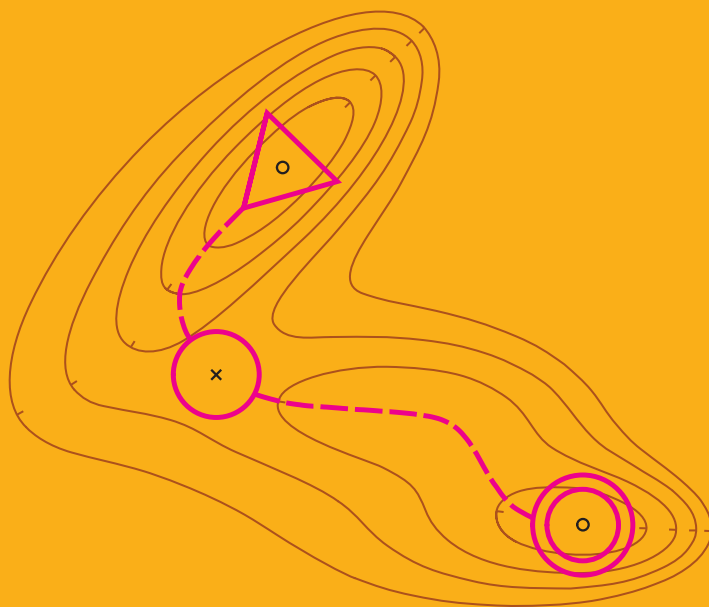


Algorithms for Finding Saddle Points and Minimum Energy Paths Using Gaussian Process Regression

Olli-Pekka Koistinen



Algorithms for Finding Saddle Points and Minimum Energy Paths Using Gaussian Process Regression

Olli-Pekka Koistinen

A doctoral dissertation completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Science, at a public examination held at the lecture hall E (Y124) of the school on 9 January 2020 at 12.

This doctoral dissertation has been conducted under a convention for the joint supervision at Aalto University (Finland) and University of Iceland (Iceland).

**Aalto University
School of Science
Department of Computer Science
University of Iceland**

Supervising professors

Professor Aki Vehtari, Aalto University, Finland

Professor Hannes Jónsson, University of Iceland, Iceland

Preliminary examiners

Professor Johannes Kästner, University of Stuttgart, Germany

Professor Andrew Peterson, Brown University, United States

Opponent

Professor Thomas Bligaard, Technical University of Denmark, Denmark

Aalto University publication series

DOCTORAL DISSERTATIONS 222/2019

© 2019 Olli-Pekka Koistinen

ISBN 978-952-60-8850-1 (printed)

ISBN 978-952-60-8851-8 (pdf)

ISSN 1799-4934 (printed)

ISSN 1799-4942 (pdf)

<http://urn.fi/URN:ISBN:978-952-60-8851-8>

Unigrafia Oy

Helsinki 2019

Finland



Author

Olli-Pekka Koistinen

Name of the doctoral dissertation

Algorithms for Finding Saddle Points and Minimum Energy Paths Using Gaussian Process Regression

Publisher School of Science

Unit Department of Computer Science

Series Aalto University publication series DOCTORAL DISSERTATIONS 222/2019

Field of research Computational science

Manuscript submitted 23 October 2019

Date of the defence 9 January 2020

Permission for public defence granted (date) 26 November 2019

Language English

Monograph

Article dissertation

Essay dissertation

Abstract

Chemical reactions and other transitions involving rearrangements of atoms can be studied theoretically by analyzing a potential energy surface defined in a high-dimensional space of atom coordinates. Local minimum points of the energy surface correspond to stable states of the system, and minimum energy paths connecting these states characterize mechanisms of possible transitions. Of particular interest is often the maximum point of the minimum energy path, which is located at a first-order saddle point of the energy surface and can be used to estimate the activation energy and rate of the particular transition.

Minimum energy paths and saddle points between two known states have been traditionally searched with iterative methods where a chain of discrete points of the coordinate space is moved and stretched towards a minimum energy path according to imaginary forces based on gradient vectors of the potential energy surface. The actual saddle point can be found by reversing the component of the gradient vector parallel to the path at one of the points of the chain and letting this point climb along the path towards the saddle point. If the end state of the transition is unknown, the saddle point can be searched correspondingly by rotating a pair of closely spaced points towards the orientation of the lowest curvature, reversing the gradient component corresponding to this direction, and moving the pair towards the saddle point. These methods may, however, require hundreds of iterations, and since accurate evaluation of the gradient vector is often computationally expensive, the information obtained from previous iterations should be utilized as efficiently as possible to decrease the number of iterations. Using statistical models, an approximation to the energy surface can be constructed, and a minimum energy path or a saddle point can be searched on the approximate surface. The accuracy of the solution can be checked with further evaluations, which can be then used to update the model for following iterations.

In this dissertation, machine learning algorithms based on Gaussian process regression are developed to enhance searches of minimum energy paths and saddle points. Gaussian process models serve here as flexible prior probability models for potential energy surfaces. Observed values of both energy and its derivatives can be used to update the model, and the posterior predictive distribution obtained as a result of Bayesian inference provides also an uncertainty estimate, which can be utilized when selecting new observation points. Separate methods are presented both for finding a minimum energy path between two known states and a saddle point located in the vicinity of a given start point. Based on simple test examples, the methods utilizing Gaussian processes may reduce the number of evaluations to a fraction of what is required by conventional methods.

Keywords saddle point, minimum energy path, Gaussian process, machine learning

ISBN (printed) 978-952-60-8850-1

ISBN (pdf) 978-952-60-8851-8

ISSN (printed) 1799-4934

ISSN (pdf) 1799-4942

Location of publisher Helsinki

Location of printing Helsinki **Year** 2019

Pages 144

urn <http://urn.fi/URN:ISBN:978-952-60-8851-8>

Tekijä

Olli-Pekka Koistinen

Väitöskirjan nimi

Gaussisia prosesseja hyödyntäviä menetelmiä satulapisteiden ja minimienergiapolkujen etsintään

Julkaisija Perustieteiden korkeakoulu**Yksikkö** Tietotekniikan laitos**Sarja** Aalto University publication series DOCTORAL DISSERTATIONS 222/2019**Tutkimusala** Laskennallinen tiede**Käsikirjoituksen pvm** 23.10.2019**Väitöspäivä** 09.01.2020**Väittelyluvan myöntämispäivä** 26.11.2019**Kieli** Englanti **Monografia** **Artikkeliväitöskirja** **Esseeväitöskirja****Tiivistelmä**

Kemiallisia reaktioita ja muita atomien liikkeisiin perustuvia tapahtumia voidaan tarkastella teoreettisesti atomien koordinaattien muodostamassa moniulotteisessa avaruudessa määritellyn potentiaalienergiapinnan avulla. Energiapinnan paikalliset minimikohdat vastaavat systeemin vakaita tiloja, ja näitä tiloja yhdistävät minimienergiapolut kuvaavat mahdollisia reaktiomekanismeja. Eriytyisen mielenkiinnon kohteena on usein minimienergiapolun globaali maksimikohta, joka sijaitsee potentiaalienergiapinnan satulapisteessä ja jonka avulla voidaan arvioida kyseisen reaktion aktivoitumisenergiaa ja reaktionopeutta.

Kahden tunnetun tilan välisiä minimienergiapolkuja ja satulapisteitä on perinteisesti etsitty iteraatiivisilla menetelmillä, joissa erillisistä koordinaattivaruuden pisteistä muodostuvaa ketjua liikutetaan ja venytetään kohti minimienergiapolkua potentiaalienergiapinnan gradienttivektorien avulla laskettujen kuvitteellisten voimien perusteella. Varsinainen satulapiste voidaan määrittää kääntämällä gradienttivektorin polun suuntainen komponentti yhdessä ketjun pisteistä, jonka annetaan nousta polun suuntaisesti kohti satulapistettä. Jos reaktion lopputila on tuntematon, voidaan satulapistettä etsiä vastaavasti kiertämällä kahden lähekkäisen pisteen muodostamaa paria potentiaalienergiapinnan pienimmän kaarevuuden suuntaiseksi, kääntämällä tätä suuntaa vastaava gradienttikomponentti, ja liikuttamalla pisteparia kohti satulapistettä. Nämä menetelmät voivat kuitenkin vaatia satoja iteraatioita, ja koska gradienttivektorin tarkka määrittäminen on usein laskennallisesti raskasta, tulisi aikaisemmista iteraatioista saatu informaatio hyödyntää mahdollisimman tehokkaasti iteraatioiden vähentämiseksi. Tilastollisten mallien avulla energiapinnalle voidaan muodostaa likimääräinen arvio, ja minimienergiapolkua tai satulapistettä voidaan etsiä likimääräiseltä pinnalta. Ratkaisu voidaan tarkistaa uusien tarkkojen havaintojen avulla, joita voidaan puolestaan käyttää mallin tarkentamiseksi mahdollisia seuraavia iteraatioita varten.

Tässä väitöskirjassa kehitetään gaussisiin prosesseihin perustuvia koneoppimisalgoritmeja minimienergiapolkujen ja satulapisteiden etsinnän nopeuttamiseksi. Gaussiset prosessit toimivat tässä tapauksessa joustavina prioritodennäköisyyksille potentiaalienergiapinnoille. Mallin päivittämiseksi voidaan käyttää sekä energian että gradienttikomponenttien havaittuja arvoja, ja Bayes-päätelyn tuloksena saatava ennustejakauma sisältää myös epävarmuusarvion, jota voidaan käyttää hyväksi uusien havaintopisteiden valinnassa. Väitöskirjassa esitetään menetelmät sekä kahden tunnetun tilan välisen minimienergiapolun määrittämiseen että annetun aloituspisteen lähistöllä sijaitsevan satulapisteen etsimiseen. Yksinkertaisten testiesimerkkien perusteella gaussisia prosesseja hyödyntävät menetelmät voivat vähentää tarkkojen havaintojen määrän murto-osaan perinteisten menetelmien vaatimista havaintomääristä.

Avainsanat satulapiste, minimienergiapolku, gaussinen prosessi, koneoppiminen**ISBN (painettu)** 978-952-60-8850-1**ISBN (pdf)** 978-952-60-8851-8**ISSN (painettu)** 1799-4934**ISSN (pdf)** 1799-4942**Julkaisupaikka** Helsinki**Painopaikka** Helsinki**Vuosi** 2019**Sivumäärä** 144**urn** <http://urn.fi/URN:ISBN:978-952-60-8851-8>

Útdráttur

Eiginleika efnahvarfa og annarra umraðana atóma er hægt að kanna með því að skoða orkuyfirborðið, skilgreint sem orka kerfisins sem fall af atómhnitunum. Staðbundin lágmark á orkuyfirborðinu samsvara ástöndum sem kerfið getur verið í og lágmarksorkuferlar milli þeirra einkenna gang mögulegra umraðana atómanna. Hámark á lágmarksorkuferlum eru sérlega mikilvæg. Þau samsvara fyrsta stigs söðulpunktum og gefa mat á virkjunarorku og þar með hraða samsvarandi umröðunar.

Lágmarksorkuferlar og söðulpunktar milli tveggja þekktra ástanda eru gjarnan fundnir með ítrekunaraðferðum þar sem röð af ímyndum af kerfinu mynda feril milli endapunktanna og eru færðar til þangað til þær liggja á lágmarksorkuferlinum. Færslan í hverri ítrekun er fundin út frá stiglinum á orkuyfirborðinu. Söðulpunktinn er hægt að finna með því að færa orkuhæstu ímyndina í átt stigilsins eftir að þætti hans í stefnu ferilsins hefur verið snúið við. Þannig færast sú ímynd upp í orku að söðulpunktinum. Ef lokaástand umröðuninnar er ekki þekkt er hægt að finna fyrsta stigs söðulpunkt með því að nota þar ímynda af kerfinu sem eru þétt saman og myndar tvennu. Henni er snúið til að finna stefnuna með lægstan krappa á orkuyfirborðinu og síðan færð í átt stigulsins eftir að þátturinn í stefnu lægsta krappans hefur verið speglaður. Þannig færast tvennan að söðulpunktinum. Þessi aðferð getur þurft hundruða ítrekana og þar eð útreikningar á orkustiglinum eru oft þungir er mikilvægt að nýta upplýsingar úr fyrri ítrekunum eins vel og hægt er til að fækka ítrekunum. Með því að nota tölfræðileg líkön er hægt að búa til nálgun fyrir orkuyfirborðið og leita að söðulpunktinum á því yfirborði. Lausnina er hægt að sanreyna með frekari útreikningum á orkustiglinum sem síðan er hægt að nota til að bæta nálgunina fyrir næstu færslur tvennunnar.

Í þessari ritgerð er vélrænn lærdómur sem byggður er á Gaussferlaaðhvarfi notaður til að hraða reikningum á lágmarksorkuferlum og söðulpunktum. Líkön fyrir orkuyfirborðið eru búin til með út frá þekktum gildum á orkunni og stiglinum með tölfræðilegum aðferðum Bayes og mat fundið á óvissunni í líkaninu sem hægt er að nýta til að ákveða hvaða punkt er best að reikna í næstu ítrekun. Mismunandi aðferðir eru þróaðar bæði til að finna lágmarksorkuferla milli tveggja þekktra ástanda og til að finna söðulpunkt í nágrenni gefins upphafspunkts. Reikningar á ýmsum mismunandi kerfum sýna að með þessu móti er hægt að fækka útreikningum á orkunni og stiglinum mjög verulega í samanburði við þær aðferðir sem nú eru notaðar.

Contents

Preface	xi
List of publications	xiii
Author's contribution	xv
Abbreviations	xvii
Symbols	xix
1. Introduction	1
2. Gaussian processes	5
2.1 Gaussian process model	5
2.2 Covariance functions	6
2.3 Gaussian process regression	8
2.4 Regression with derivatives	11
2.5 Gaussian process models for potential energy surfaces . .	13
3. Methods for finding saddle points	17
3.1 Nudged elastic band method	18
3.2 Dimer method	21
4. Summary of contributions	25
4.1 GP-NEB algorithm	25
4.2 GP-dimer algorithm	31
5. Discussion	37
Bibliography	39
Errata	47
Publications	49

Preface

The journey towards this dissertation begun already in 2011 when I applied for a summer job at the Department of Biomedical Engineering and Computational Science (BECS) and found myself in the Bayesian methodology group, developing analysis methods for brain research. After finishing my master's thesis and pondering my future for a while, I eventually decided to continue on the track towards a doctoral degree. Along the journey, BECS became NBE, the Department of Neuroscience and Biomedical Engineering, and our group moved to the Department of Computer Science to be part of the current probabilistic machine learning group. These organizational changes naturally weakened our connection to neuroscience and spread the method development to a broader range of applications. After several, more or less successful trial projects, the final topic for my dissertation was quite unexpectedly found from the field of theoretical chemistry. This connection opened an interesting opportunity for a double degree via a joint supervision agreement between Aalto University and University of Iceland. During the time of the shared supervision, I was partly employed by the Department of Applied Physics. In addition to the employer departments, I gratefully acknowledge the financial support of the Academy of Finland and the Finnish Cultural Foundation (Kari Kairamo Fund) as well as the support of the Icelandic Research Fund to partly cover the expenses of my visits to Iceland.

Although research is at times lonely work inside one's own head, it is above all collaboration and learning from others. The people that I have learned of the most about science are my two supervisors, Prof. Aki Vehtari and Prof. Hannes Jónsson. As the leader of the former Bayesian methodology group, Aki has been supporting my work from the beginning and, by opening his bottomless storage of ideas, taken care that his students are never left empty-handed. On the other hand, I thank Aki also for the freedom he has given to develop the ideas further and patience when waiting for results. Hannes joined the journey in 2016 after he had met Aki at a conference and recognized a possibility for fruitful collaboration. That meeting turned out to be a good fortune to me as I got a well-defined

goal for my theretofore unstructured doctoral research. I thank Hannes for warmly welcoming me to Iceland, introducing me to the necessary chemical details, and motivating me to find ways to tackle the methodological challenges. I want to express my gratitude also to Prof. Jouko Lampinen for the supervision in the initial phase of my doctoral studies, Prof. Johannes Kästner and Prof. Andrew Peterson for the pre-examination of this dissertation, and Prof. Thomas Bligaard for accepting the invitation to act as an opponent in the upcoming defence.

In addition to Aki and Hannes, I have the joy to acknowledge three more co-authors who have contributed to the articles included in the dissertation, Dr. Emile Maras, Freyja Dagbjartsdóttir, and Vilhjálmur Ásgeirsson. I thank Emile for introducing me to the details of the nudged elastic band method, Freyja for carefully performing the experiments for the heptamer island benchmark, and especially Villi for the close collaboration during the past three years. During the years at Aalto University, I have had a pleasure to work with many friendly and talented colleagues who have made my days easier through both work-related and more relaxed conversations and comments. I want to thank especially Akash Dhaka, Kunal Ghosh, Dr. Pasi Jylänki, Marko Järvenpää, Dr. Juho Kokkala, Sasu Mäkelä, Topi Paananen, Dr. Tomi Peltola, Dr. Juho Piironen, Prof. Michael Riis Andersen, Gabriel Riutort-Mayol, Prof. Juha Salmitaival, Eero Siivola, Tuomas Sivula, Dr. Dmitry Smirnov, and Prof. Arno Solin. I am grateful also to the service personnel at Aalto for the effort to secure the conditions for our research work.

Beside the studies, I have been lucky to have a passion towards the sport of orienteering. Although finding the balance between two demanding ambitions has been a challenge, orienteering has had an important role in keeping my thoughts away from research when needed. For the same reason, I want to thank all my friends, many of whom I have got to know in the activities of Teekkarisuunnistajat and Hiidenkiertäjät. A special mention is dedicated to Dr. Joonas Pääkkönen and Dr. Rainer Kujala, who have shared the same route choice with me also in studies and opened the way especially on the final legs of the course. Finally, and most importantly, I thank my family for all their support and encouragement. It is comforting to know that there are people you can always lean on, whatever comes about.

Espoo, December 11, 2019,

Olli-Pekka Koistinen

List of publications

This dissertation consists of an overview and of the following publications which are referred to in the text by their Roman numerals.

- I** Olli-Pekka Koistinen, Emile Maras, Aki Vehtari, and Hannes Jónsson. Minimum energy path calculations with Gaussian process regression. *Nanosystems: Physics, Chemistry, Mathematics*, volume 7, issue 6, pages 925–935, December 2016.
- II** Olli-Pekka Koistinen, Freyja B. Dagbjartsdóttir, Vilhjálmur Ásgeirsson, Aki Vehtari, and Hannes Jónsson. Nudged elastic band calculations accelerated with Gaussian process regression. *The Journal of Chemical Physics*, volume 147, issue 15, article 152720, 14 pages, September 2017.
- III** Olli-Pekka Koistinen, Vilhjálmur Ásgeirsson, Aki Vehtari, and Hannes Jónsson. Nudged elastic band calculations accelerated with Gaussian process regression based on inverse interatomic distances. *Journal of Chemical Theory and Computation*, volume 15, issue 12, pages 6738–6751, October 2019.
- IV** Olli-Pekka Koistinen, Vilhjálmur Ásgeirsson, Aki Vehtari, and Hannes Jónsson. Minimum mode saddle point searches using Gaussian process regression with inverse-distance covariance function. Accepted for publication in *Journal of Chemical Theory and Computation*, 20 pages, December 2019.

Author's contribution

Publication I: “Minimum energy path calculations with Gaussian process regression”

The initial idea of using Gaussian process regression with derivative observations in minimum energy path calculations came from Jónsson and Vehtari. Koistinen implemented and developed the GP-NEB algorithm, performed the experiments, and wrote parts of the manuscript describing Gaussian process methodology and details of the algorithm. Jónsson had the main responsibility in writing the manuscript. Vehtari contributed to the development of the algorithm and proposed suggestions to the manuscript regarding Gaussian process methodology. Maras advised in implementation of the NEB method and the experiments.

Publication II: “Nudged elastic band calculations accelerated with Gaussian process regression”

Koistinen innovated, implemented and developed the one-image-evaluated variant of the GP-NEB algorithm, performed initial experiments, and wrote parts of the manuscript describing Gaussian process methodology and details of the algorithm. Jónsson suggested extending the algorithm to climbing image NEB and including Hessian observations and had the main responsibility in writing the manuscript together with Koistinen. Vehtari supported the development of the algorithm and proposed suggestions to the manuscript regarding Gaussian process methodology. Dagbjartsdóttir performed the final experiments together with Ásgeirsson.

Publication III: “Nudged elastic band calculations accelerated with Gaussian process regression based on inverse interatomic distances”

Koistinen innovated, implemented and developed the methodological improvements to the GP-NEB algorithm, performed initial experiments and part of the final experiments, and wrote the manuscript. Jónsson supported the development of the algorithm and revised the manuscript. Vehtari supported the development of the algorithm, suggested various alternative improvements tested by Koistinen, and proposed suggestions to the manuscript regarding Gaussian process methodology. Ásgeirsson implemented the H₂/Cu(110) and H₂O potentials, performed part of the final experiments, and reviewed the manuscript.

Publication IV: “Minimum mode saddle point searches using Gaussian process regression with inverse-distance covariance function”

Koistinen implemented and developed the GP-dimer algorithm, performed the experiments, and wrote majority of the manuscript. Jónsson suggested the topic, supported the development of the algorithm, and wrote parts of the manuscript. Ásgeirsson implemented the potentials and reviewed the manuscript. Vehtari supported the development of the algorithm and reviewed the manuscript.

Abbreviations

AIE	all-images-evaluated
C	carbon
CI-NEB	climbing image nudged elastic band
Cu	copper
FCC	face-centred cubic
FIRE	fast inertial relaxation engine
GP	Gaussian process
GPR	Gaussian process regression
GPU	graphics processing unit
H	hydrogen
L-BFGS	limited-memory Broyden-Fletcher-Goldfarb-Shanno
N	nitrogen
NEB	nudged elastic band
O	oxygen
OIE	one-image-evaluated
S	sulphur
SOAP	smooth overlap of atomic positions
VPO	velocity projection optimization

Symbols

A_f	set of indices for frozen atoms
A_m	set of indices for moving atoms
$B_\nu(\cdot)$	modified Bessel function of the second kind and of order ν
C	curvature estimate
$\text{Cov}[\cdot, \cdot]$	covariance
D	dimension
$\mathcal{D}_x(\cdot, \cdot)$	regular difference measure
$\mathcal{D}_{1/r}(\cdot, \cdot)$	inverse-distance difference measure
$E(\cdot)$	energy
$E[\cdot]$	mean
$f(\cdot)$	latent function
\mathbf{f}	vector of latent function values at observation points
\mathbf{f}^*	vector of latent function values at prediction points
$\mathbf{F}(\cdot)$	atomic force vector (negative energy gradient)
$\mathbf{F}^{\parallel}(\cdot)$	parallel component of an atomic force vector
$\mathbf{F}^{\perp}(\cdot)$	perpendicular component of an atomic force vector
$\mathbf{F}_i^{\text{NEB}}$	NEB force at the $(i + 1)^{\text{th}}$ image
\mathbf{F}_i^{s}	spring force at the $(i + 1)^{\text{th}}$ image
\mathbf{F}_{rot}	rotational force
$\mathbf{F}_{\text{trans}}$	translational force
i_{CI}	index for the climbing image
\mathbf{I}_N	$N \times N$ identity matrix
$k(\cdot, \cdot)$	covariance function
k^{s}	spring constant
k_i^{s}	spring constant between i^{th} and $(i + 1)^{\text{th}}$ images

Symbols

$k_x(\cdot, \cdot)$	squared exponential covariance function
$k_x^{M-3/2}(\cdot, \cdot)$	Matérn-3/2 covariance function
$k_x^{M-5/2}(\cdot, \cdot)$	Matérn-5/2 covariance function
$k_{1/r}(\cdot, \cdot)$	inverse-distance covariance function
$K(\cdot, \cdot)$	prior covariance matrix
\mathbf{K}_{ext}	extended prior covariance matrix
$\mathbf{K}_{\text{ext}}^*$	extended prior covariance matrix
$\mathbf{K}_{\mathbf{f}}$	posterior covariance matrix
$\mathbf{K}_{\mathbf{f}^*}$	posterior predictive covariance matrix
$\mathbf{K}_{\mathbf{f} \mathbf{f}}$	conditional covariance matrix
l	length scale of a covariance function
l_d	length scale for the d^{th} input coordinate
$l_{\phi}(\cdot, \cdot)$	length scale for an atom pair
$m(\cdot)$	mean function
\mathbf{m}	prior mean vector
$\mathbf{m}_{\mathbf{f}}$	posterior mean vector
$\mathbf{m}_{\mathbf{f}^*}$	posterior predictive mean vector
$\mathbf{m}_{\mathbf{f} \mathbf{f}}$	conditional mean vector
N	number of observation points
N^*	number of prediction points
N_{im}	number of images in a nudged elastic band
N_{m}	number of moving atoms
$\hat{\mathbf{N}}$	dimer orientation
$\hat{\mathbf{N}}^*$	dimer orientation after a preliminary rotation
$\mathcal{N}(\cdot, \cdot)$	Gaussian distribution
$p(\cdot)$	probability density
$r_{i,j}(\cdot)$	distance between atoms i and j
\mathbf{R}_0	middle point of a dimer
\mathbf{R}_1	image 1 of a dimer
\mathbf{R}_1^*	image 1 of a dimer after a preliminary rotation
\mathbf{R}_2	image 2 of a dimer
\mathbf{R}_i	the $(i + 1)^{\text{th}}$ image in a nudged elastic band
\mathbb{R}	the set of real numbers
$\text{Var}[\cdot]$	variance

x_d	the d^{th} input coordinate
$x_{i,d}$	the d^{th} coordinate of atom i
\mathbf{x}	input vector
$\mathbf{x}^{(i)}$	the i^{th} observation point
$\mathbf{x}^{*(i)}$	the i^{th} prediction point
\mathbf{X}	matrix of observation points
\mathbf{X}^*	matrix of prediction points
$y^{(i)}$	the i^{th} output observation
\mathbf{y}	vector of output observations
\mathbf{y}_{ext}	extended observation vector
\mathbb{Z}	the set of integers
$\Gamma(\cdot)$	gamma function
$\Delta_{\mathbf{R}}$	dimer separation
$\boldsymbol{\theta}$	vector of hyperparameters
$\boldsymbol{\theta}_{\text{MAP}}$	maximum a posteriori estimate for hyperparameters
$\boldsymbol{\theta}_{\text{ML}}$	maximum likelihood estimate for hyperparameters
ν	smoothness parameter of the Matérn covariance function
$\boldsymbol{\rho}$	vector of parameters
σ^2	noise variance
σ_{c}^2	constant covariance
σ_{d}^2	noise variance for derivatives
σ_{m}	magnitude of a covariance function
$\boldsymbol{\Sigma}$	extended noise covariance matrix
$\boldsymbol{\tau}_i$	unnormalized path tangent at the $(i + 1)^{\text{th}}$ image
$\hat{\boldsymbol{\tau}}_i$	normalized path tangent at the $(i + 1)^{\text{th}}$ image
$\phi(\cdot, \cdot)$	atom pair type
ω	rotation angle
ω^*	preliminary rotation angle
$\hat{\boldsymbol{\Omega}}$	rotation direction
$\hat{\boldsymbol{\Omega}}^*$	rotation direction after a preliminary rotation

1. Introduction

Theoretical chemistry utilizes physics, mathematics and computer science to explain and predict structural and dynamical properties of molecules and materials. One of the key concepts in theoretical chemistry is a potential energy surface, often described as a function in a high-dimensional space of atom coordinates, which contains the essential information of the properties of the system at a finite temperature. The most interesting locations on the energy surface are its local minimum points, corresponding to stable states of the system, and first-order saddle points located at energy ridges separating those states. Transitions from one state to another, caused by thermal fluctuations, can be characterized by a minimum energy path connecting the two states, and the highest point of this path is located at a first-order saddle point. The minimum energy path cannot be considered as an actual trajectory for the transition but rather a path of maximal statistical weight. In principle, such transitions could be simulated by classical dynamics, but since the time scale of the transition is often extremely large compared to the frequency of the thermal vibrations, statistical tools such as transition state theory (Wigner, 1938; Kramers, 1940; Keck, 1967) are required. A common approach is the harmonic approximation to the transition state theory (Vineyard, 1957), where the rate of the transition is estimated based on the energy and its second derivatives at the initial state and the saddle point.

Given an initial configuration of atoms, it is straightforward to locate the nearest minimum point on the energy surface with common optimization methods. A more challenging task is to find the saddle points located along the minimum energy paths leading to other relevant states of the system. A group of iterative algorithms, called surface walking methods or mode following methods, has been developed for the task to find a saddle point without knowledge of the final state of the transition. The common principle of these algorithms is to make the problem approachable for optimization methods by reverting the gradient component in the direction of the lowest energy curvature, i.e., the direction of the eigenvector corresponding to the lowest eigenvalue of the Hessian matrix, also known

as the minimum curvature mode. With this modification, minimization of energy is supposed to lead to a first-order saddle point where the energy is maximized in the direction of the minimum energy path but minimized in all perpendicular directions. If the second derivatives of energy are easily available, all eigenvalues of the Hessian matrix can be calculated and a modified Hessian used to guide the saddle point search (Cerjan and Miller, 1981; Simons et al. 1983; Banerjee et al., 1985). A more efficient approach is to find the direction of the lowest curvature based only on the first derivatives (Henkelman and Jónsson, 1999; Munro and Wales, 1999; Malek and Mousseau, 2000). An example of such an approach is the dimer method (Henkelman and Jónsson, 1999), where a pair of closely spaced points is rotated towards the minimum curvature mode and translated towards a saddle point based on a modified gradient.

The task of finding a saddle point along a minimum energy path is easier, if also the final state of the transition has been found. In chain-of-states methods, such as the nudged elastic band (NEB) method (Mills et al., 1995; Jónsson et al., 1998), the path is represented as a discrete chain of points which is moved and stretched towards a minimum energy path so that the component of the energy gradient perpendicular to the path goes to zero at all points of the chain. In the NEB method, the distribution of the points along the path is controlled by a spring force acting parallel to the path. The actual saddle point can be found by reverting the gradient component parallel to the path at one of the points of the chain and letting this point climb along the path towards the saddle point.

Both surface walking and chain-of-states methods may require hundreds of iterations and evaluations of energy and its first derivatives. Since these evaluations typically involve computationally expensive electronic structure calculations, the information obtained from previous iterations should be utilized as efficiently as possible to decrease the number of iterations. A prominent approach for this purpose is to utilize machine learning to construct an approximate energy surface and perform the saddle point search based on the approximate model. The accuracy of the solution can be checked with further evaluations, which can then be used to update the model for the following iterations. Assuming that training of the machine learning model and evaluations of the approximate energy and derivatives are significantly cheaper than the accurate evaluations, the total number of the expensive evaluations can be reduced and the saddle point search hence accelerated. This general scheme has been introduced by Peterson (2016) with a demonstration of applying artificial neural network models to NEB calculations.

In this dissertation, similar algorithms to enhance searches of minimum energy paths and saddle points are developed using Gaussian process (GP) models as flexible prior probability models for potential energy surfaces. Observed values of both energy and its derivatives can be used to update

the model, and the posterior predictive distribution obtained as a result of Bayesian inference provides also an uncertainty estimate, which can be utilized when selecting new observation points. Whereas optimization of a large number of weights of a neural network model may be challenging due to many local minima of the cost function, optimization of the hyperparameters of a GP model is typically a much easier task. Gaussian process regression have been shown to perform well especially when learning from small training data sets (Lampinen and Vehtari, 2001; Kamath et al., 2018), which makes it an appealing approach for this application. The GP-NEB algorithm (Publications I–III), based on the nudged elastic band method, finds a minimum energy path and a saddle point between two known states, whereas the GP-dimer algorithm (Publication IV), based on the dimer method, only finds a saddle point located in the vicinity of a given start point.

The dissertation consists of four articles and this overview part. The following chapter reviews the basics of Gaussian process regression, explains how to deal with derivatives in Gaussian process models, and shows how the framework is applied to modelling of potential energy surfaces in Publications I–IV. Chapter 3 reviews the regular nudged elastic band and dimer methods, and chapter 4 summarizes the contributions of Publications I–IV by explaining the main features of the GP-NEB and GP-dimer algorithms and presenting some test results.

2. Gaussian processes

Gaussian processes are a class of stochastic processes, particularly suitable for defining flexible prior distributions for functions in a Bayesian approach to supervised learning problems. Gaussian process models have been used for decades, e.g., in signal processing and geostatistics, where methods known as Wiener-Kolmogorov filtering (Kolmogorov, 1941; Wiener, 1949) and kriging (Krige, 1951; Matheron, 1963), respectively, correspond to Gaussian process regression. Bayesian interpretation of the GP framework has been presented by Kimeldorf and Wahba (1970), Blight and Ott (1975), and O'Hagan (1978) and later adopted by neural network researchers (Neal, 1995; Williams and Rasmussen, 1996; Rasmussen, 1996) who realized that neural network models in the limit of infinite number of hidden units can be handled elegantly by replacing the networks by Gaussian processes.

This chapter reviews the basics of Gaussian process regression from the Bayesian point of view, explains how to deal with derivatives in Gaussian process models, and finally shows how the framework is applied to approximation of potential energy surfaces in Publications I–IV. A more thorough review of the Bayesian approach to Gaussian process regression, including many of the basic equations appearing in this chapter, can be found in the book of Rasmussen and Williams (2006).

2.1 Gaussian process model

By definition, a Gaussian process is a collection of random variables with a multivariate Gaussian distribution for any finite set of these random variables. The random variables are most often indexed in a continuous domain such as time or space. In that case, the probability distribution of the Gaussian process itself is the infinite-dimensional joint distribution of all the random variables, in other words, a distribution over functions in a continuous input space. In the machine learning community, the term Gaussian process is often used to refer also to the model that defines the distribution of the process (see, e.g., Rasmussen and Williams, 2006). In

this context, Gaussian processes compose a versatile modelling framework to specify prior probability distributions directly on functions and perform Bayesian inference on them based on observed data.

A Gaussian process model for the probability distribution of function $f : \mathbb{R}^D \rightarrow \mathbb{R}$ is specified by a mean function $m : \mathbb{R}^D \rightarrow \mathbb{R}$ and a covariance function $k : \mathbb{R}^D \times \mathbb{R}^D \rightarrow \mathbb{R}$. The mean function specifies the mean level of the marginal distribution of $f(\mathbf{x})$ at a given input point $\mathbf{x} \in \mathbb{R}^D$, i.e., $E[f(\mathbf{x})] = m(\mathbf{x})$, and the covariance function specifies how the values of f at any two input points, $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^D$, correlate with each other, more precisely, $E[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))] = k(\mathbf{x}, \mathbf{x}')$. Given an arbitrary set of input points $\mathbf{X} = [\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(N)}]^\top$, the joint probability distribution of function values $\mathbf{f} = [f(\mathbf{x}^{(1)}), f(\mathbf{x}^{(2)}), \dots, f(\mathbf{x}^{(N)})]^\top$ is defined as a multivariate Gaussian distribution

$$p(\mathbf{f}) = \mathcal{N}(\mathbf{m}, K(\mathbf{X}, \mathbf{X})) \quad (2.1)$$

with mean vector

$$\mathbf{m} = [m(\mathbf{x}^{(1)}), m(\mathbf{x}^{(2)}), \dots, m(\mathbf{x}^{(N)})]^\top$$

and covariance matrix

$$K(\mathbf{X}, \mathbf{X}) = \begin{bmatrix} k(\mathbf{x}^{(1)}, \mathbf{x}^{(1)}) & k(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) & \dots & k(\mathbf{x}^{(1)}, \mathbf{x}^{(N)}) \\ k(\mathbf{x}^{(2)}, \mathbf{x}^{(1)}) & k(\mathbf{x}^{(2)}, \mathbf{x}^{(2)}) & \dots & k(\mathbf{x}^{(2)}, \mathbf{x}^{(N)}) \\ \vdots & \vdots & \ddots & \vdots \\ k(\mathbf{x}^{(N)}, \mathbf{x}^{(1)}) & k(\mathbf{x}^{(N)}, \mathbf{x}^{(2)}) & \dots & k(\mathbf{x}^{(N)}, \mathbf{x}^{(N)}) \end{bmatrix}.$$

2.2 Covariance functions

From now on, the mean function of the prior GP model is assumed to be set to zero, which is a common practice and applied also in Publications I–IV after a suitable shift of the zero level of the data. The essential part of a Gaussian process model is the covariance function, which can be used to encode favourable properties of the unknown function. From the perspective of machine learning, it has a particularly important role in defining what can be learned about the function based on observed values. If a covariance function $k(\mathbf{x}, \mathbf{x}')$ depends only on the vector between the two points, $\mathbf{x} - \mathbf{x}'$, it is called stationary since it behaves similarly in all parts of the input space. If a covariance function is also isotropic, it can be written simply as a function of the distance $\|\mathbf{x} - \mathbf{x}'\| = \sqrt{\sum_{d=1}^D (x_d - x'_d)^2}$, which means that the behaviour is similar in all directions.

A common stationary example is the squared exponential (or perhaps more precisely exponentiated quadratic) covariance function

$$k(\mathbf{x}, \mathbf{x}') = \sigma_m^2 \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2l^2}\right) \quad (2.2)$$

where σ_m and l are the hyperparameters of the covariance function. The covariance is larger when the two input points are closer to each other and decreases with increasing distance. The magnitude σ_m defines the process variance, i.e., how much the values of f tend to deviate from the mean function, and the length scale l defines how far the effect of the covariance function fades out. In this isotropic form, the length scale is the same in all directions, but it is also possible to give separate length scales l_d for each input coordinate $d = 1, \dots, D$:

$$k(\mathbf{x}, \mathbf{x}') = \sigma_m^2 \exp\left(-\sum_{d=1}^D \frac{(x_d - x'_d)^2}{2l_d^2}\right). \quad (2.3)$$

A GP model with a squared exponential covariance function favours extremely smooth functions. This property stems from the fact that the covariance function is infinite times differentiable, implying that sample functions drawn from the probability model are as well infinite times differentiable.

Even though the squared exponential covariance function is one of the most popular choices for a GP model, such a demanding smoothness assumption may be unrealistic for some real-world applications (Stein, 1999). The Matérn class of covariance functions (Matérn, 1960) allows to loosen the smoothness assumptions by adjusting an additional hyperparameter ν . The general form of the isotropic Matérn covariance function is given by

$$k(\mathbf{x}, \mathbf{x}') = \sigma_m^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}\|\mathbf{x} - \mathbf{x}'\|}{l}\right)^\nu B_\nu\left(\frac{\sqrt{2\nu}\|\mathbf{x} - \mathbf{x}'\|}{l}\right), \quad (2.4)$$

where Γ denotes the gamma function and B_ν the modified Bessel function of the second kind (Olver and Maximon, 2010). A more convenient presentation is obtained when $\nu = p + 1/2$, where $0 \leq p \in \mathbb{Z}$:

$$k(\mathbf{x}, \mathbf{x}') = \sigma_m^2 \exp\left(-\frac{\sqrt{2\nu}\|\mathbf{x} - \mathbf{x}'\|}{l}\right) \frac{p!}{(2p)!} \sum_{i=0}^p \binom{p+i}{i!(p-i)!} \left(\frac{2\sqrt{2\nu}\|\mathbf{x} - \mathbf{x}'\|}{l}\right)^{p-i}.$$

Sample functions drawn from this model are n times differentiable when $n > \nu$. When ν approaches infinity, the Matérn class converges to the squared exponential covariance function, whereas a choice of $\nu = 1/2$ leads to the exponential covariance function

$$k(\mathbf{x}, \mathbf{x}') = \sigma_m^2 \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|}{l}\right) \quad (2.5)$$

and continuous but non-differentiable, roughly varying sample functions. In practice, a good compromise for the smoothness assumption is often obtained by choosing a once differentiable process with $\nu = 3/2$, so that

$$k(\mathbf{x}, \mathbf{x}') = \sigma_m^2 \left(1 + \frac{\sqrt{3}\|\mathbf{x} - \mathbf{x}'\|}{l}\right) \exp\left(-\frac{\sqrt{3}\|\mathbf{x} - \mathbf{x}'\|}{l}\right), \quad (2.6)$$

or a twice differentiable process with $\nu = 5/2$, so that

$$k(\mathbf{x}, \mathbf{x}') = \sigma_m^2 \left(1 + \frac{\sqrt{5} \|\mathbf{x} - \mathbf{x}'\|}{l} + \frac{5 \|\mathbf{x} - \mathbf{x}'\|^2}{3l^2} \right) \exp\left(-\frac{\sqrt{5} \|\mathbf{x} - \mathbf{x}'\|}{l}\right). \quad (2.7)$$

Similarly as for the squared exponential covariance function, it is possible to give separate length scales l_d for each input coordinate $d = 1, \dots, D$ by replacing the scaled distance $\|\mathbf{x} - \mathbf{x}'\|/l$ with $\sqrt{\sum_{d=1}^D ((x_d - x'_d)/l_d)^2}$.

One more simple covariance function encountered in this dissertation is the constant function

$$k(\mathbf{x}, \mathbf{x}') = \sigma_c^2, \quad (2.8)$$

often used as an auxiliary term with other covariance functions. As a constant covariance implies full correlation between all function values, adding σ_c^2 to the covariance function corresponds to adding a constant intercept term to the process so that the unknown constant has a Gaussian prior distribution with variance σ_c^2 . Thus, the constant covariance term can be used to allow variation of the global mean level even if the mean function was set to zero.

2.3 Gaussian process regression

Consider a regression problem with a training data set (\mathbf{X}, \mathbf{y}) , including output observations $\mathbf{y} = [y^{(1)}, y^{(2)}, \dots, y^{(N)}]^\top$ made at N input points $\mathbf{X} = [\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(N)}]^\top$, and an observation model

$$p(\mathbf{y}|\mathbf{f}) = \prod_{i=1}^N p(y^{(i)} | f(\mathbf{x}^{(i)})), \quad (2.9)$$

where $\mathbf{f} = [f(\mathbf{x}^{(1)}), f(\mathbf{x}^{(2)}), \dots, f(\mathbf{x}^{(N)})]^\top$ is a vector of latent function values at the input data points. In a typical Bayesian modelling approach, the latent function $f(\mathbf{x})$ would be specified by a set of unknown parameters $\boldsymbol{\rho}$ with a prior distribution $p(\boldsymbol{\rho})$. According to the Bayes' theorem, the posterior distribution of $\boldsymbol{\rho}$ conditioned on the training data would be given by

$$p(\boldsymbol{\rho} | \mathbf{X}, \mathbf{y}) = \frac{p(\mathbf{y} | \mathbf{X}, \boldsymbol{\rho}) p(\boldsymbol{\rho})}{p(\mathbf{y} | \mathbf{X})}, \quad (2.10)$$

where $p(\mathbf{y} | \mathbf{X}, \boldsymbol{\rho})$ with fixed \mathbf{y} and \mathbf{X} is the likelihood of $\boldsymbol{\rho}$ given by the observation model and the normalization constant $p(\mathbf{y} | \mathbf{X})$ is obtained by integrating over the parameters,

$$p(\mathbf{y} | \mathbf{X}) = \int_{\boldsymbol{\rho}} p(\mathbf{y} | \mathbf{X}, \boldsymbol{\rho}) p(\boldsymbol{\rho}) d\boldsymbol{\rho}. \quad (2.11)$$

In Gaussian process regression, the prior distribution is given directly to the values of the latent function f . For this reason, Gaussian process models are often called non-parametric, but sometimes also infinite-parametric

since the unlimited collection of latent function values \mathbf{f} at the training input points can be seen as the parameters of the model. When modelling the prior of f with a Gaussian process with mean function $m(\mathbf{x}) = \mathbf{0}$ and a prior covariance function $k(\mathbf{x}, \mathbf{x}' | \boldsymbol{\theta})$, the posterior distribution of \mathbf{f} , conditional on the hyperparameters $\boldsymbol{\theta}$ of the covariance function, is given by

$$p(\mathbf{f} | \mathbf{X}, \mathbf{y}, \boldsymbol{\theta}) = \frac{p(\mathbf{y} | \mathbf{f})p(\mathbf{f} | \mathbf{X}, \boldsymbol{\theta})}{p(\mathbf{y} | \mathbf{X}, \boldsymbol{\theta})}. \quad (2.12)$$

Generally, evaluation of this distribution requires approximative methods such as Monte Carlo integration (Neal, 1999), Laplace approximation (Williams and Barber, 1998), expectation propagation (Minka, 2001), or variational methods (Gibbs and MacKay, 2000), but in case of a Gaussian observation model

$$p(\mathbf{y} | \mathbf{f}) = \prod_{i=1}^N \mathcal{N}(y^{(i)} | f(\mathbf{x}^{(i)}), \sigma^2), \quad (2.13)$$

the posterior can be presented in an analytical Gaussian form:

$$p(\mathbf{f} | \mathbf{X}, \mathbf{y}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{f} | \mathbf{m}_{\mathbf{f}}, \mathbf{K}_{\mathbf{f}}), \quad (2.14)$$

where

$$\mathbf{m}_{\mathbf{f}} = K(\mathbf{X}, \mathbf{X})(K(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathbf{I}_N)^{-1} \mathbf{y}$$

and

$$\mathbf{K}_{\mathbf{f}} = K(\mathbf{X}, \mathbf{X}) - K(\mathbf{X}, \mathbf{X})(K(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathbf{I}_N)^{-1} K(\mathbf{X}, \mathbf{X})$$

with \mathbf{I}_N denoting an $N \times N$ identity matrix.

To predict function values $\mathbf{f}^* = [f(\mathbf{x}^{*(1)}), f(\mathbf{x}^{*(2)}), \dots, f(\mathbf{x}^{*(N^*)})]^\top$ at an arbitrary set of input points $\mathbf{X}^* = [\mathbf{x}^{*(1)}, \mathbf{x}^{*(2)}, \dots, \mathbf{x}^{*(N^*)}]^\top$, consider first the joint prior distribution of \mathbf{f} and \mathbf{f}^* :

$$p\left(\begin{bmatrix} \mathbf{f} \\ \mathbf{f}^* \end{bmatrix} \middle| \begin{bmatrix} \mathbf{X} \\ \mathbf{X}^* \end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix} \mathbf{f} \\ \mathbf{f}^* \end{bmatrix} \middle| \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} K(\mathbf{X}, \mathbf{X}) & K(\mathbf{X}, \mathbf{X}^*) \\ K(\mathbf{X}^*, \mathbf{X}) & K(\mathbf{X}^*, \mathbf{X}^*) \end{bmatrix}\right). \quad (2.15)$$

According to the conditionalization properties of the multivariate Gaussian distribution (Mises, 1964), the conditional distribution of \mathbf{f}^* , given \mathbf{f} , becomes

$$p(\mathbf{f}^* | \mathbf{X}^*, \mathbf{X}, \mathbf{f}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{f}^* | \mathbf{m}_{\mathbf{f}^* | \mathbf{f}}, \mathbf{K}_{\mathbf{f}^* | \mathbf{f}}), \quad (2.16)$$

where

$$\mathbf{m}_{\mathbf{f}^* | \mathbf{f}} = K(\mathbf{X}^*, \mathbf{X})K(\mathbf{X}, \mathbf{X})^{-1} \mathbf{f}$$

and

$$\mathbf{K}_{\mathbf{f}^* | \mathbf{f}} = K(\mathbf{X}^*, \mathbf{X}^*) - K(\mathbf{X}^*, \mathbf{X})K(\mathbf{X}, \mathbf{X})^{-1}K(\mathbf{X}, \mathbf{X}^*).$$

The posterior predictive distribution for the function values \mathbf{f}^* is obtained by marginalizing from the joint posterior

$$p(\mathbf{f}, \mathbf{f}^* | \mathbf{X}^*, \mathbf{X}, \mathbf{y}, \boldsymbol{\theta}) = p(\mathbf{f}^* | \mathbf{X}^*, \mathbf{X}, \mathbf{f}, \boldsymbol{\theta})p(\mathbf{f} | \mathbf{X}, \mathbf{y}, \boldsymbol{\theta}).$$

With the Gaussian observation model, also this distribution remains Gaussian:

$$p(\mathbf{f}^* | \mathbf{X}^*, \mathbf{X}, \mathbf{y}, \boldsymbol{\theta}) = \int_{\mathbf{f}} p(\mathbf{f}^* | \mathbf{X}^*, \mathbf{X}, \mathbf{f}, \boldsymbol{\theta}) p(\mathbf{f} | \mathbf{X}, \mathbf{y}, \boldsymbol{\theta}) d\mathbf{f} = \mathcal{N}(\mathbf{f}^* | \mathbf{m}_{\mathbf{f}^*}, \mathbf{K}_{\mathbf{f}^*}), \quad (2.17)$$

where

$$\mathbf{m}_{\mathbf{f}^*} = K(\mathbf{X}^*, \mathbf{X})(K(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathbf{I}_N)^{-1} \mathbf{y}$$

and

$$\mathbf{K}_{\mathbf{f}^*} = K(\mathbf{X}^*, \mathbf{X}^*) - K(\mathbf{X}^*, \mathbf{X})(K(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathbf{I}_N)^{-1} K(\mathbf{X}, \mathbf{X}^*).$$

The equations above are all conditional on the prior covariance function $k(\mathbf{x}, \mathbf{x}')$ with known hyperparameters $\boldsymbol{\theta}$. The standard way to learn the hyperparameters is to maximize the marginal likelihood of $\boldsymbol{\theta}$ given a data set $\{\mathbf{X}, \mathbf{y}\}$, appearing in the denominator of equation 2.12:

$$\boldsymbol{\theta}_{\text{ML}} = \arg \max_{\boldsymbol{\theta}} p(\mathbf{y} | \mathbf{X}, \boldsymbol{\theta}) = \arg \max_{\boldsymbol{\theta}} \int_{\mathbf{f}} p(\mathbf{y} | \mathbf{f}) p(\mathbf{f} | \mathbf{X}, \boldsymbol{\theta}) d\mathbf{f}. \quad (2.18)$$

With the Gaussian observation model, the marginal likelihood is simply given by

$$p(\mathbf{y} | \mathbf{X}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{y} | \mathbf{0}, K(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathbf{I}_N). \quad (2.19)$$

To improve stability and data efficiency, it is also possible to define a prior distribution $p(\boldsymbol{\theta})$ (hyperprior), as done in Publications I–IV, and maximize the marginal posterior probability density $p(\boldsymbol{\theta} | \mathbf{y}, \mathbf{X}) \propto p(\boldsymbol{\theta}) p(\mathbf{y} | \mathbf{X}, \boldsymbol{\theta})$:

$$\boldsymbol{\theta}_{\text{MAP}} = \arg \max_{\boldsymbol{\theta}} p(\boldsymbol{\theta} | \mathbf{y}, \mathbf{X}) = \arg \max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}) p(\mathbf{y} | \mathbf{X}, \boldsymbol{\theta}). \quad (2.20)$$

An alternative to a maximum a posteriori estimate would be to integrate over the uncertainty of the marginal posterior $p(\boldsymbol{\theta} | \mathbf{y}, \mathbf{X})$ using approximations based on, e.g., Monte Carlo or grid sampling or a central composite design (Rue et al., 2009; Vanhatalo et al., 2010). In addition to the hyperparameters of the covariance function, the parameters of the observation model such as the noise variance σ^2 can be similarly treated as unknown hyperparameters and incorporated in the optimization or integration.

The elegance of Gaussian process regression relies on the implicit encoding of the function properties via selection of the covariance function, which allows flexible models without restricting to simple parametric forms. The strength of the framework is most apparent in prediction based on small training data sets (Lampinen and Vehtari, 2001; Kamath et al., 2018). The price of the elegance, however, is realized as computational challenges with large data sets, since training of the model involves solving a linear system associated with the training covariance matrix. This is typically performed via a Cholesky decomposition with a cubic computational cost with respect to the number of training observations, which makes large data sets infeasible (Rasmussen and Williams, 2006). Common ways to alleviate the problem involve compactly supported covariance functions leading to

sparse covariance matrices with zero covariance between data points far away from each other (Sansò and Schuh, 1987; Wu, 1995; Wendland, 1995; Vanhatalo and Vehtari, 2008), sparse approximations by representing the training data set with a smaller set of inducing points (Csató and Opper, 2002; Seeger et al., 2003; Quiñonero-Candela and Rasmussen, 2005; Snelson and Ghahramani, 2006; Titsias, 2009; Hensman et al., 2013), and mixture-of-experts models where the computation can be distributed over several local data sets (Deisenroth and Ng, 2015). However, a more recent inference approach (Gardner et al., 2018; Wang et al., 2019) avoids the Cholesky decomposition by using a modified batched conjugate gradients algorithm and allows quadratic scaling without compromising the accuracy of the inference. In this approach, the covariance matrix is accessed through matrix-matrix multiplications which can be computed efficiently with GPU (graphics processing unit) hardware.

2.4 Regression with derivatives

In many applications, it is desirable to predict also the derivatives of f or incorporate information about the derivatives into the model. For Gaussian process models with differentiable covariance functions, this turns out to be straightforward since the linearity of differentiation implies that the derivative of a Gaussian process is another Gaussian process (O’Hagan, 1992; Rasmussen, 2003; Solak et al., 2003; Riihimäki and Vehtari, 2010). The covariance between a partial derivative at \mathbf{x} and a function value at \mathbf{x}' is simply given by differentiating the covariance function,

$$\text{Cov}\left[\frac{\partial f(\mathbf{x})}{\partial x_d}, f(\mathbf{x}')\right] = \frac{\partial}{\partial_1 x_d} \text{Cov}[f(\mathbf{x}), f(\mathbf{x}')] = \frac{\partial k(\mathbf{x}, \mathbf{x}')}{\partial_1 x_d}, \quad (2.21)$$

and similarly,

$$\text{Cov}\left[\frac{\partial f(\mathbf{x})}{\partial x_{d_1}}, \frac{\partial f(\mathbf{x}')}{\partial x'_{d_2}}\right] = \frac{\partial^2}{\partial_1 x_{d_1} \partial_2 x'_{d_2}} \text{Cov}[f(\mathbf{x}), f(\mathbf{x}')] = \frac{\partial^2 k(\mathbf{x}, \mathbf{x}')}{\partial_1 x_{d_1} \partial_2 x'_{d_2}}. \quad (2.22)$$

The notation ∂_1 indicates here that the covariance is differentiated with respect to a component of the first argument \mathbf{x} , and ∂_2 correspondingly refers to the second argument \mathbf{x}' .

To predict partial derivatives of f , vector \mathbf{f}^* and covariance matrices $K(\mathbf{X}^*, \mathbf{X}^*)$ and $K(\mathbf{X}^*, \mathbf{X})$ in equations 2.15–2.17 can be extended as

$$\begin{bmatrix} \mathbf{f}^* \\ \frac{\partial f(\mathbf{X}^*)}{\partial x_1^*} \\ \frac{\partial f(\mathbf{X}^*)}{\partial x_2^*} \\ \vdots \\ \frac{\partial f(\mathbf{X}^*)}{\partial x_D^*} \end{bmatrix}, \begin{bmatrix} K(\mathbf{X}^*, \mathbf{X}^*) & \frac{\partial K(\mathbf{X}^*, \mathbf{X}^*)}{\partial_2 x_1^*} & \frac{\partial K(\mathbf{X}^*, \mathbf{X}^*)}{\partial_2 x_2^*} & \dots & \frac{\partial K(\mathbf{X}^*, \mathbf{X}^*)}{\partial_2 x_D^*} \\ \frac{\partial K(\mathbf{X}^*, \mathbf{X}^*)}{\partial_1 x_1^*} & \frac{\partial K(\mathbf{X}^*, \mathbf{X}^*)}{\partial_1 x_1^* \partial_2 x_1^*} & \frac{\partial K(\mathbf{X}^*, \mathbf{X}^*)}{\partial_1 x_1^* \partial_2 x_2^*} & \dots & \frac{\partial K(\mathbf{X}^*, \mathbf{X}^*)}{\partial_1 x_1^* \partial_2 x_D^*} \\ \frac{\partial K(\mathbf{X}^*, \mathbf{X}^*)}{\partial_1 x_2^*} & \frac{\partial K(\mathbf{X}^*, \mathbf{X}^*)}{\partial_1 x_2^* \partial_2 x_1^*} & \frac{\partial K(\mathbf{X}^*, \mathbf{X}^*)}{\partial_1 x_2^* \partial_2 x_2^*} & \dots & \frac{\partial K(\mathbf{X}^*, \mathbf{X}^*)}{\partial_1 x_2^* \partial_2 x_D^*} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial K(\mathbf{X}^*, \mathbf{X}^*)}{\partial_1 x_D^*} & \frac{\partial K(\mathbf{X}^*, \mathbf{X}^*)}{\partial_1 x_D^* \partial_2 x_1^*} & \frac{\partial K(\mathbf{X}^*, \mathbf{X}^*)}{\partial_1 x_D^* \partial_2 x_2^*} & \dots & \frac{\partial K(\mathbf{X}^*, \mathbf{X}^*)}{\partial_1 x_D^* \partial_2 x_D^*} \end{bmatrix}, \text{ and } \begin{bmatrix} K(\mathbf{X}^*, \mathbf{X}) \\ \frac{\partial K(\mathbf{X}^*, \mathbf{X})}{\partial_1 x_1^*} \\ \frac{\partial K(\mathbf{X}^*, \mathbf{X})}{\partial_1 x_2^*} \\ \vdots \\ \frac{\partial K(\mathbf{X}^*, \mathbf{X})}{\partial_1 x_D^*} \end{bmatrix},$$

respectively. Often the primary interest is in the marginal posterior predictive distribution of individual variables, whereupon the covariances between predictions of different partial derivatives and between predictions at different input points can be ignored. For example, the posterior predictive distribution of the partial derivative of f with respect to input coordinate d at \mathbf{x}^* , assuming the Gaussian observation model, is a Gaussian distribution with mean

$$\mathbb{E}\left[\frac{\partial f(\mathbf{x}^*)}{\partial x_d^*} \middle| \mathbf{X}, \mathbf{y}, \boldsymbol{\theta}\right] = \frac{\partial K(\mathbf{x}^*, \mathbf{X})}{\partial x_d^*} (K(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathbf{I}_N)^{-1} \mathbf{y} \quad (2.23)$$

and variance

$$\text{Var}\left[\frac{\partial f(\mathbf{x}^*)}{\partial x_d^*} \middle| \mathbf{X}, \mathbf{y}, \boldsymbol{\theta}\right] = \frac{\partial^2 k(\mathbf{x}^*, \mathbf{x}^*)}{\partial_1 x_d^* \partial_2 x_d^*} - \frac{\partial K(\mathbf{x}^*, \mathbf{X})}{\partial x_d^*} (K(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathbf{I}_N)^{-1} \frac{\partial K(\mathbf{X}, \mathbf{x}^*)}{\partial x_d^*}. \quad (2.24)$$

Similarly, derivative observations can be included in the model by extending the observation vector \mathbf{y} to include partial derivative observations and by extending the covariance matrices correspondingly. Assuming a Gaussian noise model also for the derivative observations, the posterior predictive mean and variance for f at \mathbf{x}^* are then given as

$$\mathbb{E}[f(\mathbf{x}^*) | \mathbf{y}_{\text{ext}}, \mathbf{X}, \boldsymbol{\theta}] = \mathbf{K}_{\text{ext}}^* (\mathbf{K}_{\text{ext}} + \boldsymbol{\Sigma})^{-1} \mathbf{y}_{\text{ext}} \quad (2.25)$$

and

$$\text{Var}[f(\mathbf{x}^*) | \mathbf{y}_{\text{ext}}, \mathbf{X}, \boldsymbol{\theta}] = k(\mathbf{x}^*, \mathbf{x}^*) - \mathbf{K}_{\text{ext}}^* (\mathbf{K}_{\text{ext}} + \boldsymbol{\Sigma})^{-1} \mathbf{K}_{\text{ext}}^{*\top}, \quad (2.26)$$

where

$$\mathbf{y}_{\text{ext}} = \begin{bmatrix} \mathbf{y} \\ \frac{\partial f(\mathbf{X})}{\partial x_1} \\ \frac{\partial f(\mathbf{X})}{\partial x_2} \\ \vdots \\ \frac{\partial f(\mathbf{X})}{\partial x_D} \end{bmatrix}, \quad \mathbf{K}_{\text{ext}} = \begin{bmatrix} K(\mathbf{X}, \mathbf{X}) & \frac{\partial K(\mathbf{X}, \mathbf{X})}{\partial_2 x_1} & \frac{\partial K(\mathbf{X}, \mathbf{X})}{\partial_2 x_2} & \dots & \frac{\partial K(\mathbf{X}, \mathbf{X})}{\partial_2 x_D} \\ \frac{\partial K(\mathbf{X}, \mathbf{X})}{\partial_1 x_1} & \frac{\partial^2 K(\mathbf{X}, \mathbf{X})}{\partial_1 x_1 \partial_2 x_1} & \frac{\partial^2 K(\mathbf{X}, \mathbf{X})}{\partial_1 x_1 \partial_2 x_2} & \dots & \frac{\partial^2 K(\mathbf{X}, \mathbf{X})}{\partial_1 x_1 \partial_2 x_D} \\ \frac{\partial K(\mathbf{X}, \mathbf{X})}{\partial_1 x_2} & \frac{\partial^2 K(\mathbf{X}, \mathbf{X})}{\partial_1 x_2 \partial_2 x_1} & \frac{\partial^2 K(\mathbf{X}, \mathbf{X})}{\partial_1 x_2 \partial_2 x_2} & \dots & \frac{\partial^2 K(\mathbf{X}, \mathbf{X})}{\partial_1 x_2 \partial_2 x_D} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial K(\mathbf{X}, \mathbf{X})}{\partial_1 x_D} & \frac{\partial^2 K(\mathbf{X}, \mathbf{X})}{\partial_1 x_D \partial_2 x_1} & \frac{\partial^2 K(\mathbf{X}, \mathbf{X})}{\partial_1 x_D \partial_2 x_2} & \dots & \frac{\partial^2 K(\mathbf{X}, \mathbf{X})}{\partial_1 x_D \partial_2 x_D} \end{bmatrix},$$

$$\mathbf{K}_{\text{ext}}^* = \begin{bmatrix} K(\mathbf{x}^*, \mathbf{X}) & \frac{\partial K(\mathbf{x}^*, \mathbf{X})}{\partial_2 x_1} & \frac{\partial K(\mathbf{x}^*, \mathbf{X})}{\partial_2 x_2} & \dots & \frac{\partial K(\mathbf{x}^*, \mathbf{X})}{\partial_2 x_D} \end{bmatrix},$$

and

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2 \mathbf{I}_N & \mathbf{0} \\ \mathbf{0} & \sigma_d^2 \mathbf{I}_{ND} \end{bmatrix}$$

is the extended noise covariance matrix with noise variance σ_d^2 for the derivative observations. Correspondingly, the mean and variance of the posterior predictive distribution of the partial derivative of f with respect to coordinate d at \mathbf{x}^* are given as

$$\mathbb{E}\left[\frac{\partial f(\mathbf{x}^*)}{\partial x_d^*} \middle| \mathbf{y}_{\text{ext}}, \mathbf{X}, \boldsymbol{\theta}\right] = \frac{\partial \mathbf{K}_{\text{ext}}^*}{\partial x_d^*} (\mathbf{K}_{\text{ext}} + \boldsymbol{\Sigma})^{-1} \mathbf{y}_{\text{ext}} \quad (2.27)$$

and

$$\text{Var} \left[\frac{\partial f(\mathbf{x}^*)}{\partial x_d^*} \middle| \mathbf{y}_{\text{ext}}, \mathbf{X}, \boldsymbol{\theta} \right] = \frac{\partial^2 k(\mathbf{x}^*, \mathbf{x}^*)}{\partial_1 x_d^* \partial_2 x_d^*} - \frac{\partial \mathbf{K}_{\text{ext}}^*}{\partial_1 x_d^*} (\mathbf{K}_{\text{ext}} + \boldsymbol{\Sigma})^{-1} \frac{\partial \mathbf{K}_{\text{ext}}^*}{\partial_1 x_d^*}{}^{\top}. \quad (2.28)$$

Equation 2.27 is the central result used in Publications I–IV when predicting gradients of a potential energy surface based on a training data set including derivative observations.

2.5 Gaussian process models for potential energy surfaces

In this dissertation, Gaussian processes are used to model parts of potential energy surfaces in order to accelerate algorithms that aim to find minimum energy paths and saddle points on the energy surfaces. Following the notation of Publication III,

$$\mathbf{x} = [x_{1,1}, x_{1,2}, x_{1,3}, x_{2,1}, x_{2,2}, x_{2,3}, \dots, x_{N_m,1}, x_{N_m,2}, x_{N_m,3}]^{\top}$$

represents now a $3N_m$ -dimensional configuration vector including coordinates for moving atoms $1, 2, \dots, N_m \in A_m$ and f is the unknown energy of the system as a function of \mathbf{x} . The training data set consists of both the energy and its first derivatives with respect to the components of \mathbf{x} . The observations are here regarded as accurate up to floating point presentation accuracy, and thus only a really small Gaussian noise term is included in the observation model to avoid numerical issues. An approximation to the energy surface is given by the mean of the posterior predictive distribution of f (equation 2.25), and the variance of the distribution (equation 2.26) can be used as an uncertainty estimate for the GP approximation. As the algorithm proceeds, more observations are made and the model is updated until it is accurate enough to allow convergence to a minimum energy path and/or a saddle point.

In Publications I and II, a simple model with a stationary squared exponential covariance function k_x is successfully applied to meet the goals of the algorithms in a benchmark case involving rearrangements of a heptamer island on a crystal surface (Henkelman, Jóhannesson, and Jónsson, 2000; Chill et al., 2014):

$$k_x(\mathbf{x}, \mathbf{x}') = \sigma_c^2 + \sigma_m^2 \exp \left(-\frac{1}{2} \mathcal{D}_x^2(\mathbf{x}, \mathbf{x}') \right), \quad (2.29)$$

where

$$\mathcal{D}_x(\mathbf{x}, \mathbf{x}') = \sqrt{\sum_{i=1}^{N_m} \sum_{d=1}^3 \frac{(x_{i,d} - x'_{i,d})^2}{l^2}} \quad (2.30)$$

is a difference measure defined as a regular Euclidean distance between configuration vectors in the $3N_m$ -dimensional space of atom coordinates.

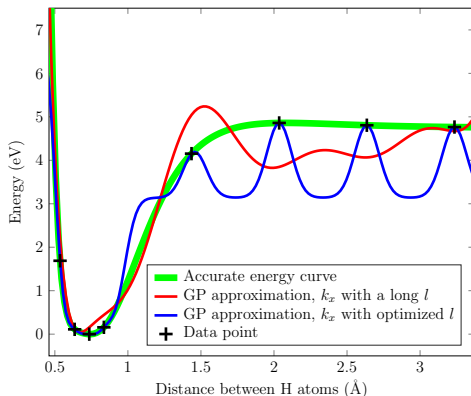


Figure 2.1. Illustration of problems in fitting a one-dimensional energy curve of a hydrogen molecule when using a Gaussian process model with stationary squared exponential covariance function k_x (equation 2.29). The training data points include accurate values for both energy and its first derivative. When the length scale of the covariance function is too long, the dominating data from the steep parts of the curve disturb the predictions at longer distances where the GP approximation does not match with the training data points even if the noise variance is assumed to be really small. When optimized, however, the length scale becomes too short for interpolation of the data points at the flat parts of the curve. Figure reproduced with permission from Publication III.

In some systems, however, strong and quickly changing repulsive forces may cause problems for stationary covariance functions, as demonstrated in Publication III. Figure 2.1 shows a simple example involving a pair of hydrogen atoms, where fitting a one-dimensional energy curve turns out to be a challenging task for covariance function k_x . With a too long length scale l , the dominating data from the steep part of the curve disturb the predictions at longer distances. To accommodate the data, the model hence favours small values of l . A short length scale, however, leads to oscillations in the GP approximation at the flat parts of the energy curve as the predictive mean between the observation points approaches the mean of the whole data.

In addition to the stationarity of the covariance function, part of the problem is due to the strong smoothness assumptions of the infinitely differentiable squared exponential covariance function. The infinitely differentiable model tends to avoid abrupt changes not only in energy and its first derivatives but also in derivatives of any order. As shown in the Supporting Information of Publication III and also by Denzel and Kästner (2018a), Matérn covariance functions with smoothness parameter $\nu = 3/2$ or $\nu = 5/2$ may perform somewhat better in modelling chemical systems than the squared exponential covariance function but are not able to fully resolve the problem. These covariance functions are here denoted by $k_x^{M-3/2}$ and $k_x^{M-5/2}$, respectively:

$$k_x^{M-3/2}(\mathbf{x}, \mathbf{x}') = \sigma_c^2 + \sigma_m^2 (1 + \sqrt{3}\mathcal{D}_x(\mathbf{x}, \mathbf{x}')) \exp(-\sqrt{3}\mathcal{D}_x(\mathbf{x}, \mathbf{x}')), \quad (2.31)$$

and

$$k_x^{M-5/2}(\mathbf{x}, \mathbf{x}') = \sigma_c^2 + \sigma_m^2 \left(1 + \sqrt{5} \mathcal{D}_x(\mathbf{x}, \mathbf{x}') + \frac{5}{3} \mathcal{D}_x^2(\mathbf{x}, \mathbf{x}') \right) \exp(-\sqrt{5} \mathcal{D}_x(\mathbf{x}, \mathbf{x}')). \quad (2.32)$$

Since potential energy typically changes faster with respect to atom coordinates when the atoms are close to each other, a modified difference measure based on inverse interatomic distances is introduced in Publication III and used also in Publication IV to replace $\mathcal{D}_x(\mathbf{x}, \mathbf{x}')$ in the squared exponential covariance function:

$$\mathcal{D}_{1/r}(\mathbf{x}, \mathbf{x}') = \sqrt{\sum_{i \in A_m} \sum_{\substack{j \in A_m, j > i \\ j \in A_f}} \frac{\left(\frac{1}{r_{i,j}(\mathbf{x})} - \frac{1}{r_{i,j}(\mathbf{x}')} \right)^2}{l_{\phi(i,j)}^2}}, \quad (2.33)$$

where

$$r_{i,j}(\mathbf{x}) = \sqrt{\sum_{d=1}^3 (x_{i,d} - x_{j,d})^2}$$

is the distance between atoms i and j and $l_{\phi(i,j)}$ denotes the length scale for atom pair type $\phi(i,j)$. The outer summation goes through the set of moving atoms A_m , and the inner summation includes all other moving atoms and the possible set of frozen atoms A_f with fixed coordinates. The closer an atom is to another atom, the larger effect a displacement of the atom towards or away from the other atom has on the difference measure. Thus, the difference measure can be interpreted to be stretched when atoms approach each other, which makes the covariance function nonstationary with respect to the atom coordinates and allows faster variation of energy in those directions.

With the modified difference measure $\mathcal{D}_{1/r}(\mathbf{x}, \mathbf{x}')$, the squared exponential covariance function gets the following form:

$$k_{1/r}(\mathbf{x}, \mathbf{x}') = \sigma_c^2 + \sigma_m^2 \exp \left(-\frac{1}{2} \sum_{i \in A_m} \sum_{\substack{j \in A_m, j > i \\ j \in A_f}} \frac{\left(\frac{1}{r_{i,j}(\mathbf{x})} - \frac{1}{r_{i,j}(\mathbf{x}')} \right)^2}{l_{\phi(i,j)}^2} \right). \quad (2.34)$$

Expressions for the partial derivatives of k_x , $k_x^{M-5/2}$, $k_x^{M-3/2}$, and $k_{1/r}$ required when dealing with derivative observations and predicting the energy gradient (as described in section 2.4) are given in the Appendix and Supporting Information of Publication III. Figure 2.2 shows a two-dimensional illustration where using the stationary squared exponential covariance function k_x leads to oscillations in spite of a dense grid of observations. The Matérn covariance functions $k_x^{M-5/2}$ and $k_x^{M-3/2}$ perform better, but the interpolation is poor especially at the lower left corner of

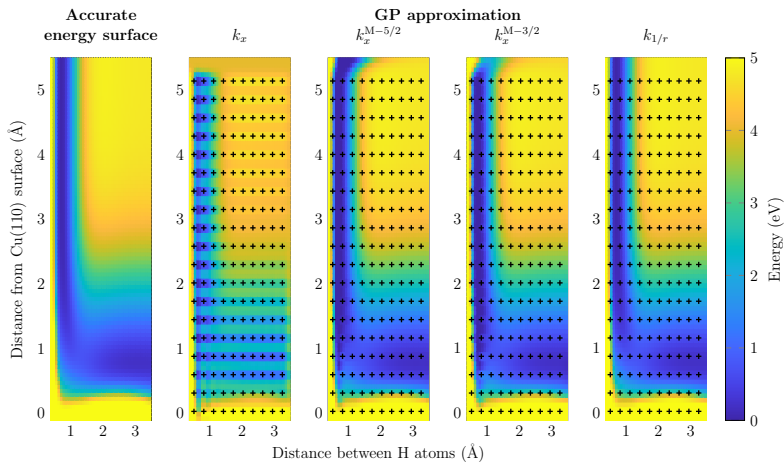


Figure 2.2. A two-dimensional cut through the potential energy surface for a pair of hydrogen atoms near a copper surface (Mills et al., 1995). The GP approximations with optimized hyperparameters are based on accurate values of energy and its first derivatives with respect to coordinates of the hydrogen atoms at the training data points shown with black crosses. With stationary squared exponential (k_x) and Matérn ($k_x^{M-5/2}$ and $k_x^{M-3/2}$) covariance functions, the high-gradient observations on the left induce oscillations in the GP approximation. When using covariance function $k_{1/r}$, based on the inverse-distance measure $\mathcal{D}_{1/r}$, the training data are interpolated without problems. Figure reproduced with permission from Publication III.

the graph. With the inverse-distance covariance function $k_{1/r}$, the high-gradient observations on the left do not disturb the fitting of the energy surface.

Another advantage of the difference measure $\mathcal{D}_{1/r}$ is that it modifies the similarity structure of the coordinate space in a natural way, which allows more efficient learning with fewer observations. Even more efficient representations for entire potential energy surfaces could be obtained by using some of the carefully designed descriptors or covariance functions associated with the Gaussian approximation potential framework (Bartók et al., 2010; Bartók and Csányi, 2015). These models approximate the total energy of the system with a sum over local atomic environments where the local energy is assumed to be invariant with respect to rotations and translations of the environment and permutations of identical atoms. For example, the SOAP (smooth overlap of atomic positions) covariance function between local environments is based on measuring the overlap in smooth density functions centred at the locations of the neighbouring atoms. In this dissertation, however, the ultimate goal is automated and accurate modelling of the surroundings of a minimum energy path or a saddle point, and the models are therefore kept fairly simple.

3. Methods for finding saddle points

A saddle point of a smooth function is a critical point with a zero gradient but neither local minimum nor maximum point of the function. In this dissertation, the interest is in first-order saddle points of potential energy surfaces located along minimum energy paths, where the Hessian matrix has exactly one negative eigenvalue. In practice, this means that the saddle point is a local maximum point along the direction of the minimum energy path but at a local minimum point along all directions perpendicular to the path.

The two main groups of saddle point search algorithms are chain-of-states methods and surface walking methods. In chain-of-states methods, such as the nudged elastic band method (Mills et al., 1995; Jónsson et al., 1998) or the string method (E et al., 2002; Ren, 2003; E et al., 2007), the task is to find a minimum energy path between the known initial and final states of a transition and to locate the saddle point at the maximum point of that path. The path is represented as a discrete chain of points in the coordinate space, i.e., a chain of states of the system, which is optimized so that the component of the energy gradient perpendicular to the path goes to zero at all points of the chain.

Another group of algorithms, called surface walking methods or mode following methods, aims at finding saddle points without knowing the final state of the transition. The start point for these algorithms is often varied close to a known initial state to search for possible transitions, but it is also common to start closer to the saddle point with an initial guess based for example on approximative minimum energy path calculations. Early examples of this group are based on calculating all eigenvectors of the Hessian matrix and, by modifying the Hessian, maximizing the energy in the direction of the lowest curvature corresponding to the smallest eigenvalue while minimizing the energy in other directions (Cerjan and Miller, 1981; Simons et al. 1983; Banerjee et al., 1985). Some later algorithms, such as the dimer method, find out only the eigenvector corresponding to the smallest eigenvalue without observing the Hessian matrix (Henkelman and Jónsson, 1999; Munro and Wales, 1999; Malek and Mousseau, 2000)

and proceed towards the saddle point based on a modified gradient. This approach is often referred to as the minimum mode following method.

Recent advances of saddle point search methods are reviewed in a book chapter by Ásgeirsson and Jónsson (2018). In this dissertation, the focus is on the nudged elastic band method and the dimer method, which are common representatives of the two main groups, both based on the first derivatives of the energy surface. In publications I–IV, these methods are used as parts of the GP-NEB and GP-dimer algorithms, where the search of a minimum energy path and/or a saddle point is enhanced using Gaussian process regression. Similar algorithms can, however, be applied to accelerate practically any other stable saddle point search method.

3.1 Nudged elastic band method

Consider a system of N_m moving atoms with configurations represented by $3N_m$ -dimensional vectors including the atom coordinates. In the nudged elastic band method (Mills et al., 1995; Jónsson et al., 1998), two given local minimum points representing the initial and final states of a transition are connected with a discrete chain of N_{im} configurations, often referred to as images of the system. The first image of the chain, \mathbf{R}_0 , is fixed to the initial state and the last image, $\mathbf{R}_{N_{im}-1}$, to the final state, whereas the intermediate images, $\mathbf{R}_i, i = 1, 2, \dots, N_{im} - 2$, are iteratively moved towards a minimum energy path. The simplest path to begin with is obtained by placing the intermediate images regularly along a straight line between \mathbf{R}_0 and $\mathbf{R}_{N_{im}-1}$. In some cases, however, this may lead to unphysical configurations with overlapping atoms. A better initial guess that avoids the overlapping can be obtained with the IDPP (image dependent pair potential) method (Smidstrup et al., 2014), which aims to place the intermediate images so that the distances between neighbouring atoms change as linearly as possible along the chain, or the geodesic approach recently introduced by Zhu et al. (2019).

The movements of the intermediate images $\mathbf{R}_i, i = 1, 2, \dots, N_{im} - 2$, are based on the energy $E(\mathbf{R}_i)$, the atomic force vector $\mathbf{F}(\mathbf{R}_i) = -\nabla E(\mathbf{R}_i)$ given by the negative gradient of the energy, and the tangent of the path, $\hat{\tau}_i$. The goal of the movements is to zero an effective force vector, here referred to as the NEB force:

$$\mathbf{F}_i^{\text{NEB}} = \mathbf{F}^\perp(\mathbf{R}_i) + \mathbf{F}_i^s, \quad (3.1)$$

where

$$\mathbf{F}^\perp(\mathbf{R}_i) = \mathbf{F}(\mathbf{R}_i) - (\mathbf{F}(\mathbf{R}_i) \cdot \hat{\tau}_i) \hat{\tau}_i \quad (3.2)$$

is the component of $\mathbf{F}(\mathbf{R}_i)$ perpendicular to the normalized path tangent $\hat{\tau}_i$ at \mathbf{R}_i and \mathbf{F}_i^s is a spring force parallel to $\hat{\tau}_i$. In the original formulation (Jónsson et al., 1998), the spring force is defined as

$$\mathbf{F}_i^s = \left((k_{i+1}^s(\mathbf{R}_{i+1} - \mathbf{R}_i) - k_i^s(\mathbf{R}_i - \mathbf{R}_{i-1})) \cdot \hat{\boldsymbol{\tau}}_i \right) \hat{\boldsymbol{\tau}}_i, \quad (3.3)$$

where k_i^s is a spring constant that determines the relative length desired for the interval between images \mathbf{R}_i and \mathbf{R}_{i-1} . A common choice of equal intervals is made for applications of NEB in Publication I, where the spring force is defined according to Henkelman and Jónsson (2000) as

$$\mathbf{F}_i^s = k^s (|\mathbf{R}_{i+1} - \mathbf{R}_i| - |\mathbf{R}_i - \mathbf{R}_{i-1}|) \hat{\boldsymbol{\tau}}_i. \quad (3.4)$$

The word *nudged* refers to the separation of the forces into two orthogonal components, which is an essential feature of the NEB method. Removal of the atomic force component parallel to the path prevents the images from sliding down towards the minimum energy points and leaves the control of the distribution of the images along the path to the spring forces. On the other hand, projection of the spring force on the path tangent prevents corner cutting since the perpendicular spring forces would tend to straighten the path at curves. A small perpendicular spring force can sometimes stabilize the path optimization by preventing kinks of the path in regions where the atomic forces perpendicular to the path are small compared to the forces along the path, but these solutions require some sort of switching function for the magnitude of the force (Jónsson et al., 1998; Trygubenko and Wales, 2004; Sheppard et al., 2008; Maras et al., 2016). Another cure for this behaviour is obtained by modifying the estimate of the path tangent (Henkelman and Jónsson, 2000). Whereas a simple estimate for the path tangent is parallel to a line segment connecting the previous and the following image,

$$\hat{\boldsymbol{\tau}}_i = \frac{\mathbf{R}_{i+1} - \mathbf{R}_{i-1}}{\|\mathbf{R}_{i+1} - \mathbf{R}_{i-1}\|}, \quad (3.5)$$

a better-behaved estimate for the direction of the tangent, used also in Publications I–III, can be achieved by

$$\boldsymbol{\tau}_i = \begin{cases} \mathbf{R}_{i+1} - \mathbf{R}_i, & \text{if } E(\mathbf{R}_{i-1}) < E(\mathbf{R}_i) < E(\mathbf{R}_{i+1}) \\ \mathbf{R}_i - \mathbf{R}_{i-1}, & \text{if } E(\mathbf{R}_{i+1}) < E(\mathbf{R}_i) < E(\mathbf{R}_{i-1}) \\ \Delta E_- (\mathbf{R}_{i+1} - \mathbf{R}_i) + \Delta E_+ (\mathbf{R}_i - \mathbf{R}_{i-1}), & \text{if } E(\mathbf{R}_{i\pm 1}) < E(\mathbf{R}_i) \\ \Delta E_+ (\mathbf{R}_{i+1} - \mathbf{R}_i) + \Delta E_- (\mathbf{R}_i - \mathbf{R}_{i-1}), & \text{if } E(\mathbf{R}_i) < E(\mathbf{R}_{i\pm 1}), \end{cases} \quad (3.6)$$

where $\Delta E_- = |E(\mathbf{R}_i) - E(\mathbf{R}_{i-1})|$ and $\Delta E_+ = |E(\mathbf{R}_{i+1}) - E(\mathbf{R}_i)|$. If the energy at an image is either higher or lower than at both of its neighbours, the direction of the tangent is defined as a weighted average of two line segments. Otherwise, only the line segment to the neighbouring image with higher energy is taken into account.

The ultimate goal of NEB calculations is often to locate the saddle point at the maximum point of the minimum energy path. However, the maximum

energy may be under- or overestimated if interpolated based on the discrete representation of the path. The climbing image nudged elastic band (CI-NEB) method (Henkelman, Uberuaga, and Jónsson, 2000) provides a solution to this problem by letting one of the images climb upwards along the path towards the saddle point. The method is often started with regular NEB iterations to find a rough shape of the path, and the image with the highest energy is then selected as the climbing image $\mathbf{R}_{i_{\text{Cl}}}$. This special image is not exposed to any spring forces, but a component of the atomic force perpendicular to the path tangent is restored and reverted to point towards the direction of increasing energy along the path. The effective force on the climbing image is hence given as

$$\mathbf{F}_{i_{\text{Cl}}}^{\text{NEB}} = \mathbf{F}(\mathbf{R}_{i_{\text{Cl}}}) - 2(\mathbf{F}(\mathbf{R}_{i_{\text{Cl}}}) \cdot \hat{\mathbf{t}}_{i_{\text{Cl}}})\hat{\mathbf{t}}_{i_{\text{Cl}}}. \quad (3.7)$$

With equal spring constants, the images leaving on each side of the climbing image are then distributed evenly on each subpath. Since the saddle point is usually the most interesting part of the minimum energy path, it is common to set a tighter convergence threshold for the NEB force of the climbing image than for the rest of the images. Figure 3.1 illustrates the effect of the climbing image on a NEB calculation on an artificial two-dimensional energy surface (Müller and Brown, 1979). Without the climbing image feature, the images of the converged path are evenly distributed and miss the saddle point found by the climbing image. The CI-NEB method has a central role in Publication II, where the details of the GP-NEB algorithm are modified to take the climbing image into account.

The simplest way to define the NEB iterations is to move the images in the direction of the NEB force with a step length proportional to the magnitude of the NEB force. This steepest descent approach may, however, require an excessively large number of iterations. A more efficient control of the step length is obtained by the velocity projection optimization

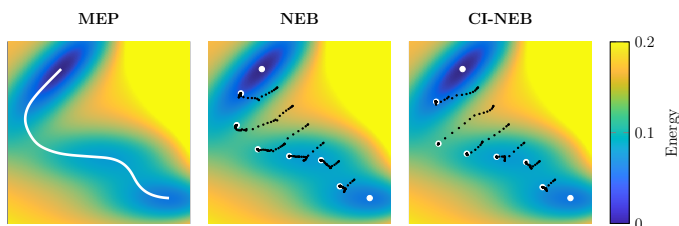


Figure 3.1. Progression of a NEB calculation on a two-dimensional Müller-Brown energy surface (Müller and Brown, 1979) with and without the climbing image feature. The white dots represent images of the converged path, and the small black dots represent earlier locations of intermediate images where energy and its first derivatives have been evaluated during the process. With CI-NEB, the third image of the path converges to the saddle point, whereas the evenly distributed NEB path takes a shortcut on the critical area. The continuous minimum energy path is presented on the left.

(VPO) algorithm (also known as quick-min) based on molecular dynamics (Jónsson et al., 1998). The movement of the images is accelerated based on the velocity Verlet algorithm (Andersen, 1980; Swope et al., 1982), or alternatively a simpler Euler integrator (Sheppard et al., 2008), with the exception that the velocity vector is projected on the direction of the NEB force or zeroed if the direction of the projected velocity would be opposite to the NEB force. This optimization method is used in the implementation of the GP-NEB algorithm in Publications I–III.

The lack of a well-defined objective function due to the force projections make NEB challenging for more advanced optimization methods, such as nonlinear conjugate gradient (Fletcher and Reeves, 1964; Polak and Ribière, 1969) or limited memory Broyden-Fletcher-Goldfarb-Shanno (LBFGS) algorithms (Nocedal, 1980; Liu and Nocedal, 1989), that modify also the search direction based on previous iterations and often perform a line search along that direction based on a finite difference step. These methods have a potential to faster convergence to the minimum energy path especially when the convergence threshold is tight but may be unstable during the early phases of the optimization (Sheppard et al., 2008). Better convergence properties can be achieved also by using the fast inertial relaxation engine (Bitzek et al., 2006), shortened as FIRE, which is an extension of the VPO algorithm involving adaptive time steps and additional modifications to the velocity. Since acquisition of the second derivatives of energy is usually too expensive, the optimization algorithm is required to be based only on the first derivatives. In case accurate observations of the second derivatives were easily available, the NEB optimization could be done efficiently with a Newton-Raphson method using analytical calculations of the derivatives of the NEB forces (Bohner et al., 2013).

3.2 Dimer method

The dimer method (Henkelman and Jónsson, 1999) is an example of a minimum mode following algorithm with the objective to find a saddle point by following the direction of the lowest curvature on the energy surface without knowing the final state of the transition. Inspired by the idea of Voter (1997), the lowest curvature mode is found by rotating a dimer consisting of a pair of images, \mathbf{R}_1 and \mathbf{R}_2 , with respect to its middle point \mathbf{R}_0 , and the whole dimer is then translated towards the saddle point based on a force vector where the component parallel to the direction of the dimer is reverted to point towards the direction of increasing energy, similarly as in the CI-NEB method. During the algorithm, rotation and translation phases alternate until the magnitude of the translational force is below some convergence threshold. Figure 3.2 shows a simple two-dimensional

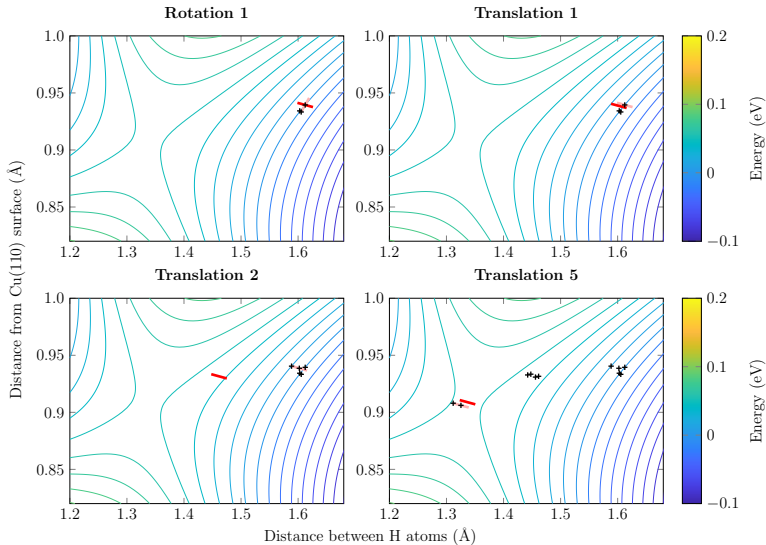


Figure 3.2. Progression of the dimer method in a simple example where two hydrogen atoms are free to move near a fixed copper surface (Mills et al., 1995). The saddle point and the initial dimer coincide with the same two-dimensional cut of the coordinate space as shown in figure 2.2. The pink and red bars represent the dimer before and after the rotation or translation, respectively, and the black crosses represent locations where energy and its first derivatives have been evaluated during the process. In this case, the orientation of the dimer after the first rotation turns out to be close enough to the lowest curvature mode of the saddle point so that no further rotations are needed.

illustration of the progression of the dimer method in the same system as shown in figure 2.2.

The rotations towards the lowest curvature mode are based on the atomic force vectors $\mathbf{F}(\mathbf{R}_1) = -\nabla E(\mathbf{R}_1)$ and $\mathbf{F}(\mathbf{R}_2) = -\nabla E(\mathbf{R}_2)$, given by the negative energy gradient at the two images. The distance from the middle point \mathbf{R}_0 to \mathbf{R}_1 and \mathbf{R}_2 , referred to as the dimer separation $\Delta_{\mathbf{R}}$, is fixed to a small value in order to estimate the second derivative of energy along the dimer as accurately as possible. The direction of the lowest curvature corresponds to the orientation where the dimer energy, defined as $E(\mathbf{R}_1) + E(\mathbf{R}_2)$, is minimized. The minimum curvature mode is thus found by zeroing a scaled rotational force given as

$$\mathbf{F}_{\text{rot}} = \frac{\mathbf{F}^{\perp}(\mathbf{R}_1) - \mathbf{F}^{\perp}(\mathbf{R}_2)}{\Delta_{\mathbf{R}}}, \quad (3.8)$$

where

$$\mathbf{F}^{\perp}(\mathbf{R}_i) = \mathbf{F}(\mathbf{R}_i) - (\mathbf{F}(\mathbf{R}_i) \cdot \hat{\mathbf{N}})\hat{\mathbf{N}} \quad (3.9)$$

is the component of $\mathbf{F}(\mathbf{R}_i)$ perpendicular to the orientation vector $\hat{\mathbf{N}}$, which is a unit vector pointing from \mathbf{R}_0 towards \mathbf{R}_1 . Instead of evaluating the force at both \mathbf{R}_1 and \mathbf{R}_2 between subsequent rotations, it is more efficient

to evaluate the force at the fixed middle point \mathbf{R}_0 and extrapolate the force at \mathbf{R}_2 as $\mathbf{F}(\mathbf{R}_2) \approx 2\mathbf{F}(\mathbf{R}_0) - \mathbf{F}(\mathbf{R}_1)$, as suggested by Olsen et al. (2004).

The rotational plane for each rotation iteration is spanned by the orientation vector $\hat{\mathbf{N}}$ and another unit vector $\hat{\mathbf{\Omega}}$, which defines the direction of the rotation for \mathbf{R}_1 . The steepest descent direction is simply the direction of the rotational force: $\hat{\mathbf{\Omega}} = \mathbf{F}_{\text{rot}}/|\mathbf{F}_{\text{rot}}|$. It is also possible to use a nonlinear conjugate gradient (Fletcher and Reeves, 1964; Polak and Ribière, 1969; Henkelman and Jónsson, 1999) or a more efficient L-BFGS (Nocedal, 1980; Liu and Nocedal, 1989; Kästner and Sherwood, 2008) approach, where $\hat{\mathbf{\Omega}}$ is modified based on previous rotation iterations.

In the original formulation (Henkelman and Jónsson, 1999), a small preliminary step with a rotation angle ω^* is first taken to get a finite difference approximation to the derivative of the rotational force, and the optimal rotation angle ω is then obtained based on a local quadratic approximation to the energy surface. Heyden et al. (2005) prefer a larger preliminary rotation instead of a finite difference step in order to avoid possible problems with noisy data. They suggest the following rough estimate, used also in Publication IV, for the preliminary rotation angle:

$$\omega^* = \frac{1}{2} \arctan \frac{(\mathbf{F}(\mathbf{R}_1) - \mathbf{F}(\mathbf{R}_0)) \cdot \hat{\mathbf{\Omega}}}{\Delta_{\mathbf{R}} |C|}, \quad (3.10)$$

where

$$C = (\mathbf{F}(\mathbf{R}_0) - \mathbf{F}(\mathbf{R}_1)) \cdot \hat{\mathbf{N}} / \Delta_{\mathbf{R}} \quad (3.11)$$

is an estimate for the curvature of energy along the dimer. The dimer orientation and rotation direction after the preliminary rotation step are given by

$$\hat{\mathbf{N}}^* = \hat{\mathbf{N}} \cos \omega^* + \hat{\mathbf{\Omega}} \sin \omega^* \quad (3.12)$$

and

$$\hat{\mathbf{\Omega}}^* = -\hat{\mathbf{N}} \sin \omega^* + \hat{\mathbf{\Omega}} \cos \omega^*, \quad (3.13)$$

and the force $\mathbf{F}(\mathbf{R}_1^*)$ is then evaluated at $\mathbf{R}_1^* = \mathbf{R}_0 + \Delta_{\mathbf{R}} \hat{\mathbf{N}}^*$. Based on a local quadratic approximation, the final rotation angle that minimizes the dimer energy on the rotational plane is given as

$$\omega = \begin{cases} \frac{1}{2} \arctan \frac{b_1}{a_1}, & \text{if } \frac{b_1}{a_1} \geq 0 \\ \frac{1}{2} \arctan \frac{b_1}{a_1} + \frac{\pi}{2}, & \text{if } \frac{b_1}{a_1} < 0, \end{cases} \quad (3.14)$$

where

$$b_1 = (\mathbf{F}(\mathbf{R}_0) - \mathbf{F}(\mathbf{R}_1)) \cdot \hat{\mathbf{\Omega}} / \Delta_{\mathbf{R}} \quad (3.15)$$

and

$$a_1 = \frac{b_1 \cos(2\omega^*) - (\mathbf{F}(\mathbf{R}_0) - \mathbf{F}(\mathbf{R}_1^*)) \cdot \hat{\mathbf{\Omega}}^* / \Delta_{\mathbf{R}}}{\sin(2\omega^*)}. \quad (3.16)$$

The new dimer orientation after the rotation is given by

$$\hat{\mathbf{N}}^{\text{new}} = \hat{\mathbf{N}} \cos \omega + \hat{\mathbf{\Omega}} \sin \omega \quad (3.17)$$

and the new \mathbf{R}_1 by

$$\mathbf{R}_1^{\text{new}} = \mathbf{R}_0 + \Delta_{\mathbf{R}} \hat{\mathbf{N}}^{\text{new}}. \quad (3.18)$$

In some implementations, no more than one rotation iteration is performed between the translation steps. The rotation iterations can be also repeated until rotational convergence, defined based on the preliminary or final rotation angle or the magnitude of rotational force, or until some maximum number of consecutive rotations is reached. In that case, the number of force evaluations between consecutive rotation iterations can be reduced by extrapolating $\mathbf{F}(\mathbf{R}_1^{\text{new}})$ from $\mathbf{F}(\mathbf{R}_0)$, $\mathbf{F}(\mathbf{R}_1)$, and $\mathbf{F}(\mathbf{R}_1^*)$ and using this estimate when calculating the rotational force for the following rotation iteration (Kästner and Sherwood, 2008).

After each rotation phase, a translation step is performed to move the middle point of the dimer towards the saddle point. The nature of the translation step depends on the curvature along the current orientation vector $\hat{\mathbf{N}}$, estimated either by the quadratic approximation (Olsen et al., 2004; Heyden et al., 2005) or equation 3.11. If the curvature is positive, the dimer is assumed to be in a convex region with positive second derivatives of energy in all directions, which is often the case if the start point is chosen to be close to an minimum energy point. In this case, a step with some predefined length is taken to the direction of increasing energy along $\hat{\mathbf{N}}$ to make the dimer climb up from the convex region. If the curvature along the dimer is negative, the translational force is obtained by reverting the component of $\mathbf{F}(\mathbf{R}_0)$ parallel to the dimer:

$$\mathbf{F}_{\text{trans}} = \mathbf{F}(\mathbf{R}_0) - 2\mathbf{F}^{\parallel}(\mathbf{R}_0), \quad (3.19)$$

where

$$\mathbf{F}^{\parallel}(\mathbf{R}_0) = (\mathbf{F}(\mathbf{R}_0) \cdot \hat{\mathbf{N}}) \hat{\mathbf{N}}. \quad (3.20)$$

This allows the dimer to climb upwards on the energy surface following the direction of the minimum curvature mode mode while minimizing the energy in directions perpendicular to the dimer. The displacement of \mathbf{R}_0 can be performed using any gradient-based optimization approach, including nonlinear conjugate gradient (Fletcher and Reeves, 1964; Polak and Ribière, 1969) and L-BFGS (Nocedal, 1980; Liu and Nocedal, 1989) algorithms. Similarly as in the rotation phase, a preliminary step can be taken to estimate a proper step length for the translation. In the L-BFGS approach, however, a good estimate for the translation step length is provided by an inverse Hessian approximated implicitly based on information stored during previous translation iterations. As suggested by Kästner and Sherwood (2008), the L-BFGS approach is applied to both translations and rotations in the GP-dimer algorithm presented in Publication IV.

4. Summary of contributions

The contribution of this dissertation consists of development and testing of two algorithms that utilize Gaussian process regression to enhance searches of saddle points and minimum energy paths. The GP-NEB algorithm aims to find a minimum energy path between two known minimum energy configurations and the saddle point located at the maximum point of the path, whereas the GP-dimer algorithm only searches for the saddle point starting somewhere from its vicinity.

4.1 GP-NEB algorithm

The general idea of using machine learning methods to enhance saddle point search algorithms has been introduced by Peterson (2016), who applied artificial neural networks to nudged elastic band calculations. In the iterative procedure, a minimum energy path is first found on an approximate energy surface based on a machine learning model, and accurate evaluations of energy and its first derivatives are then performed to check if the path has converged also on the accurate energy surface. If the convergence criteria are not satisfied, the new observations are included in the training data set, the machine learning model is updated, and the path is relaxed again on the approximate energy surface. The iterations are repeated until final convergence is confirmed by accurate evaluations. The advantage of this approach is based on the assumption that the accurate evaluations are significantly more expensive than training of the machine learning model or approximation of energy and its derivatives based on the model. By performing the path relaxation on the approximate energy surface, the total number of accurate evaluations required for convergence can be reduced and the minimum energy path search hence accelerated.

Publication I presents an initial step in the development of a similar algorithm where Gaussian process regression is applied instead of neural networks as a machine learning approach to model the energy surface. Whereas optimization of a large number of weights of a neural network

model may be challenging due to many local minima of the cost function, optimization of the hyperparameters of a Gaussian process model is a much easier task. As described in section 2.4, Gaussian process models allow straightforward ways to handle derivatives, which is beneficial when learning from the derivative observations and predicting NEB forces for the path relaxation. Furthermore, GP models have been shown to perform well especially when learning from small training data sets, which makes them appealing for this application. In Publication I, the feasibility of the GP-NEB approach is demonstrated for three simple benchmark transitions, where two atoms of an heptamer island move to adjacent sites on the (111) surface of a FCC (face-centred cubic) crystal (Henkelman, Jóhannesson, and Jónsson, 2000; Chill et al., 2014). As compared with a regular NEB method, the number of evaluations required for convergence to the minimum energy path is decreased to less than fifth with a simple implementation of the GP-NEB algorithm using the stationary squared exponential covariance function k_x (see equation 2.29 in section 2.5).

Publication II extends the GP-NEB method to CI-NEB calculations and presents detailed descriptions for two variants of the algorithm. The simpler one, referred to as the all-images-evaluated (AIE) algorithm, follows the original idea of Peterson (2016) by evaluating accurate energy and its first derivatives at all intermediate images of the NEB path relaxed on the approximate energy surface. Figure 4.1 shows the progression of the AIE algorithm in the same two-dimensional task as shown in figure 3.1 for the regular CI-NEB method. Knowing the coordinates of the initial path, $\mathbf{R}_i, i = 0, 1, \dots, N_{\text{im}} - 1$, and accurate energy and its (zero) gradient at the two end points, $E(\mathbf{R}_0)$, $\nabla E(\mathbf{R}_0)$, $E(\mathbf{R}_{N_{\text{im}}-1})$, and $\nabla E(\mathbf{R}_{N_{\text{im}}-1})$, the algorithm is started by evaluating $E(\mathbf{R}_i)$ and $\nabla E(\mathbf{R}_i)$ at the intermediate images $\mathbf{R}_i, i = 1, \dots, N_{\text{im}} - 2$. Based on these data, a GP model for the energy surface is trained by optimizing the hyperparameters of the covariance function, and a CI-NEB calculation is performed using the mean of the

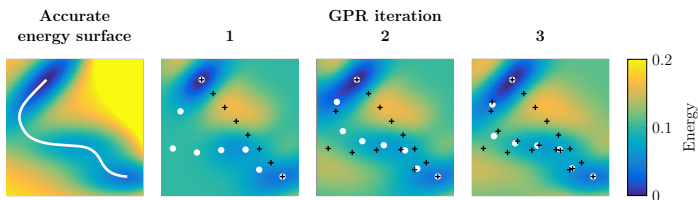


Figure 4.1. Progression of the simpler all-images-evaluated (AIE) version of the GP-NEB algorithm on a two-dimensional Müller-Brown energy surface (Müller and Brown, 1979). The white dots represent images of the relaxed CI-NEB path on an approximate energy surface obtained by GP regression. After each GPR iteration, final convergence of the path is checked by evaluating accurate energy and its first derivatives at all intermediate images, and those observations are then added to the training data set (observed locations marked with black crosses). Figure reproduced with permission from Publication II.

posterior predictive distribution of energy and its derivatives (see equations 2.25 and 2.27 in section 2.4) when calculating the NEB forces. After the CI-NEB path has relaxed on the approximate energy surface, new energy and gradient evaluations are made at the intermediate images of the relaxed path and added to the training data set for the following GPR iteration. The algorithm is continued until final convergence criteria for the accurate NEB forces are satisfied after three GPR iterations and a total of 24 evaluations.

Due to the probabilistic nature of Gaussian process regression, the predictions of energy and its derivatives are expressed as probability distributions. The more advanced variant of GP-NEB, referred to as the one-image-evaluated (OIE) algorithm, utilizes the variance of the posterior distribution of energy (see equation 2.26 in section 2.4) as a measure of uncertainty to direct the evaluations to locations where they are most useful. According to the main rule, accurate energy and derivatives are evaluated only at the image with the highest uncertainty before updating the GP model and relaxing the path. However, since confirmation of the final convergence requires accurate energy gradient to be known for all images of the path, also the other intermediate images are included in the evaluations one by one without moving the path as long as there is a chance that final convergence might have been reached based on the mixture of accurate and approximated NEB forces. Since the convergence criterion may be tighter for the climbing image and since its location affects

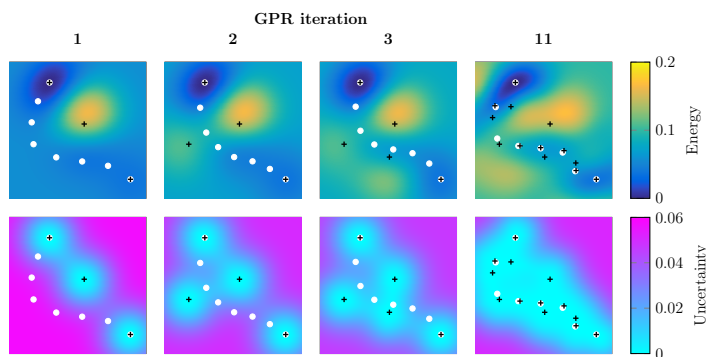


Figure 4.2. Progression of the more advanced one-image-evaluated (OIE) version of the GP-NEB algorithm on a two-dimensional Müller-Brown energy surface (Müller and Brown, 1979). The white dots represent images of the relaxed CI-NEB path on an approximate energy surface obtained by GP regression. The lower panel shows the standard deviation of the posterior distribution of energy representing the uncertainty of the predictions according to the GP model. After GPR iterations 1, 2, and 3, accurate energy and its first derivatives are evaluated at the image with the largest uncertainty, and the information is then added to the training data set (observed locations marked with black crosses). After GPR iteration 11, the path is not moved anymore but the final convergence is confirmed by accurate evaluations at each of the intermediate images. Figure reproduced with permission from Publication II.

on the distribution of the other images, the climbing image is favoured over other images in the evaluation order of the convergence check. Figure 4.2 shows the progression of the OIE algorithm in the two-dimensional example task. The final shape of the path is here obtained after eleven energy and gradient evaluations, and the final convergence is then confirmed by six more evaluations.

In Publication II, the two variants of the GP-NEB algorithm are tested in CI-NEB calculations for the whole heptamer island benchmark (Henkelman, Jóhannesson, and Jónsson, 2000; Chill et al., 2014) including thirteen transitions. As compared with a regular CI-NEB method, the number of evaluations required for convergence to the minimum energy path is decreased by an order of magnitude. The OIE algorithm reduces the number of evaluations to about a half of what is required by the AIE algorithm. As an additional test feature, information about the second derivatives of energy at the two end points are included by adding finite difference data points in the initial training data set. Since the Hessian of energy is often evaluated anyway at the initial and final states of the transition when calculating transition rates using the harmonic approximation to the transition state theory (Vineyard, 1957), these evaluations may be considered available without additional effort. The use of the Hessian data reduces the number of observations by about 20% when using the AIE algorithm, but the effect is smaller for the OIE algorithm.

In Publications I and II, a simple GP model with a stationary squared exponential covariance function k_x is successfully used in the GP-NEB calculations. In some systems, however, the stationarity of the covariance function with respect to the atom coordinates may lead to problems as illustrated in section 2.5. In Publication III, these problems are avoided by defining a modified covariance function $k_{1/r}$ where the difference measure fed to the squared exponential covariance function is based on differences in the inverse interatomic distances (see equation 2.34 in section 2.5). This difference measure stretches when atoms are closer to each other, which makes it easier to model large repulsive forces. In addition, the more informative covariance structure allows more efficient learning of the potential energy surface.

Even though the modified GP model handles well also large repulsive forces, avoiding unphysical configurations and constraining the exploration of uncertain regions may still stabilize the algorithm. Another modification introduced in Publication III concerns the early stopping criteria that define the allowed region for the images of the path during the NEB relaxation phase. The early stopping criterion used in Publication II is based on the distance to the nearest observed data point according to the regular difference measure \mathcal{D}_x (see equation 2.30 in section 2.5) with the limit set to a half of the length of the initial path. If the limit is exceeded, then the last step of the NEB relaxation phase is rejected, the

relaxation phase is stopped, and the following evaluation is performed at the image that violated the condition. In Publication III, an additional early stopping criterion is introduced based on relative changes in the interatomic distances. The condition requires that for each image of the current path, there exists an observed data point with all interatomic distances between $2/3$ and $3/2$ of the corresponding distance in the current image. Accompanied with the inverse-distance covariance function $k_{1/r}$,

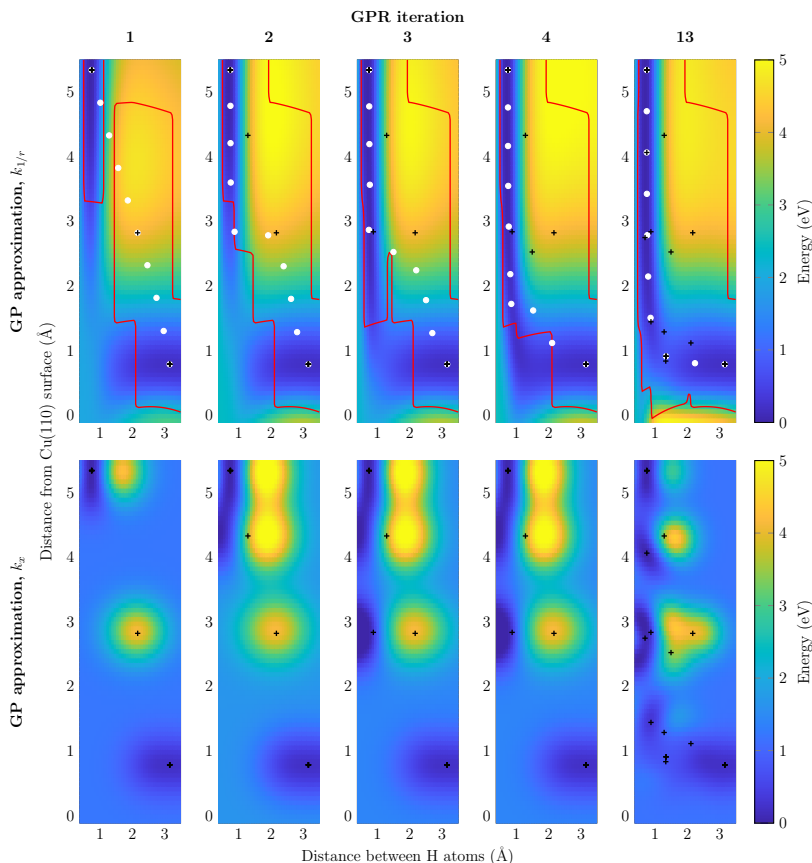


Figure 4.3. A two-dimensional cut through the potential energy surface for a pair of hydrogen atoms near a fixed copper surface (Mills et al., 1995). The upper panel shows the progression of the modified GP-NEB algorithm. The white dots are projections of the images of the relaxed CI-NEB path on an approximate energy surface obtained by GP regression with the inverse-distance covariance function $k_{1/r}$, and the red line shows the border of the allowed region defined by the accompanying early stopping criterion. The black crosses are projections of the training data points. In the first four GPR iterations, the NEB relaxation phase is terminated by the early stopping rule, and the final path is obtained after thirteen GPR iterations. For comparison, the lower panel shows GP approximations with the stationary covariance function k_x using the same training data sets as in the upper panel. Figure reproduced with permission from Publication III.

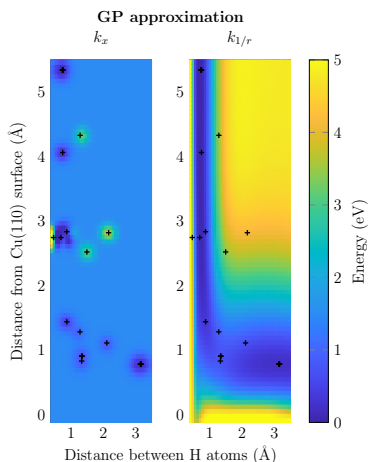


Figure 4.4. GP approximations based on covariance functions k_x and $k_{1/r}$ corresponding to the rightmost graphs in figure 4.3 with a high-gradient data point close to the left border of the graph added to the training data set. Figure reproduced with permission from Publication III.

this criterion practically prevents the distance between two atoms from becoming shorter than $2/3$ of the unknown bond length (notice that this does not apply with a stationary covariance function).

The upper panel of figure 4.3 shows the progression of the modified GP-NEB algorithm in a CI-NEB calculation for a dissociation of a hydrogen molecule on a fixed copper surface (Mills et al., 1995). The initial and final states coincide with the same two-dimensional cut of the six-dimensional coordinate space as shown in figure 2.2. The GP approximation based on the inverse-distance covariance function looks quite realistic already in the beginning, when the training data include the energy and its first derivatives at one intermediate image and the two end points in addition to the Hessian data at the end points. Since the third image of the initial path is outside the allowed regions, the early stopping rule is triggered already before moving the images, and also the NEB relaxations in the following three GPR iterations are terminated by the early stopping rule. The final convergence is confirmed after nineteen energy and gradient evaluations, whereas a regular CI-NEB calculation requires about 500 evaluations.

For comparison, the lower panel of figure 4.3 shows what the GP approximation with the same training data would look like if the stationary squared exponential covariance function k_x was used instead of the inverse-distance covariance function $k_{1/r}$. Since the stationary GP model extrapolates the attractive forces acting on the hydrogen atoms to regions where the atoms collide, it would be difficult to keep the images away from regions of large repulsive forces without a too restrictive stopping rule. As

shown in figure 4.4, an additional data point from the repulsive region would make interpolation of the training data set more difficult for the stationary model and lead to a short length scale. With the inverse-distance covariance function, the additional data point would not cause problems.

In addition to demonstrations in systems that are challenging for stationary GP models, Publication III reports results also for the heptamer island benchmark for which the squared exponential covariance function k_x works well. Figure 4.5 shows the average number of energy and gradient evaluations required for GP-NEB calculations with a varying number of degrees of freedom. Depending on the algorithm variant, the inverse-distance covariance function with the accompanying early stopping criterion reduces the number of energy and force evaluations by about 30–50% when compared with the squared exponential covariance function.

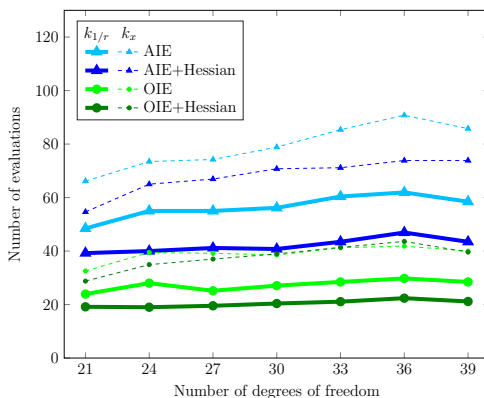


Figure 4.5. Number of energy and gradient evaluations required for convergence of CI-NEB calculations in a heptamer island benchmark (Henkelman, Jóhannesson, and Jónsson, 2000; Chill et al., 2014) with variants of the GP-NEB algorithm. The average over thirteen different transitions is presented as a function of the number of degrees of freedom, increased by allowing a larger number of substrate atoms to move. The narrow dashed lines present results with the stationary squared exponential covariance function k_x , and the thick solid lines present the corresponding results when using the inverse-distance covariance function $k_{1/r}$ with the accompanying stopping criterion. The blue triangles represent the all-images-evaluated (AIE) algorithm, and the green dots represent the one-image-evaluated (OIE) algorithm. The use of Hessian data at the two end points is indicated by darker colour. Figure reproduced with permission from Publication III.

4.2 GP-dimer algorithm

Publication IV applies the Gaussian process regression approach used in the GP-NEB algorithm to the dimer method in a saddle point search task where only a start point is known. A similar general scheme connecting Gaussian process regression with surface walking methods has

been recently applied by Denzel and Kästner (2018b), who use a stationary Matérn-5/2 covariance function to build a computationally efficient multi-level Gaussian process model (Denzel and Kästner, 2018a). The GP-dimer algorithm in Publication IV applies the more expressive inverse-distance covariance function $k_{1/r}$ coupled with the robust stopping criterion as suggested for the GP-NEB algorithm in Publication III, and the performance is compared with corresponding results obtained with stationary covariance functions.

With the middle point \mathbf{R}_0 of the initial dimer set to the given start point and images \mathbf{R}_1 and \mathbf{R}_2 aligned with the a possibly randomized start orientation, the GP-dimer algorithm is started by evaluating accurate energy and its gradient at \mathbf{R}_0 and \mathbf{R}_1 , i.e., $E(\mathbf{R}_0)$, $\nabla E(\mathbf{R}_0)$, $E(\mathbf{R}_1)$, and $\nabla E(\mathbf{R}_1)$. If no information is available about the energy surface or the direction of the lowest energy curvature at the start point, it is useful to perform initial rotations with accurate evaluations before translating the dimer. In the GP-dimer algorithm, this initial phase is performed by repeated initial rotation rounds on an approximate energy surface obtained by GP regression based on the evaluations made so far. During each initial rotation round, the direction of the lowest curvature on the approximate energy surface is found according to a regular rotation scheme using the mean of the posterior predictive distribution of the energy gradient to calculate the rotational force (see equation 2.27 in section 2.4), and accurate energy $E(\mathbf{R}_1)$ and gradient $\nabla E(\mathbf{R}_1)$ are then evaluated at the new location of \mathbf{R}_1 . The initial rotation phase is stopped when the preliminary rotation angle ω^* (see equation 3.10 in section 3.2) based on the accurate gradients $\nabla E(\mathbf{R}_0)$ and $\nabla E(\mathbf{R}_1)$ or the angle between the relaxed orientations of two subsequent rounds is below a given threshold. As shown in Publication IV, this approach requires fewer evaluations for rotational convergence than regular rotation schemes. A similar initial phase where GP regression is utilized to find the direction of the lowest energy curvature is applied also by Denzel and Kästner (2018b).

In the actual GPR iterations started after the initial rotation phase, the dimer is both rotated and translated based on the GP approximation. During each GPR iteration, a saddle point on the approximate energy surface is found according to a regular dimer method, and final convergence of the dimer is then checked by evaluating accurate energy $E(\mathbf{R}_0)$ and gradient $\nabla E(\mathbf{R}_0)$ at the middle point \mathbf{R}_0 of the relaxed dimer. Figure 4.6 shows the progression of the GP-dimer algorithm in the same example task as shown in figure 3.2 for the regular dimer method. In this simple example, the direction of the lowest curvature of the accurate energy surface is found after two initial rotation rounds, and a saddle point close to the correct location is formed on the approximate energy surface based only on the four data points around the start point. After evaluating the accurate energy and gradient at this predicted saddle point, the GP

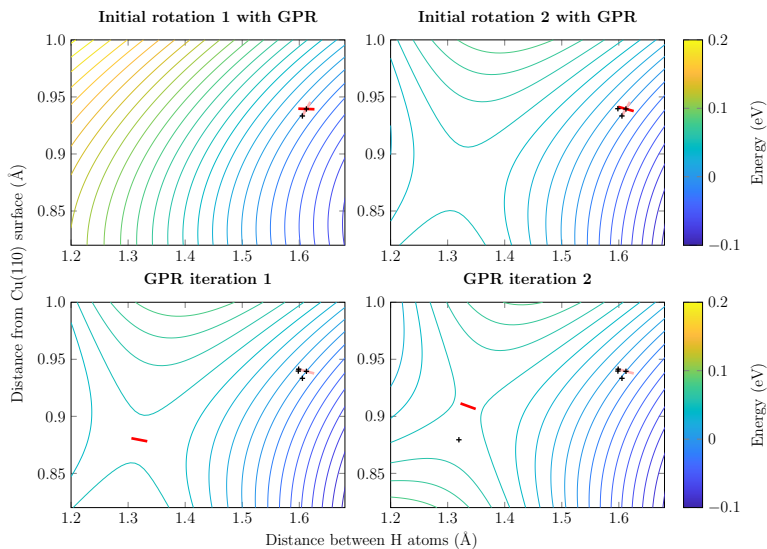


Figure 4.6. Progression of the GP-dimer algorithm in a simple example where two hydrogen atoms are free to move near a fixed copper surface (Mills et al., 1995). The saddle point and the initial dimer coincide with the same two-dimensional cut of the coordinate space as shown in figure 2.2. The pink and red bars represent the dimer in the beginning and end of the initial rotation round or GPR iteration, respectively. During the two rotation rounds, the orientation of the dimer is aligned with the direction of the lowest curvature on an approximate energy surface obtained by GP regression. After each round, accurate energy and its first derivatives are evaluated at one of the images of the dimer, and the information is then added to the training data set (observed locations marked with black crosses). In the actual GPR iterations started after reaching rotational convergence at the start point, the dimer is both rotated and translated to find a saddle point on the approximate energy surface, and final convergence is then checked by evaluating accurate energy and its first derivatives at the middle point of the relaxed dimer. Figure reproduced with permission from Publication IV.

approximation becomes accurate enough for convergence to the correct saddle point.

In addition to the dissociative adsorption of a hydrogen molecule on a copper surface (Mills et al., 1995), used also in GP-NEB calculations in Publication III, the tests of the GP-dimer algorithm in Publication IV involve three gas phase chemical reactions (Birkholtz and Schlegel, 2015) with saddle point configurations illustrated in figure 4.7 alongside the corresponding result graphs. A set of start points is chosen randomly with a varying distance from the saddle point of each example reaction, and the number of energy and gradient evaluations required for convergence is reported for two variants of the regular dimer method and for the GP-dimer algorithm with the inverse-distance covariance function $k_{1/r}$ and stationary squared exponential (k_x) and Matérn-5/2 ($k_x^{M-5/2}$) covariance functions. As shown in figure 4.7, the variants of GP-dimer require fewer evaluations than the regular methods, and the difference increases when

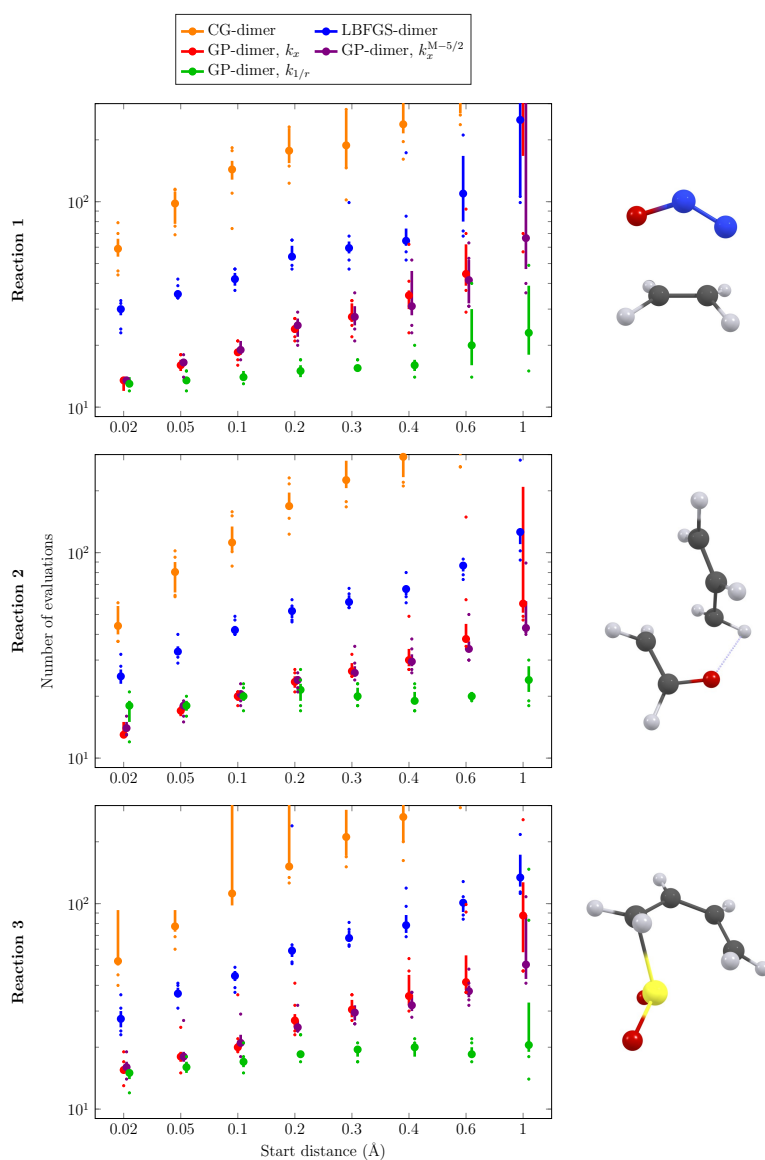


Figure 4.7. Number of energy and gradient evaluations required for convergence with a regular dimer method, based on conjugate gradients (Heyden et al., 2005) or L-BFGS (Kästner and Sherwood, 2008), and the GP-dimer algorithm using the inverse-distance ($k_{1/r}$), squared exponential (k_x), or Matérn-5/2 ($k_x^{M-5/2}$) co-variance function. The saddle point configuration of each of the three example reactions (Birkholtz and Schlegel, 2015) is visualized with the following atom colours: C, dark gray; H, light gray; O, red; N, blue; S, yellow. The distance of the start point from this configuration is shown on the horizontal axis. The large dots present the median number of evaluations among ten randomly chosen start positions, the bars present the interval between the third and eighth largest numbers, and the two smallest and largest numbers are presented with small dots. Figure reproduced with permission from Publication IV.

the start point is farther from the example saddle point. With start points closer than 0.1 Å to the saddle point, only small differences are observed between the three variants of the GP-dimer algorithm, but the benefits of using the inverse-distance covariance function become apparent with larger start distances.

5. Discussion

This dissertation presents the first steps in utilizing Gaussian process regression to enhance saddle point search algorithms on potential energy surfaces. In the GP-NEB algorithm, a minimum energy path between two known minimum energy configurations and a saddle point located at the maximum point of the path are found on an approximate energy surface based on a Gaussian process model, which is updated with accurate observations of energy and its derivatives until convergence of the path on the accurate energy surface is confirmed. In the GP-dimer algorithm, a similar approach is applied to minimum mode following calculations, where only a start point for a saddle point search is given in the beginning. Based on simple test examples, the Gaussian process regression approach may reduce the required number of accurate energy and force evaluations by an order of magnitude when compared with conventional methods.

In Gaussian process regression, the predictions of energy and its derivatives are expressed as probability distributions obtained as a result of Bayesian inference. The variance of the predictive distribution can be utilized in the GP-NEB algorithm as an uncertainty estimate when selecting new observation points from the discretized path. This approach has similarities with Bayesian optimization, where an acquisition function based on the predictive distribution of the objective function is defined for the selection of observation points in a global optimization task (Shahriari et al., 2016). A major difference is that saddle point search algorithms are typically satisfied with a local type of convergence, which means that exploration of uncertain regions far from the predicted minimum energy path or saddle point is not necessary. When the task is to find a minimum energy path with convergence confirmed by accurate evaluations at all points of the discretized path, it is often most efficient to select one of those points as the new observation point. However, if accurate convergence is important only for the saddle point, the convergence of the rest of the path can be defined based on the estimated uncertainty without restricting to any number of discretization points (Garrido Torres et al., 2019). The search for the energy maximum along the path can be then defined as

a Bayesian optimization problem with various possible choices for the acquisition function.

Besides algorithmic development, the dissertation shows that automated and accurate modelling of the surroundings of a minimum energy path is possible with rather simple Gaussian process models. While stationary covariance functions with similar properties in all parts of the space of atom coordinates turn out to be insufficient in many systems involving large repulsive forces, a good representation can be obtained by defining the difference measure between two configurations based on inverse inter-atomic distances. More sophisticated descriptors designed for modelling entire potential energy surfaces are often based on approximation of the total energy of the system with a sum over local atomic environments and may require larger noise variance to be assumed for the observations (Bartók et al., 2010; Bartók and Csányi, 2015). With reduced convergence requirements, however, such models may provide useful properties also for the GP-NEB and GP-dimer algorithms.

The advantage of the Gaussian process regression approach to saddle point searches relies on the assumption that the accurate energy and gradient evaluations are significantly more expensive than predictions based on the Gaussian process model or training of the model. In large systems, however, the applicability of the approach is limited due to the poor scaling of the computational cost of Gaussian process regression with respect to the number of training observations. Since the number of available derivative observations depends on the number of moving atoms, the computational cost increases fast with the system size if full advantage is taken of the derivative information. While many of the attempts to make Gaussian process models more applicable to large data sets rely on approximations, recent development on exact Gaussian process inference is reducing the training cost from cubic to quadratic without compromising the accuracy (Gardner et al., 2018; Wang et al., 2019). This sort of advancement paves the way for further development of efficient saddle point search methods.

Bibliography

- Andersen, H. C. (1980). Molecular dynamics simulations at constant pressure and/or temperature. *J. Chem. Phys.*, volume 72, issue 4, pages 2384–2393.
- Ásgeirsson, V. and Jónsson, H. (2018). Exploring potential energy surfaces with saddle point searches. In Andreoni, W. and Yip, S. (editors), *Handbook of Materials Modeling: Methods: Theory and Modeling*, Springer: Cham.
- Banerjee, A., Adams, N., Simons, J., and Shepard, R. (1985). Search for stationary points on surfaces. *J. Phys. Chem.*, volume 89, issue 1, pages 52–57.
- Bartók, A. P. and Csányi, G. (2015). Gaussian approximation potentials: a brief tutorial introduction. *Int. J. Quantum Chem.*, volume 115, issue 16, pages 1051–57.
- Bartók, A. P., Payne, M. C., Condor, R., and Csányi, G. (2010). Gaussian approximation potentials: the accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.*, volume 104, issue 13, article 136403.
- Birkholz, A. B. and Schlegel, H. B. (2015). Using bonding to guide transition state optimization. *J. Comput. Chem.*, volume 36, issue 15, pages 1157–1166.
- Bitzek, E., Koskinen, P., Gähler, F., Moseler, M., and Gumbusch, B. (2006). Structural relaxation made simple. *Phys. Rev. Lett.*, volume 97, issue 17, article 170201.
- Blight, B. J. N. and Ott, L. (1975). A Bayesian approach to model inadequacy for polynomial regression. *Biometrika*, volume 62, issue 1, pages 79–88.
- Bohner, M. U., Meisner, J., and Kästner, J. (2013). A quadratically-converging nudged elastic band optimizer. *J. Chem. Theory Comput.*, volume 9, issue 8, pages 3498–3504.
- Cerjan, C. J. and Miller, W. H. (1981). On finding transition states. *J. Chem. Phys.*, volume 75, issue 6, pages 2800–2806.
- Chill, S. T., Stevenson, J., Ruhle, V., Shang, C., Xiao, P., Farrell, J. D., Wales, D. J., and Henkelman, G. (2014). Benchmarks for characterization of minima, transition states and pathways in atomic, molecular, and condensed matter systems. *J. Chem. Theory Comput.*, volume 10, issue 12, pages 5476–5482.

- Csató, L. and Opper, M. (2002). Sparse on-line Gaussian processes. *Neural Comput.*, volume 14, issue 3, pages 641–668.
- Deisenroth, M. P. and Ng, J. W. (2015). Distributed Gaussian processes. In Bach, F. and Blei, D. (editors), *Proceedings of the Thirty-Second International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1481–1490.
- Denzel, A. and Kästner, J. (2018a). Gaussian process regression for geometry optimization. *J. Chem. Phys.*, volume 148, issue 9, article 94114.
- Denzel, A. and Kästner, J. (2018b). Gaussian process regression for transition state search. *J. Chem. Theory Comput.*, volume 14, issue 11, pages 5777–5786.
- E, W., Ren, W., and Vanden-Eijnden, E. (2002). String method for the study of rare events. *Phys. Rev. B*, volume 66, issue 5, article 52301.
- E, W., Ren, W., and Vanden-Eijnden, E. (2007). Simplified and improved string method for computing the minimum energy paths in barrier-crossing events. *J. Chem. Phys.*, volume 126, issue 16, article 164103.
- Fletcher, R. and Reeves, C. M. (1964). Function minimization by conjugate gradients. *Comput. J.*, volume 7, issue 2, pages 149–154.
- Gardner, J. R., Pleiss, G., Weinberger, K. Q., Bindel, D., and Wilson, A. G. (2018). GPYtorch: blackbox matrix-matrix Gaussian process inference with GPU acceleration. In Bengio, S., Wallach, H. M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (editors), *Advances in Neural Information Processing Systems 31*, Curran Associates: Red Hook, pages 7576–7586.
- Garrido Torres, J. A., Jennings, P. C., Hansen, M. H., Boes, J. R., and Bligaard, T. (2019). Low-scaling algorithm for nudged elastic band calculations using a surrogate machine learning model. *Phys. Rev. Lett.*, volume 122, issue 15, article 156001.
- Gibbs, M. N. and MacKay, D. J. C. (2000). Variational Gaussian process classifiers. *IEEE Trans. Neural Netw.*, volume 11, issue 6, pages 1458–1464.
- Henkelman, G., Jóhannesson, G. H., and Jónsson, H. (2000). Methods for finding saddle points and minimum energy paths. In Schwartz, S. D. (editor), *Theoretical Methods in Condensed Phase Chemistry*, Kluwer Academic: New York, pages 269–300.
- Henkelman, G. and Jónsson, H. (1999). A dimer method for finding saddle points on high dimensional potential surfaces using only first derivatives. *J. Chem. Phys.*, volume 111, issue 15, pages 7010–7022.
- Henkelman, G. and Jónsson, H. (2000). Improved tangent estimate in the nudged elastic band method for finding minimum energy paths and saddle points. *J. Chem. Phys.*, volume 113, issue 22, pages 9978–9985.
- Henkelman, G., Uberuaga, B. P., and Jónsson, H. (2000). A climbing image nudged elastic band method for finding saddle points and minimum energy paths. *J. Chem. Phys.*, volume 113, issue 22, pages 9901–9904.

- Hensman, J., Fusi, N., and Lawrence, N. D. (2013). Gaussian processes for big data. In Nicholson, A. and Smyth, P. (editors), *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, AUAI Press: Corvallis, pages 282–290.
- Heyden, A., Bell, A. T., and Keil, F. J. (2005). Efficient methods for finding transition states in chemical reactions: comparison of improved dimer method and partitioned rational function optimization method. *J. Chem. Phys.*, volume 123, issue 22, article 224101.
- Jónsson, H., Mills, G., and Jacobsen, K. W. (1998). Nudged elastic band method for finding minimum energy paths of transitions. In Berne, B. J., Ciccotti, G., and Coker, D. F. (editors), *Classical and Quantum Dynamics in Condensed Phase Simulations*, World Scientific: Singapore, pages 385–404.
- Kamath, A., Vargas-Hernández, R. A., Krems, R. V., Carrington, T., and Manzhos, S. (2018). Neural networks vs Gaussian process regression for representing potential energy surfaces: a comparative study of fit quality and vibrational spectrum accuracy. *J. Chem. Phys.*, volume 148, issue 24, article 241702.
- Kästner, J. and Sherwood, P. (2008). Superlinearly converging dimer method for transition state search. *J. Chem. Phys.*, volume 128, issue 1, article 14106.
- Keck, J. C. (1967). Variational theory of reaction rates. In Prigogine, I. (editor), *Advances in Chemical Physics*, volume 13, Wiley: New York, pages 85–121.
- Kimeldorf, G. S. and Wahba, G. (1970). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Ann. Math. Stat.*, volume 41, issue 2, pages 495–502.
- Kolmogorov, A. N. (1941). Interpolation und extrapolation von stationären zufälligen folgen. *Izv. Akad. Nauk SSSR Ser. Mat.*, volume 5, issue 1, pages 3–14.
- Kramers, H. A. (1940). Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica*, volume 7, issue 4, pages 284–304.
- Krige, D. G. (1951). A statistical approach to some basic mine valuation problems on the Witwatersrand. *J. South. Afr. Inst. Min. Metall.*, volume 52, issue 6, pages 119–139.
- Lampinen, J. and Vehtari, A. (2001). Bayesian approach for neural networks: review and case studies. *Neural Netw.*, volume 14, issue 3, pages 257–274.
- Liu, D. C. and Nocedal, J. (1989). On the limited memory BFGS method for large scale optimization. *Math. Program.*, volume 45, issues 1–3, pages 503–528.
- Malek, R. and Mousseau, N. (2000). Dynamics of Lennard-Jones clusters: a characterization of the activation-relaxation technique. *Phys. Rev. E*, volume 62, issue 6, pages 7723–7728.
- Maras, E., Trushin, O., Stukowski, A., Ala-Nissilä, T., and Jónsson, H. (2016). Global transition path search for dislocation formation in Ge on Si(001). *Comput. Phys. Commun.*, volume 205, pages 13–21.

- Matérn, B. (1960). *Spatial Variation*. Allmänna förlaget: Stockholm.
- Matheron, G. (1963). Principles of geostatistics. *Econ. Geol.*, volume 58, issue 8, pages 1246–1266.
- Mills, G., Jónsson, H., and Schenter, G. K. (1995). Reversible work based transition state theory: application to H₂ dissociative adsorption. *Surf. Sci.*, volume 324, issues 2–3, pages 305–337.
- Minka, T. P. (2001). Expectation propagation for approximate Bayesian inference. In Breese, J. S., Koller, D. (editors), *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann Publishers: San Francisco, pages 362–369.
- Mises, R. von (1964). *Mathematical Theory of Probability and Statistics*. Academic Press: New York.
- Müller, K. and Brown, L. D. (1979). Location of saddle points and minimum energy paths by a constrained simplex optimization procedure. *Theor. Chim. Acta*, volume 53, issue 1, pages 75–93.
- Munro, L. J. and Wales, D. J. (1999). Defect migration in crystalline silicon. *Phys. Rev. B*, volume 59, issue 6, pages 3969–3980.
- Neal, R. M. (1995). *Bayesian Learning for Neural Networks*. Ph.D. thesis, University of Toronto.
- Neal, R. M. (1999). Regression and classification using Gaussian process priors. In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M. (editors), *Bayesian Statistics 6*, Clarendon Press: Oxford, pages 475–501.
- Nocedal, J. (1980). Updating quasi-Newton matrices with limited storage. *Math. Comput.*, volume 35, issue 151, pages 773–782.
- O’Hagan, A. (1978). Curve fitting and optimal design for prediction. *J. Royal Stat. Soc. B*, volume 40, issue 1, pages 1–42.
- O’Hagan, A. (1992). Some Bayesian numerical analysis. In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M. (editors), *Bayesian Statistics 4*, Clarendon Press: Oxford, pages 345–363.
- Olsen, R. A., Kroes, G. J., Henkelman, G., Arnaldsson, A., and Jónsson, H. (2004). Comparison of methods for finding saddle points without knowledge of the final states. *J. Chem. Phys.*, volume 121, issue 20, pages 9776–9792.
- Olver, F. W. J. and Maximon, L. C. (2010). Bessel functions. In Olver, F. W. J., Lozier, D. W., Boisvert, R. F., and Clark, C. W. (editors), *NIST Handbook of Mathematical Functions*, Cambridge University Press: Cambridge, pages 215–286.
- Peterson, A. A. (2016). Acceleration of saddle-point searches with machine learning. *J. Chem. Phys.*, volume 145, issue 7, article 74106.

- Polak, E. and Ribière, G. (1969). Note sur la convergence de méthodes de directions conjuguées. *ESAIM: Mathematical Modelling and Numerical Analysis – Modélisation Mathématique et Analyse Numérique*, volume 3, issue R1, pages 35–43.
- Quiñonero-Candela, J. and Rasmussen, C. E. (2005). A unifying view of sparse approximate Gaussian process regression. *J. Mach. Learn. Res.*, volume 6, pages 1939–1959.
- Rasmussen, C. E. (1996). *Evaluations of Gaussian Processes and Other Methods for Non-Linear Regression*. Ph.D. thesis, University of Toronto.
- Rasmussen, C. E. (2003). Gaussian processes to speed up hybrid Monte Carlo for expensive Bayesian integrals. In Bernardo, J. M., Dawid, A. P., Berger, J. O., West, M., Heckerman, D., Bayarri, M. J., and Smith, A. F. M. (editors), *Bayesian Statistics 7*, Clarendon Press: Oxford, pages 651–659.
- Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press: Cambridge.
- Ren, W. (2003). Higher order string method for finding minimum energy paths. *Comm. Math. Sci.*, volume 1, issue 2, pages 377–384.
- Riihimäki, J. and Vehtari, A. (2010). Gaussian processes with monotonicity information. In Teh, Y. W. and Titterton, M. (editors), *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 645–652.
- Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. Royal Stat. Soc. B*, volume 71, issue 2, pages 319–392.
- Sansò, F. and Schuh, W.-D. (1987). Finite covariance functions. *Bull. Geodesique*, volume 61, issue 4, pages 331–347.
- Seeger, M., Williams, C. K. I., and Lawrence, N. D. (2003). Fast forward selection to speed up sparse Gaussian process regression. In Bishop, C. M. and Frey, B. J. (editors), *Proceedings of the Ninth International Conference on Artificial Intelligence and Statistics*, Society for Artificial Intelligence and Statistics.
- Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., and de Freitas, N. (2016). Taking the human out of the loop: a review of Bayesian optimization. *Proc. IEEE*, volume 104, issue 1, pages 148–175.
- Sheppard, D., Terrell, R., and Henkelman, G. (2008). Optimization methods for finding minimum energy paths. *J. Chem. Phys.*, volume 128, issue 13, article 134106.
- Simons, J., Jørgensen, P., Taylor, H., and Ozment, J. (1983). Walking on potential energy surfaces. *J. Phys. Chem.*, volume 87, issue 15, pages 2745–2753.
- Smidstrup, S., Pedersen, A., Stokbro, K., and Jónsson, H. (2014). Improved initial guess for minimum energy path calculations. *J. Chem. Phys.*, volume 140, issue 21, article 214106.

- Snelson, E. and Ghahramani, Z. (2006). Sparse Gaussian processes using pseudo-inputs. In Weiss, Y., Schölkopf, B., and Platt, J. C. (editors), *Advances in Neural Information Processing Systems 18*, MIT Press: Cambridge, pages 1257–1264.
- Solak, E., Murray-Smith, R., Leithead, W. E., Leith, D. J., and Rasmussen, C. E. (2003). Derivative observations in Gaussian process models of dynamic systems. In Becker, S., Thrun, S., and Obermayer, K. (editors), *Advances in Neural Information Processing Systems 15*, MIT Press: Cambridge, pages 1057–1064.
- Stein, M. L. (1999). *Interpolation of Spatial Data*. Springer-Verlag: New York.
- Swope, W. C., Andersen, H. C., Berens, P. H., and Wilson, K. R. (1982). A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: application to small water clusters. *J. Chem. Phys.*, volume 76, issue 1, pages 637–649.
- Titsias, M. K. (2009). Variational learning of inducing variables in sparse Gaussian processes. In van Dyk, D. and Welling, M. (editors), *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 567–574.
- Trygubenko, S. A. and Wales, D. J. (2004). A doubly nudged elastic band method for finding transition states. *J. Chem. Phys.*, volume 120, issue 5, pages 2082–2094.
- Vanhatalo, J., Pietiläinen, V., and Vehtari, A. (2010). Approximate inference for disease mapping with sparse Gaussian processes. *Stat. Med.*, volume 29, issue 15, pages 1580–1607.
- Vanhatalo, J. and Vehtari, A. (2008). Modelling local and global phenomena with sparse Gaussian processes. In McAllester, D. A. and Myllymäki, P. (editors), *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, AUAI Press: Corvallis, pages 571–578.
- Vineyard, G. H. (1957). Frequency factors and isotope effects in solid state processes. *J. Phys. Chem. Solids*, volume 3, issues 1–2, pages 121–127.
- Voter, A. F. (1997). Hyperdynamics: accelerated molecular dynamics of infrequent events. *Phys. Rev. Lett.*, volume 78, issue 20, pages 3908–3911.
- Wang, K. A., Pleiss, G., Gardner, J. R., Tyree, S., and Wilson, A. G. (2019). Exact Gaussian processes on a million data points. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d’Alché-Buc, F., Fox, E., and Garnett, R. (editors), *Advances in Neural Information Processing Systems 32*, Curran Associates: Red Hook, pages 14622–14632.
- Wiener, N. (1949). *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. Wiley: New York.
- Wigner, E. (1938). The transition state method. *Trans. Faraday Soc.*, volume 34, pages 29–41.

- Williams, C. K. I. and Barber, D. (1998). Bayesian classification with Gaussian processes. *IEEE Trans. Pattern Anal. Mach. Intell.*, volume 20, issue 12, pages 1342–1351.
- Williams, C. K. I. and Rasmussen, C. E. (1996). Gaussian processes for regression. In Touretzky, D. S., Mozer, M. C., and Hasselmo, M. E. (editors), *Advances in Neural Information Processing Systems 8*, MIT Press: Cambridge, pages 514–520.
- Wu, Z. (1995). Compactly supported positive definite radial functions. *Adv. Comput. Math.*, volume 4, issue 1, pages 283–292.
- Zhu, X., Thompson, K. C., and Martínez, T. J. (2019). Geodesic interpolation for reaction pathways. *J. Chem. Phys.*, volume 150, issue 16, article 164103.

Errata

Publication I

The early stopping and convergence criteria are described correctly in the Methods section (section 2.3), but there are unfortunate mistakes in the following two sentences in the Results section. A corrected version of Publication I is available as e-print arXiv:1703.10423.

Original sentence (rows 12–15 of section 3.1):

The relaxation of the images on this rough estimate of the energy surface does not, however, bring the images too far from the initial placement because of the condition that images cannot be moved in a single iteration by more than a half of the initial distance between the images.

Corrected sentence:

[...] because of the condition that the relaxation phase is stopped early if the convergence measure, i.e., the mean of the magnitudes of the force components perpendicular to the path at the intermediate images, increases.

Original sentence (rows 3–5 of the caption of figure 4):

The convergence tolerance is $0.001 \text{ eV}/\text{\AA}$ for the magnitude of the perpendicular component of the force on any one of the images.

Corrected sentence:

[...] for the mean of the magnitudes of the perpendicular force components at the intermediate images.

Publications I and II

The computational complexity of one inner iteration of the GP-NEB algorithm is claimed to be linear with respect to the number of degrees of freedom D (row 25 of section 2.3 in Publication I and the last three rows of section IV.A in Publication II). The computational cost of prediction of energy or any gradient component indeed scales linearly with respect to D , but since moving the images requires prediction of the whole gradient vector, the complexity of one inner iteration becomes quadratic with respect to D .



ISBN 978-952-60-8850-1 (printed)
ISBN 978-952-60-8851-8 (pdf)
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)

Aalto University
School of Science
Department of Computer Science
www.aalto.fi

**BUSINESS +
ECONOMY**

**ART +
DESIGN +
ARCHITECTURE**

**SCIENCE +
TECHNOLOGY**

CROSSOVER

**DOCTORAL
DISSERTATIONS**