A., Pimenta, P. M., and Wriggers, P. (2008). “A 3D contact investigation of rough surfaces considering elastoplasticity”. In: *Exacta* 6 (1), pp. 109–118.

Bandeira, A. A., Wriggers, P., and Pimenta, P. d. M. (2004). “Numerical derivation of contact mechanics interface laws using a finite element approach for large 3D deformation”. In: *International Journal for Numerical Methods in Engineering* 59 (2), pp. 173–195.

Belgacem, F. B., Hild, P., and Laborde, P. (1998). “The mortar finite element method for contact problems”. In: *Mathematical and Computer Modelling*. Recent Advances in Contact Mechanics 28 (4), pp. 263–271.

Bernardi, C., Maday, Y., and Patera, A. T. (1993). “Domain Decomposition by the Mortar Element Method”. In: *Asymptotic and Numerical Methods for Partial Differential Equations with Critical Parameters.* Ed. by H. G. Kaper, M. Garbey, and G. W. Pieper. NATO ASI Series. Dordrecht: Springer Netherlands, pp. 269–286.

Berthier, Y. *et al.* (2004). “The role and effects of the third body in the wheel–rail interaction”. In: *Fatigue & Fracture of Engineering Materials & Structures* 27 (5), pp. 423–436.

Bhushan, B. (1998). *Handbook of Micro/Nano Tribology, Second Edition.* CRC Press. 882 pp.

Bonet, J. and Wood, R. D. (2008). *Nonlinear Continuum Mechanics for Finite Element Analysis.* Cambridge University Press. 317 pp.

Borri-Brunetto, M., Carpinteri, A., and Chiaia, B. (1999). “Scaling phenomena due to fractal contact in concrete and rock fractures”. In: *International Journal of Fracture* 95 (1), p. 221.

Bowden, F. P., Bowden, F. P., and Tabor, D. (2001). *The Friction and Lubrication of Solids.* Clarendon Press. 432 pp.

Bruzzone, A. A. G. *et al.* (2008). “Advances in engineered surfaces for functional performance”. In: *CIRP Annals* 57 (2), pp. 750–769.

Bush, A. W., Gibson, R. D., and Keogh, G. P. (1976). “The limit of elastic deformation in the contact of rough surfaces”. In: *Mechanics Research Communications* 3 (3), pp. 169–174.

Bush, A. W., Gibson, R. D., and Keogh, G. P. (1979). “Strongly Anisotropic Rough Surfaces”. In: *Journal of Lubrication Technology* 101 (1), pp. 15–20.

Bush, A. W., Gibson, R. D., and Thomas, T. R. (1975). “The elastic contact of a rough surface”. In: *Wear* 35 (1), pp. 87–111.

Campañá, C. and Müser, M. H. (2007). “Contact mechanics of real vs. randomly rough surfaces: A Green\textquotesingles function molecular dynamics study”. In: *Europhysics Letters (EPL)* 77 (3), p. 38005.

Campañá, C., Persson, B. N. J., and Müser, M. H. (2011). “Transverse and normal interfacial stiffness of solids with randomly rough surfaces”. In: *Journal of Physics: Condensed Matter* 23 (8), p. 085001.

Campañá, C. and Müser, M. H. (2006). “Practical Green’s function approach to the simulation of elastic semi-infinite solids”. In: *Physical Review B* 74 (7), p. 075420.

Campañá, C., Müser, M. H., and Robbins, M. O. (2008). “Elastic contact between self-affine surfaces: comparison of numerical stress and contact correlation functions with analytic predictions”. In: *Journal of Physics: Condensed Matter* 20 (35), p. 354013.

Carbone, G. and Bottiglione, F. (2008). “Asperity contact theories: Do they predict linearity between contact area and load?” In: *Journal of the Mechanics and Physics of Solids* 56 (8), pp. 2555–2572.

Carbone, G., Lorenz, B., *et al.* (2009). “Contact mechanics and rubber friction for randomly rough surfaces with anisotropic statistical properties”. In: *The European Physical Journal E* 29 (3), pp. 275–284.

Carbone, G., Scaraggi, M., and Tartaglino, U. (2009). “Adhesive contact of rough surfaces: Comparison between numerical calculations and analytical theories”. In: *The European Physical Journal E* 30 (1), p. 65.

Chaparro, L. (2010). *Signals and Systems using MATLAB.* Academic Press. 769 pp.

Ciavarella, M. (2016). “On the Significance of Asperity Models Predictions of Rough Contact With Respect to Recent Alternative Theories”. In: *Journal of Tribology* 139 (2), pp. 021402–021402–4.

Ciavarella, M., Delfine, V., and Demelio, G. (2006). “A “re-vitalized” Greenwood and Williamson model of elastic contact between fractal surfaces”. In: *Journal of the Mechanics and Physics of Solids* 54 (12), pp. 2569–2591.

Ciavarella, M., Greenwood, J. A., and Paggi, M. (2008). “Inclusion of “interaction” in the Greenwood and Williamson contact theory”. In: *Wear* 265 (5), pp. 729–734.

Dapp, W. B., Lücke, A., *et al.* (2012). “Self-Affine Elastic Contacts: Percolation and Leakage”. In: *Physical Review Letters* 108 (24), p. 244301.

Dapp, W. B., Prodanov, N., and Müser, M. H. (2014). “Systematic analysis of Persson’s contact mechanics theory of randomly rough elastic surfaces”. In: *Journal of Physics: Condensed Matter* 26 (35), p. 355002.

De Lorenzis, L. and Wriggers, P. (2013). “Computational homogenization of rubber friction on rough rigid surfaces”. In: *Computational Materials Science* 77, pp. 264–280.

DeVries, W. R. (1979). “Autoregressive Time Series Models ofr Surface Profile Characterization”. In: *CIRP Annals* 28 (1), pp. 437–440.

Dieterich, J. H. and Kilgore, B. D. (1994). “Direct observation of frictional contacts: New insights for state-dependent properties”. In: *Pure and Applied Geophysics* 143 (1), pp. 283–302.

Dodds, C. J. and Robson, J. D. (1973). “The description of road surface roughness”. In: *Journal of Sound and Vibration* 31 (2), pp. 175–183.

Einax, M., Dieterich, W., and Maass, P. (2013). “Colloquium: Cluster growth on surfaces - densities, size distributions and morphologies”. In: *Reviews of Modern Physics* 85 (3), pp. 921–939.

Elderton, W. P. and Johnson, N. L. (1969). *Systems of frequency curves.* At the University Press. 216 pp.

Farah, P., Popp, A., and Wall, W. A. (2015). "Segment-based vs. element-based integration for mortar methods in computational contact mechanics". In: *Computational Mechanics* 55 (1), pp. 209–228.

Flemisch, B. and Wohlmuth, B. I. (2007). "Stable Lagrange multipliers for quadrilateral meshes of curved interfaces in 3D". In: *Computer Methods in Applied Mechanics and Engineering*. Domain Decomposition Methods: recent advances and new challenges in engineering 196 (8), pp. 1589–1602.

Francavilla, A. and Zienkiewicz, O. C. (1975). "A note on numerical computation of elastic contact problems". In: *International Journal for Numerical Methods in Engineering* 9 (4), pp. 913–924.

Francisco, A. and Brunetière, N. (2016). "A hybrid method for fast and efficient rough surface generation". In: *Proceedings of the Institution of Mechanical Engineers, Part J: Journal of Engineering Tribology* 230 (7), pp. 747–768.

Ganti, S. and Bhushan, B. (1995). "Generalized fractal analysis and its applications to engineering surfaces". In: *Wear* 180 (1), pp. 17–34.

Geuzaine, C. and Remacle, J.-F. (2009). "Gmsh: A 3-D finite element mesh generator with built-in pre- and post-processing facilities". In: *International Journal for Numerical Methods in Engineering* 79 (11), pp. 1309–1331.

Geuzaine, C. and Remacle, J.-F. (2019). *Gmsh Reference Manual.*

Greenwood J. A. and Langstreth J. K. (1984). "A unified theory of surface roughness". In: *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences* 393 (1804), pp. 133–157.

Greenwood J. A. and Williamson J. B. P. (1966). "Contact of nominally flat surfaces". In: *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* 295 (1442), pp. 300–319.

Greenwood, J. A. (2006). "A simplified elliptic model of rough surface contact". In: *Wear* 261 (2), pp. 191–200.

Greenwood, J. A., Putignano, C., and Ciavarella, M. (2011). "A Greenwood & Williamson theory for line contact". In: *Wear* 270 (3), pp. 332–334.

Greenwood, J. and Wu, J. (2001). "Surface Roughness and Contact: An Apology". In: *Meccanica* 36 (6), pp. 617–630.

Gu, X. and Huang, Y. (1990). "The modelling and simulation of a rough surface". In: *Wear* 137 (2), pp. 275–285.

Hall, P. and Davies, S. (1995). "On direction-invariance of fractal dimension on a surface". In: *Applied Physics A* 60 (3), pp. 271–274.

Hallquist, J. O. (1979). *NIKE2D: an implicit, finite-deformation, finite-element code for analyzing the static and dynamic response of two-dimensional solids.* UCRL-52678. California Univ., Livermore (USA). Lawrence Livermore Lab.

Hartmann, S. *et al.* (2009). "A contact domain method for large deformation frictional contact problems. Part 2: Numerical aspects". In: *Computer Methods in Applied Mechanics and Engineering* 198 (33), pp. 2607–2631.

HENKEL (2017). *Thermal Interface Materials - Selection Guide.*

Hertz, H. (1882). "Ueber die Berührung fester elastischer Körper". In: *Journal fur die Reine und Angewandte Mathematik* 1882 (92), pp. 156–171.

Hill, I. D. (1976). "Algorithm AS 100: Normal-Johnson and Johnson-Normal Transformations". In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 25 (2), pp. 190–192.

Hill, I. D., Hill, R., and Holder, R. L. (1976). "Algorithm AS 99: Fitting Johnson Curves by Moments". In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 25 (2), pp. 180–189.

Hill, I. D. and Wheeler, E. (1981). "Remark AS R33: A Remark on Algorithms AS 99: Fitting Johnson Curves by Moments and AS 100: Normal-Johnson and Johnson-Normal Transformations". In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 30 (1), pp. 105–105.

Hill, R. (1963). "Elastic properties of reinforced solids: Some theoretical principles". In: *Journal of the Mechanics and Physics of Solids* 11 (5), pp. 357–372.

Holzapfel, G. A. (2000). *Nonlinear Solid Mechanics: A Continuum Approach for Engineering.* Wiley. 480 pp.

Hu, Y. Z. and Tonder, K. (1992). "Simulation of 3-D random rough surface by 2-D digital filter and fourier analysis". In: *International Journal of Machine Tools and Manufacture.* Proceedings of the 5th International Conference on Metrology and Properties of Engineering Surfaces 32 (1), pp. 83–90.

Hüeber, S. and Wohlmuth, B. I. (2005). "A primal–dual active set strategy for non-linear multibody contact problems". In: *Computer Methods in Applied Mechanics and Engineering* 194 (27), pp. 3147–3166.

Hüeber, S. (2008). "Discretization techniques and efficient algorithms for contact problems". PhD thesis. Universität Stuttgart.

Hughes, T. J. R. *et al.* (1976). "A finite element method for a class of contact-impact problems". In: *Computer Methods in Applied Mechanics and Engineering* 8 (3), pp. 249–276.

Hyun, S., Pei, L., *et al.* (2004). "Finite-element analysis of contact between elastic self-affine surfaces". In: *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics* 70 (2), p. 026117.

Hyun, S. and Robbins, M. O. (2007). “Elastic contact between rough surfaces: Effect of roughness at large and small wavelengths”. In: *Tribology International.* Tribology at the Interface: Proceedings of the 33rd Leeds-Lyon Symposium on Tribology (Leeds, 2006) 40 (10), pp. 1413–1422.

Iordanoff, I., Seve, B., and Berthier, Y. (2002). “Solid Third Body Analysis Using a Discrete Approach: Influence of Adhesion and Particle Size on Macroscopic Properties”. In: *Journal of Tribology* 124 (3), pp. 530–538.

ISO 25178 (2016). *Geometrical Product Specifications (GPS)—Surface texture: Areal.*

ISO 4287 (1997). *Geometrical Product Specifications (GPS)—Surface texture: Profile method.*

Israelachvili, J. N. (2015). *Intermolecular and Surface Forces.* Academic Press. 706 pp.

Jackson, R. L. and Green, I. (2011). “On the Modeling of Elastic Contact between Rough Surfaces”. In: *Tribology Transactions* 54 (2), pp. 300–314.

Jacobs, T. D. B., Junge, T., and Pastewka, L. (2017). “Quantitative characterization of surface topography using spectral analysis”. In: *Surface Topography: Metrology and Properties* 5 (1), p. 013001.

Johnson, K. L. (1987). *Contact Mechanics.* Cambridge University Press. 472 pp.

Johnson, N. L. (1949). “Systems of Frequency Curves Generated by Methods of Translation”. In: *Biometrika* 36 (1), pp. 149–176.

Jost, H. P. (1966). *Lubrication: Tribology; Education and Research; Report on the Present Position and Industry’s Needs (submitted to the Department of Education and Science by the Lubrication Engineering and Research) Working Group.* HM Stationery Office.

Kikuchi, N. and Oden, J. T. (1988). *Contact problems in elasticity: a study of variational inequalities and finite element methods.* SIAM. 516 pp.

Kogut, L. and Jackson, R. L. (2005). “A Comparison of Contact Modeling Utilizing Statistical and Fractal Approaches”. In: *Journal of Tribology* 128 (1), pp. 213–217.

Kreyszig, E. (2010). *Advanced Engineering Mathematics.* John Wiley & Sons. 267 pp.

Laursen, T. A. and Simo, J. C. (1993). “A continuum-based finite element formulation for the implicit solution of multibody, large deformation-frictional contact problems”. In: *International Journal for Numerical Methods in Engineering* 36 (20), pp. 3451–3485.

Liao, D. *et al.* (2018). “An improved rough surface modeling method based on linear transformation technique”. In: *Tribology International* 119, pp. 786–794.

Longuet-Higgins, M. S. (1957a). “Statistical Properties of an Isotropic Rough Surface”. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 250, pp. 157–174.

Longuet-Higgins, M. S. (1957b). “The Statistical Analysis of a Random, Moving Surface”. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 249 (966), pp. 321–387.

Lu, W.-S. (1992). *Two-Dimensional Digital Filters.* CRC Press. 420 pp.

Madhusudana, C. V. (2014). *Thermal Contact Conductance.* 2nd ed. Mechanical Engineering Series. Springer International Publishing.

Mainsah, E., Greenwood, J. A., and Chetwynd, D. G. (2013). *Metrology and Properties of Engineering Surfaces.* Springer Science & Business Media. 470 pp.

Majumdar, A. and Bhushan, B. (1991). "Fractal Model of Elastic-Plastic Contact Between Rough Surfaces". In: *Journal of Tribology* 113 (1), pp. 1–11.

Majumdar, A. and Tien, C. L. (1990). "Fractal characterization and simulation of rough surfaces". In: *Wear* 136 (2), pp. 313–327.

Mandelbrot, B. (1967). "How Long Is the Coast of Britain? Statistical Self-Similarity and Fractional Dimension". In: *Science* 156 (3775), pp. 636–638.

Mandelbrot, B. B. (1983). *The Fractal Geometry of Nature.* Henry Holt and Company. 504 pp.

Mandelbrot, B. B., Passoja, D. E., and Paullay, A. J. (1984). "Fractal character of fracture surfaces of metals". In: *Nature* 308 (5961), p. 721.

Manners, W. and Greenwood, J. A. (2006). "Some observations on Persson's diffusion theory of elastic contact". In: *Wear* 261 (5), pp. 600–610.

McCool, J. I. (1978). "Characterization of surface anisotropy". In: *Wear* 49 (1), pp. 19–31.

McCool, J. I. (1986). "Comparison of models for the contact of rough surfaces". In: *Wear* 107 (1), pp. 37–60.

Meirovitch, L. (2001). *Fundamentals of Vibrations.* McGraw-Hill. 806 pp.

Minet, C. *et al.* (2010). "Analysis and Modeling of the Topography of Mechanical Seal Faces". In: *Tribology Transactions* 53 (6), pp. 799–815.

Misbah, C., Pierre-Louis, O., and Saito, Y. (2010). "Crystal surfaces in and out of equilibrium: A modern view". In: *Reviews of Modern Physics* 82 (1), pp. 981–1040.

Mummery, L. (1992). *Surface Texture Analysis: The Handbook.* Hommelwerke GmbH. 106 pp.

Nayak, P. R. (1971). "Random Process Model of Rough Surfaces". In: *Journal of Lubrication Technology* 93 (3), p. 398.

Nayak, P. R. (1973). "Some aspects of surface roughness measurement". In: *Wear* 26 (2), pp. 165–174.

Newland, D. E. (1984). *An introduction to random vibrations and spectral analysis.* Longman. 414 pp.

Nitsche, R. (2011). "A multiscale projection method for contact on rough surfaces". PhD thesis. Hannover: Institut für Kontinuumsmechanik. 153 pp.

Oden, P. I. *et al.* (1992). “AFM Imaging, Roughness Analysis and Contact Mechanics of Magnetic Tape and Head Surfaces”. In: *Journal of Tribology* 114 (4), pp. 666–674.

Oliver, J. *et al.* (2009). “A contact domain method for large deformation frictional contact problems. Part 1: Theoretical basis”. In: *Computer Methods in Applied Mechanics and Engineering* 198 (33), pp. 2591–2606.

Orfanidis, S. J. (1996). *Introduction to Signal Processing*. Prentice Hall. 824 pp.

Paggi, M. and Ciavarella, M. (2010). “The coefficient of proportionality {κ} between real contact area and load, with new asperity models”. In: *Wear* 268 (7), pp. 1020–1029.

Panda, S. *et al.* (2016). “Spectral Approach on Multiscale Roughness Characterization of Nominally Rough Surfaces”. In: *Journal of Tribology* 139 (3), pp. 031402–031402–10.

Papadopoulos, P. and Taylor, R. L. (1992). “A mixed formulation for the finite element solution of contact problems”. In: *Computer Methods in Applied Mechanics and Engineering* 94 (3), pp. 373–389.

Pastewka, L. and Robbins, M. O. (2014). “Contact between rough surfaces and a criterion for macroscopic adhesion”. In: *Proceedings of the National Academy of Sciences* 111 (9), pp. 3298–3303.

Patir, N. (1978). “A numerical procedure for random generation of rough surfaces”. In: *Wear* 47 (2), pp. 263–277.

Patrikar, R. M. (2004). “Modeling and simulation of surface roughness”. In: *Applied Surface Science* 228 (1), pp. 213–220.

Pawlus, P. (2008). “Simulation of stratified surface topographies”. In: *Wear*. 10th International Conference on Metrology and Properties of Engineering Surfaces 264 (5), pp. 457–463.

Pei, L. *et al.* (2005). “Finite element modeling of elasto-plastic contact between rough surfaces”. In: *Journal of the Mechanics and Physics of Solids* 53 (11), pp. 2385–2409.

Peitgen, H.-O. and Saupe, D. (2012). *The Science of Fractal Images*. Springer Science & Business Media. 328 pp.

Persson, B. N. J. (2001a). “Elastoplastic Contact between Randomly Rough Surfaces”. In: *Physical Review Letters* 87 (11), p. 116101.

Persson, B. N. J. (2001b). “Theory of rubber friction and contact mechanics”. In: *The Journal of Chemical Physics* 115 (8), pp. 3840–3861.

Persson, B. N. J. (2006). “Contact mechanics for randomly rough surfaces”. In: *Surface Science Reports* 61 (4), pp. 201–227.

Persson, B. N. J. (2007). “Relation between interfacial separation and load: a general theory of contact mechanics”. In: *Physical Review Letters* 99 (12), p. 125502.

Persson, B. N. J. (2014). “On the Fractal Dimension of Rough Surfaces”. In: *Tribology Letters* 54 (1), pp. 99–106.

Persson, B. N. J., Albohr, O., *et al.* (2005). “On the nature of surface roughness with application to contact mechanics, sealing, rubber friction and adhesion”. In: *Journal of Physics: Condensed Matter* 17 (1), R1.

Persson, B. N. J., Bucher, F., and Chiaia, B. (2002). “Elastic contact between randomly rough surfaces: Comparison of theory with numerical results”. In: *Physical Review B* 65 (18), p. 184106.

Persson, B. N. J., Sivebaek, I. M., *et al.* (2008). “On the origin of Amonton’s friction law”. In: *Journal of Physics: Condensed Matter* 20 (39), p. 395006.

Pinto Carvalho, R. (2018). *Finite deformation contact modelling at different scales.* PhD Seminar. Faculdade de Engenharia da Universidade do Porto.

Popova, E. and Popov, V. L. (2015). “The research works of Coulomb and Amontons and generalized laws of friction”. In: *Friction* 3 (2), pp. 183–190.

Popp, A. (2012). “Mortar Methods for Computational Contact Mechanics and General Interface Problems”. PhD thesis. Technische Universität München.

Popp, A., Gee, M. W., and Wall, W. A. (2009). “A finite deformation mortar contact formulation using a primal–dual active set strategy”. In: *International Journal for Numerical Methods in Engineering* 79 (11), pp. 1354–1391.

Popp, A., Gitterle, M., *et al.* (2010). “A dual mortar approach for 3D finite deformation contact with consistent linearization”. In: *International Journal for Numerical Methods in Engineering* 83 (11), pp. 1428–1465.

Proakis, J. G. and Manolakis, D. G. (2007). *Digital Signal Processing.* Pearson Prentice Hall. 1112 pp.

Proakis, J. G. and Salehi, M. (2002). *Communication Systems Engineering.* Prentice Hall. 801 pp.

Prodanov, N., Dapp, W. B., and Müser, M. H. (2014). “On the Contact Area and Mean Gap of Rough, Elastic Contacts: Dimensional Analysis, Numerical Corrections, and Reference Data”. In: *Tribology Letters* 53 (2), pp. 433–448.

Puso, M. (2004). “A mortar segment-to-segment contact method for large deformation solid mechanics”. In: *Computer Methods in Applied Mechanics and Engineering* 193, pp. 601–629.

Puso, M. A. and Laursen, T. A. (2002). “A 3D contact smoothing method using Gregory patches”. In: *International Journal for Numerical Methods in Engineering* 54 (8), pp. 1161–1194.

Rabinowicz, E. (1965). *Friction and wear of materials.* Wiley. 264 pp.

Rao, K. R., Kim, D. N., and Hwang, J. J. (2011). *Fast Fourier Transform - Algorithms and Applications.* Springer Science & Business Media. 437 pp.

Reinelt, J. (2009). "Frictional Contact of Elastomer Materials on Rough Rigid Surfaces". PhD thesis. Inst. für Kontinuumsmechanik. 121 pp.

Reinelt, J. and Wriggers, P. (2010). "Multi-scale Approach for Frictional Contact of Elastomers on Rough Rigid Surfaces". In: *Elastomere Friction: Theory, Experiment and Simulation.* Ed. by D. Besdo *et al.* Lecture Notes in Applied and Computational Mechanics. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 53–94.

Reizer, R. (2011). "Simulation of 3D Gaussian surface topography". In: *Wear.* The 12th International Conference on Metrology and Properties of Engineering Surfaces 271 (3), pp. 539–543.

Russ, J. C. (1994). *Fractal Surfaces.* Springer US.

Sayles, R. S. and Thomas, T. R. (1976). "Thermal conductance of a rough elastic contact". In: *Applied Energy* 2 (4), pp. 249–267.

Sayles, R. S. and Thomas, T. R. (1978). "Surface topography as a nonstationary random process". In: *Nature* 271 (5644), pp. 431–434.

SEMI MF1811 (2010). *Guide for Estimating the Power Spectral Density Function and Related Finish Parameters from Surface Profile Data.*

Shohat, J. (1929). "Inequalities for Moments of Frequency Functions and for Various Statistical Constants". In: *Biometrika* 21 (1), pp. 361–375.

Simo, J. C., Wriggers, P., and Taylor, R. L. (1985). "A perturbed Lagrangian formulation for the finite element solution of contact problems". In: *Computer Methods in Applied Mechanics and Engineering* 50 (2), pp. 163–180.

Sokolnikoff, I. S. (1951). *Tensor analysis: theory and applications.* Wiley. 360 pp.

Souza Neto, E. A. de *et al.* (1996). "Design of simple low order finite elements for large strain analysis of nearly incompressible solids". In: *International Journal of Solids and Structures* 33 (20), pp. 3277–3296.

Stanley, H. M. and Kato, T. (1997). "An FFT-Based Method for Rough Surface Contact". In: *Journal of Tribology* 119 (3), pp. 481–485.

Staufert, G. (1979). "Characterization of Random Roughness Profiles-A Comparison of AR-Modeling Technique and Profile Description by Means of Commonly Used Parameters". In: *CIRP Annals* 28 (1), pp. 431–435.

Stupkiewicz, S. (2007). *Micromechanics of Contact and Interphase Layers.* Lecture Notes in Applied and Computational Mechanics. Berlin Heidelberg: Springer-Verlag.

Suh, J., Dilon, R. P., and Tseng, S. (2015). *Thermal Interface Materials Selection and ApplicationGuidelines: In Perspective of Xilinx Virtex-5QV Thermal Management.* Pasadena, California: Jet Propulsion Laboratory.

Temizer, İ. (2011). “Thermomechanical contact homogenization with random rough surfaces and microscopic contact resistance”. In: *Tribology International* 44 (2), pp. 114–124.

Temizer, İ. and Wriggers, P. (2008). “A multiscale contact homogenization technique for the modeling of third bodies in the contact interface”. In: *Computer Methods in Applied Mechanics and Engineering* 198 (3), pp. 377–396.

Thomas, T. R. (1999). *Rough Surfaces.* Imperial College Press. 300 pp.

Tworzydlo, W. W. *et al.* (1998). “Computational micro- and macroscopic models of contact and friction: formulation, approach and applications”. In: *Wear* 220 (2), pp. 113–140.

Tzanakis, I. *et al.* (2012). “Future perspectives on sustainable tribology”. In: *Renewable and Sustainable Energy Reviews* 16 (6), pp. 4126–4140.

Urzică, A. C., Bălan, M. R. D., and Creţu, S. S. (2012). “Pressures Distributions and Depth Stresses Developed in Concentrated Contacts Between Elements With Non-Gaussian Rough Surfaces”. In: pp. 547–554.

Vakis, A. I. *et al.* (2018). “Modeling and simulation in tribology across scales: An overview”. In: *Tribology International* 125, pp. 169–199.

Vallet, C. *et al.* (2009). “Real versus synthesized fractal surfaces: Contact mechanics and transport properties”. In: *Tribology International* 42 (2), pp. 250–259.

Vieira, R. (2018). *Numerical Micromechanical Analysis on the Influence of Monocrystalline Parameters on the Elastic and Yielding Response of Polycristalline Aggregates.* M.Sc. Thesis. Porto: Faculdade de Engenharia da Universidade do Porto.

Wagner, P. (2018). “A Multiscale FEM Approach for Rubber Friction on Rough Surfaces”. PhD thesis. Institut für Kontinuumsmechanik, Gottfried Wilhelm Leibniz Universität Hannover. book.

Wagner, P., Wriggers, P., Klapproth, C., *et al.* (2015). “Multiscale FEM approach for hysteresis friction of rubber on rough surfaces”. In: *Computer Methods in Applied Mechanics and Engineering* 296, pp. 150–168.

Wagner, P., Wriggers, P., Veltmaat, L., *et al.* (2017). “Numerical multiscale modelling and experimental validation of low speed rubber friction on rough road surfaces including hysteretic and adhesive effects”. In: *Tribology International* 111, pp. 243–253.

Watson, W. and Spedding, T. (1982). “The time series modelling of non-gaussian engineering processes”. In: *Wear* 83 (2), pp. 215–231.

Whitehouse David J., Phillips M. J., and Tabor David (1978). “Discrete properties of random surfaces”. In: *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences* 290 (1369), pp. 267–298.

Whitehouse David J., Phillips M. J., and Tabor David (1982). “Two-dimensional properties of random surfaces”. In: *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences* 305 (1490), pp. 441–468.

Whitehouse, D. J. (1978). “The Digital Measurement of Peak Parameters on Surface Profiles”. In: *Journal of Mechanical Engineering Science* 20 (4), pp. 221–227.

Whitehouse, D. J. (1982). “The parameter rash — is there a cure?” In: *Wear* 83 (1), pp. 75–78.

Whitehouse, D. J. (1983). “The Generation of Two Dimensional Random Surfaces Having a Specified Function”. In: *CIRP Annals* 32 (1), pp. 495–498.

Whitehouse, D. J. and Archard, J. F. (1970). “The Properties of Random Surfaces of Significance in their Contact”. In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 316 (1524), pp. 97–121.

Whitehouse, D. J. (1994). *Handbook of Surface Metrology.* CRC Press. 1048 pp.

Wohlmuth, B. (2000). “A Mortar Finite Element Method Using Dual Spaces for the Lagrange Multiplier”. In: *SIAM Journal on Numerical Analysis* 38 (3), pp. 989–1012.

Wriggers, P. (2006). *Computational Contact Mechanics.* 2nd ed. Berlin Heidelberg: Springer-Verlag.

Wriggers, P. and Laursen, T. A. (2008). *Computational Contact Mechanics.* Springer. 252 pp.

Wriggers, P. and Reinelt, J. (2009). “Multi-scale approach for frictional contact of elastomers on rough rigid surfaces”. In: *Computer Methods in Applied Mechanics and Engineering.* Advances in Simulation-Based Engineering Sciences – Honoring J. Tinsley Oden 198 (21), pp. 1996–2008.

Wu, J.-J. (2000a). “Characterization of fractal surfaces”. In: *Wear* 239 (1), pp. 36–47.

Wu, J.-J. (2000b). “Simulation of rough surfaces with FFT”. In: *Tribology International* 33 (1), pp. 47–58.

Wu, J.-J. (2002). “Analyses and simulation of anisotropic fractal surfaces”. In: *Chaos, Solitons & Fractals* 13 (9), pp. 1791–1806.

Wu, J.-J. (2004). “Simulation of non-Gaussian surfaces with FFT”. In: *Tribology International* 37 (4), pp. 339–346.

Yang, C. and Persson, B. N. J. (2008). “Contact mechanics: contact area and interfacial separation from small contact to full contact”. In: *Journal of Physics: Condensed Matter* 20 (21), p. 215214.

Yastrebov, V. A. (2011). “Numerical Methods in Contact Mechanics”. PhD thesis. Centre des Matériaux, MINES ParisTech. 277 pp.

Yastrebov, V. A., Anciaux, G., and Molinari, J.-F. (2012). “Contact between representative rough surfaces”. In: *Physical Review E* 86 (3), p. 035601.

Yastrebov, V. A., Anciaux, G., and Molinari, J.-F. (2015). “From infinitesimal to full contact between rough surfaces: Evolution of the contact area”. In: *International Journal of Solids and Structures* 52, pp. 83–102.

Yastrebov, V. A., Anciaux, G., and Molinari, J.-F. (2017). “The role of the roughness spectral breadth in elastic contact of rough surfaces”. In: *Journal of the Mechanics and Physics of Solids* 107, pp. 469–493.

Yastrebov, V. A., Durand, J., *et al.* (2011). “Rough surface contact analysis by means of the Finite Element Method and of a new reduced model”. In: *Comptes Rendus Mécanique*. Surface mechanics : facts and numerical models 339 (7), pp. 473–490.

Zahouani, H., Vargiolu, R., and Loubet, J. .-.-L. (1998). “Fractal models of surface topography and contact mechanics”. In: *Mathematical and Computer Modelling*. Recent Advances in Contact Mechanics 28 (4), pp. 517–534.

Zavarise, G. and Paggi, M. (2007). “Reliability of Micromechanical Contact Models: a Still Open Issue”. In: *Computational Contact Mechanics*. Ed. by P. Wriggers and T. A. Laursen. CISM International Centre for Mechanical Sciences. Vienna: Springer Vienna, pp. 39–82.

Zhang, S., Wang, W., and Zhao, Z. (2014). “The effect of surface roughness characteristics on the elastic–plastic contact performance”. In: *Tribology International* 79, pp. 59–73.

Zienkiewicz, O. C., Taylor, R. L., and Taylor, R. L. (2000a). *The Finite Element Method: The basis*. Butterworth-Heinemann. 724 pp.

Zienkiewicz, O. C., Taylor, R. L., and Taylor, R. L. (2000b). *The Finite Element Method: Solid mechanics*. Butterworth-Heinemann. 482 pp.

Zou, M. *et al.* (2007). “A Monte Carlo method for simulating fractal surfaces”. In: *Physica A: Statistical Mechanics and its Applications* 386 (1), pp. 176–186.