Numerical Analysis of Optimization-Constrained Differential Equations: Applications to Atmospheric Chemistry

THÈSE N^{O} 4345 (2009)

PRÉSENTÉE LE 27 MARS 2009

À LA FACULTÉ SCIENCES DE BASE

CHAIRE D'ANALYSE ET DE SIMULATION NUMÉRIQUE

PROGRAMME DOCTORAL EN MATHÉMATIQUES

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Chantal LANDRY

acceptée sur proposition du jury:

Prof. F. Eisenbrand, président du jury
Prof. J. Rappaz, Prof. A. Caboussat, directeurs de thèse
Dr A. Clappier, rapporteur
Prof. E. Hairer, rapporteur
Prof. J. He, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2009

Abstract

The modeling of a system composed by a gas phase and organic aerosol particles, and its numerical resolution are studied. The gas-aerosol system is modeled by ordinary differential equations coupled with a mixed-constrained optimization problem. This coupling induces discontinuities when inequality constraints are activated or deactivated.

Two approaches for the solution of the optimization-constrained differential equations are presented. The first approach is a time splitting scheme together with a fixed-point method that alternates between the differential and optimization parts. The ordinary differential equations are approximated by the Crank-Nicolson scheme and a primal-dual interior-point method combined with a warm-start strategy is used to solve the minimization problem. The second approach considers the set of equations as a system of differential algebraic equations after replacing the minimization problem by its first order optimality conditions. An implicit 5th-order Runge-Kutta method (RADAU5) is then used. Both approaches are completed by numerical techniques for the detection and computation of the events (activation and deactivation of inequality constraints) when the system evolves in time. The computation of the events is based on continuation techniques and geometric arguments. Moreover the first approach completes the computation with extrapolation polynomials and sensitivity analysis, whereas the second approach uses dense output formulas.

Numerical results for gas-aerosol system made of several chemical species are proposed for both approaches. These examples show the efficiency and accuracy of each method. They also indicate that the second approach is more efficient than the first one. Furthermore theoretical examples show that the method for the computation of the activation is of second order for the first approach and exact for the second one.

Keywords: Initial value problems, differential algebraic equations, constrained optimization, event detection, discontinuity points, computational chemistry.

Version abrégée

Cette thèse s'intéresse à la modélisation d'un système composé de gaz et de particules aérosols organiques ainsi qu'à sa résolution numérique. Le système gaz-aérosol est modélisé par des équations différentielles ordinaires auxquelles on couple un problème d'optimisation avec des contraintes d'égalité et d'inégalité. De ce couplage naissent des discontinuités. Elles apparaissent dans le système lorsque les constraintes d'inégalité s'activent ou se désactivent.

Deux approches sont présentées pour résoudre ce système d'équations. La première approche suit un schéma de *splitting* en séparant la résolution des équations différentielles de celle du problème d'optimisation. Concernant la résolution des équations différentielles, la méthode de Crank-Nicolson est utilisée. Le problème d'optimisation est quant à lui résolu à l'aide d'une méthode de point intérieur à laquelle on ajoute une stratégie *warm-start* pour son initialisation. La seconde approche considère le modèle comme un système d'équations différentielles algébriques après avoir substitué le problème d'optimisation par ses conditions d'optimalité du premier ordre. Elle utilise ensuite une méthode de Runge-Kutta implicite d'ordre 5 (RADAU5) pour résoudre ce nouveau système. Chacune des approches est complétée par des techniques qui détectent l'activation et la désactivation des contraintes d'inégalité, et qui calculent l'instant auquel ces événements se produisent. Ces techniques font appel à des méthodes de continuation et aux caractéristiques géométriques du modèle. Pour la première approche une extrapolation polynomiale et des analyses de sensiblité viennent compléter le calcul des événements, alors que dans la seconde approche des formules de *dense output* sont utilisées.

Des exemples numériques sont présentés pour chacune des approches sur différents systèmes gaz-aérosol. Ces résultats illustrent la rapidité et la précision de chaque méthode ainsi qu'une plus grande efficacité de la part de la seconde. Finalement des exemples théoriques montrent que le calcul des temps d'activation est d'ordre deux pour la première méthode et exacte pour la seconde.

Mots clés: Problèmes de Cauchy, équations différentielles algébriques, optimisation sous contraintes, détection d'événements, points de discontinuité, chimie computationnelle.

Acknowledgements

First of all, I would like to thank Prof. Jacques Rappaz. He has given me the opportunity to go to the USA and accepted my thesis subject which is far from his usual topics. I am very grateful to him for the confidence he has always had in me. I also thank him for introducing me to Prof. Ernst Hairer.

My second thanks are addressed to Prof. Alexandre Caboussat who was more than a thesis supervisor. I am deeply indebted to him for his availability, kindness, and scientific and moral support. Many thanks for all the discussions, the mathematical ones and the others, that we shared during these years, and for welcoming me in Houston so well at each time. Thanks to him and his wife, Anouck, I enjoyed Houston and the Texas a lot.

I am most grateful to Prof. Ernst Hairer for his numerous good advices and our fruitfull discussions. I particularly appreciate his availability, kindness and permanent interest in my research. I would also like to thank him for honouring me by being part of my jury.

A special thank to Prof. Jiwen He for receiving me in Houston when I was there for the first time and giving me the opportunity to work on the project UHAERO. I am grateful to him for all his help on the atmospheric chemistry. I also thank him for accepting to be member of my jury.

I would like to thank Dr Alain Clappier who accepted to be member of my jury and read this thesis. Thanks also to Prof. Friedrich Eisenbrand, president of the jury.

I acknowledge the partial support of the U.S. Environmental Protection Agency, under grant X-83234201. I also thank the department of mathematics at the University of Houston for having partially supported this work, in particular for having sponsored my visits at the University of Houston.

Many thanks to all the colleagues at EPFL, especially to Annalisa, Gonçalo and Guillaume with whom I have shared so many amazing and unforgettable moments, inside and outside the EPFL. A special thank to my flatmate, Andrea, for our numerous discussions and UNO games, the amarettos and the portos which were a very good way to decompress. I would like to thank Mireille, Glenn, François and all my friends for their sincere friendship and their supports.

Last but not least, a very special thank to my parents, my sister Marie-Claude and my brother Christian for their support, love and patience.

Contents

Introduction **1**

1 Modeling **7**
- 1.1 Atmospheric aerosol particles 7
- 1.2 The gas-aerosol system 8
- 1.3 Mass transfer 10
- 1.4 Phase equilibrium problem 12
  - 1.4.1 Liquid-liquid separation 12
  - 1.4.2 Formulation of the phase equilibrium problem 13
  - 1.4.3 First-order optimality conditions 16
- 1.5 Computation of the gas-concentration vector $\mathbf{c}_g^{surf}$ 20
- 1.6 Formulation of the coupled model 21
- 1.7 Geometric interpretation 24

2 Solving the phase equilibrium problem **33**
- 2.1 Mathematical characterizations of the local minima 33
- 2.2 Solution of the phase equilibrium problem 37
  - 2.2.1 Active set method 37
  - 2.2.2 Primal-dual interior-point method 39
  - 2.2.3 Initialization procedure 43

3 An optimization-based numerical method **47**
- 3.1 Numerical method for a fixed number of inactive constraints 48
  - 3.1.1 A time splitting scheme 48
  - 3.1.2 A warm-start strategy 51
- 3.2 Detection of the discontinuity times and points 52
  - 3.2.1 Criterion for the detection 53
  - 3.2.2 Computation of the minimal distance criterion 56
- 3.3 Computation of the discontinuity times and points 60
  - 3.3.1 Activation of an inequality constraint 61

3.3.2 Deactivation of an inequality constraint 69
3.4 A priori error estimation 75
3.4.1 Theoretical results in 3 dimensions 76
3.4.2 Theoretical results in s dimensions 82
3.4.3 Theoretical example 82
3.5 Numerical results 88
3.5.1 Numerical results in one dimension 89
3.5.2 Numerical results in two dimensions 93
3.5.3 Numerical results in higher dimensions 117

4 A numerical method based on differential algebraic systems 123
4.1 Numerical method for a fixed number of inactive constraints 124
4.1.1 The differential algebraic equations 124
4.1.2 Runge-Kutta methods and the RADAU5 method 127
4.1.3 Numerical method for differential algebraic equations 130
4.2 Detection of discontinuity times and points 131
4.2.1 Activation of an inequality constraint 131
4.2.2 Deactivation of an inequality constraint 132
4.3 Computation of the discontinuity times and points 132
4.4 Numerical Results 135
4.4.1 Numerical results in one dimension 135
4.4.2 Numerical results in two dimensions 139
4.4.3 Numerical results in higher dimensions 147

Conclusions 153

Introduction

Climate change and air pollution are among the main environmental preoccupations of this century. Both problems are very closely related while they both depend on the composition of the atmosphere.

The Earth's climate changes when the global energy balance between incoming energy from the Sun and outgoing from the Earth is upset. There are a number of natural mechanisms that can upset this balance, for example fluctuations in the Earth's orbit, variations in ocean circulation and changes in the composition of the Earth's atmosphere [96, 102]. In recent times, the latter has been evident as a consequence of the increasing emissions of greenhouse gases in the atmosphere [78]. Moreover atmospheric physicists now recognize aerosol particles as a new agent in the climate change [92]. Atmospheric aerosols influence the transfer of energy in the atmosphere in two ways referred to as direct forcing and indirect forcing. In the direct forcing mechanism, aerosols reflect sunlight back to space and cool the planet. The indirect forcing modifies the optical properties and lifetimes of clouds. Aerosol particles act as additional cloud condensation nuclei, spreading the cloud's liquid water over more, smaller droplets. This makes clouds more reflective and longer lasting [53, 90, 96, 99, 102].

Air pollution may be defined as the presence of substances in the atmosphere causing adverse effects to man and the environment. Natural air pollution has occurred on Earth since the planet's formation. Fires, volcanic eruptions, meteorite impacts and high winds all cause natural air pollution [53, 96]. Anthropogenic air pollution problems have existed on urban scales for centuries and have resulted from burning of wood, vegetation, coal, oil, natural gas, waste and chemicals. In the nineteenth and early twentieth centuries, most air pollution was due to chimney and smokestack emission of coal and chemical-factory combustion products [102]. In the early twentieth century, the widespread use of automobiles and the increase in industrial activity increased the prevalence of air pollution. Air pollutants consist of either trace gases or aerosol particles [53, 96]. They can have detrimental effects on human health such as asthma [13, 77, 111]. Aerosol particles may be seen as the most critical of all pollutants, and some estimates have suggested that they are responsible for up to 10,000 premature deaths in the UK each year. Air pollutants also affect acid deposition [58, 76] and their high number reduces the visibility in urban and regional areas [63, 65].

In conclusion, atmospheric aerosol particles have impacts on both air pollution and climate change. Their impacts do not only depend on their concentration in the atmosphere, but also on their physical state and chemical composition [95]. Since their amount in the atmosphere is still growing, a better prediction of their behaviour is essential. A better understanding leads to develop tools that can be used for policy making. With accurate models, policy makers can try to mitigate pollution and climate problems [78].

Atmospheric particles undergo physical and chemical transformations in the atmosphere that alter their number, size, composition and physical state. These transformations are due to the following processes: condensation, evaporation, coagulation, gravitational settling, nucleation, advection, turbulent transport, emission, deposition and chemical reactions [53, 96, 111]. Gelbard and Seinfeld in [43] introduced the General Dynamic Equation (GDE) for aerosol particles. This equation describes mathematically all the above-mentioned processes and is the reference equation for the dynamics of aerosols. The variables are only the particle size and composition. Several resolution methods are proposed [85, 86, 106, 111] and aerosol dynamics models exist (MADM [83], MOSAIC [110] and MAM [94, 98]). In these models, significant simplifications are employed. The first reason is that little experimental data are available [106]. The measurement of aerosol particles and their properties is still an ongoing challenge in atmospheric sciences [59, 67]. This lack of data are sources of uncertainty in models. The second reason is that the simulation of aerosol dynamics is the slowest part in air quality models [54, 74, 84]. In order to keep reasonable computational time, simplifications in the modeling of the aerosol dynamics have to be considered.

A first simplification is to suppose that the aerosol particles are internally mixed, that is all the particles of a given size have the same composition, even if a few investigators have shown that ambient aerosol particles are not internally mixed [26, 70]. With this assumption, it remains only one independent variable in the GDE: the size of the particles. In early aerosol models another simplification was assumed: the instantaneous thermodynamic equilibrium for the condensation and evaporation processes. Then these two processes are reduced to a set of stiff ordinary differential equations [85, 86, 106]. However, Wexler and Seinfeld [107], and Meng and Seinfeld [71] show that this hypothesis is wrong. Furthermore Allen et al. [1] and Wexler and Seinfeld [107] indicate that equilibrium alone cannot uniquely determine the size, and both mass transport and thermodynamics must be considered to accurately predict the size distribution of aerosol particles. Actual aerosol models [83, 110] couple the thermodynamics with condensation and evaporation processes. However all simplify or even neglect the part dedicated to the prediction of the physical and chemical properties of atmospheric aerosols. Only the size distribution is modeled.

The purpose of this thesis is to develop a model and a numerical method for the prediction of the dynamics of atmospheric aerosol particles. Since in the GDE the most important part is due to condensation and evaporation [106], only these two processes are considered. Although, unlike in the above-mentioned models, the aerosol particles are not supposed internally mixed and the size of the aerosol particles is not the sole variable in the model. We are also interested in the prediction of the composition and physical state of the aerosols, which is determined by the thermodynamic equilibrium. The final aim of this work is to incorporate the model in 3D air quality models such as the U.S. Models3/CMAQ [17] and the Harvard GEOS-CHEM model [80]. In order to be competitive the numerical method has to be fast, accurate and must not consume too much memory.

The system under study is made of atmospheric aerosol particles and a surrounding gas of same chemical composition. The system is supposed to be closed in the sense that no matter is created nor destroyed. Gas and particles interact with each other in order to reach the equilibrium in the system. These interactions lead to changes in the gas-particle partitioning and in the thermodynamic equilibrium inside the particles. The partitioning is mathematically represented by ordinary differential equations, whereas the thermodynamic equilibrium inside the aerosols is obtained by solving a mixed-constrained global optimization problem. Thus a mathematical model for this problem can be written as follows [18, 19]: for $p, q > 0$, $T > 0$ and $\mathbf{b}_0$ given, find $\mathbf{b} : t \to \mathbf{b}(t) \in \mathbb{R}^p$ and $\mathbf{x} : t \to \mathbf{x}(t) \in \mathbb{R}^q$ satisfying

$$\begin{aligned} \frac{d}{dt}\mathbf{b}(t) &= \mathbf{f}(t, \mathbf{b}(t), \mathbf{x}(t)), \qquad\qquad \mathbf{b}(0) = \mathbf{b}_0 \\ \mathbf{x}(t) &= \arg\min_{\bar{\mathbf{x}}}\ \mathcal{G}(\bar{\mathbf{x}}) \\ &\qquad\quad \text{s.t. } c_{\mathcal{I}}(\bar{\mathbf{x}}, \mathbf{b}(t)) \geq 0, \quad \forall \mathcal{I} = 1, \ldots, n_{\mathcal{I}}, \\ &\qquad\qquad\quad c_{\mathcal{E}}(\bar{\mathbf{x}}, \mathbf{b}(t)) = 0, \quad \forall \mathcal{E} = 1, \ldots, n_{\mathcal{E}}, \end{aligned} \tag{0.0.1}$$

where $t \in (0, T)$. The vector $\mathbf{b}$ represents the composition of the aerosol particle and $\mathbf{x}$ describes the thermodynamic equilibrium. The first equation in (0.0.1) is a stiff and nonlinear ordinary differential equation where $\mathbf{f}$ is a continuous vector-valued function. The second part of (0.0.1) corresponds to a global minimization problem subjected to $n_{\mathcal{I}}$ inequality constraints and $n_{\mathcal{E}}$ equality constraints. The objective function $\mathcal{G}$ represents the Gibbs free energy of the aerosol. This function is nonconvex, nonlinear and uniquely depends on $\mathbf{x}$. The equality constraints can be nonlinear functions while the inequality constraints are supposed to be linear.

Note that problem (0.0.1) is not an optimal control problem. The main difference between problems arising in control systems theory [104] and the present problem resides in the fact that the function $\mathcal{G}$ is minimized for a.e. $t \in (0, T)$ along the trajectory, and not only at the final time T for instance.

The purpose of this thesis is to present an efficient numerical method that solves optimization-constrained differential equations like (0.0.1) in the framework of atmospheric particles, together with the corresponding appropriate model. System (0.0.1) is such that as soon as an inequality constraint is activated or deactivated, the variable $\mathbf{x}$ is "truncated" and loses regularity. The time at which the "truncation" occurs is called a *discontinuity time* and $\mathbf{x}$ evaluated at this time is a *discontinuity point.* The numerical method has to accurately detect and compute the times of activation and deactivation of constraints in order to (i) compute the discontinuity time and point and (ii) guarantee the correctness and accuracy of the numerical solution of (0.0.1).

The first part of this thesis concerns the origin and the mathematical formulation of the model (Chapter 1). The variables and a set of notations are defined. The chemical and physical laws that govern a system composed of *organic aerosol particles* are presented in order to obtain the mathematical system of the form (0.0.1). The geometrical interpretation of the problem is explained. The dynamics of the system are interpreted as the dynamic computation of the convex envelope of the objective function $\mathcal{G}$. A series of definitions and notations is given.

Chapter 2 treats the resolution of the optimization problem. This problem is treated separately for several reasons. The first one is because this thesis is the continuation of the works of [4, 5] where Amundson et al. developed an efficient algorithm for the computation of the thermodynamic equilibrium for a given organic aerosol particle. Their method is based on primal-dual interior-point techniques [10, 72, 73, 75]. The second reason is that its resolution appears at several times in the numerical method to solve (0.0.1).

Two different approaches for the resolution of (0.0.1) are presented. The first one follows the optimization techniques developed in Chapter 2 whereas the second one reads the system (0.0.1) as a differential algebraic system. For both approaches the resolution method without any tracking of discontinuities is first introduced. Then the strategy for the detection and computation of a deactivation or activation is explained. The difficulty in both cases resides in the impossibility of defining an explicit event function that characterizes the activation or deactivation of a constraint. Finally numerical and some theoretical results follow each method.

The first method is studied in Chapter 3. This method has the aim of keeping the optimization technique developed in Chapter 2. Therefore a time splitting idea is considered to solve the system (0.0.1). The ordinary differential equations are discretized in time with the Crank-Nicolson scheme and combined with the first order optimality conditions of the minimization problem. The resulting system is then solved with a fixed point technique. The detection of events is based on the behaviour of the resolution of the optimization problem. Classical event detection algorithms couple a discontinuity locking approach [23] with interpolation techniques [35, 81, 97]. However, because of the truncation of the variable $\mathbf{x}$, the interpolation polynomials are inefficient. Extrapolation techniques are well-adapted to compute the discontinuity points. Our computation of the discontinuity points follow the works of Esposito and Kumar in [33]. The idea is to extrapolate an approximation of the function that describes the activation or the deactivation. The order of convergence of this first method is proved when inequality constraints are activated and several numerical examples emphasize the efficiency of this first approach.

The second method is presented in Chapter 4. If the number of active inequality constraints is fixed, the considered system can be associated to a system of differential algebraic equations (DAE), by replacing the minimization problem by its first order optimality conditions. The second method is based on this observation.

The strategy for solving (0.0.1) is similar to the first method in the sense that it splits the resolution in two steps: (i) solve the DAE system when the number of active inequality constraints is fixed, (ii) verify at each time step if an inequality constraint has to be activated or deactivated. If it is the case, compute the discontinuity time and point, define the new DAE system and restart in (i).

Efficient techniques to solve DAE systems, relying on implicit Runge-Kutta methods, have been developed in [7, 49, 50, 61]. The DAE system stemmed from (0.0.1) is solved by using the implicit Runge-Kutta method of order 5, *RADAU5*, developed by Hairer and Wanner in [50]. The detection of an activation or deactivation follows the idea developed for the first method. Concerning the computation of the discontinuity time and point, the strategies differ. Guglielmi and Hairer suggest in [47] an efficient technique to compute breaking points when the system to solve is a DAE system and a Runge-Kutta method is used. Once the time interval that contains the breaking point is defined, the insight of this technique is to insert the step size needed to reach the breaking point as a variable in the set of equations and to solve this new augmented system with a splitting idea in order to keep the efficient resolution method for the DAE. This technique is applied for the computation of the discontinuity time and point in (0.0.1).

Since the discontinuity corresponds either to the activation, or to the deactivation, of an inequality constraint, the number of active inequality constraints changes at the discontinuity time. Hence before restarting the resolution of the DAE system, one needs to adapt the DAE system to the new number of active constraints. In particular the number of unknowns and the size of the DAE system change after activation or deactivation.

As for the first method, several numerical results are presented to illustrate the efficiency and accuracy of the algorithm.

The final chapter of this thesis summarizes each method with their advantages and limits. Some perspectives of the work are also discussed.

Chapter 1

Modeling

This first chapter presents the gas-aerosol system and the corresponding mathematical model. The first part of this chapter proposes a presentation of the atmospheric aerosol particles. Definition, characteristics and impacts of these particles are introduced which also gives the motivations to model the existing interactions between the gas phase and the aerosol particles. Then the transcription of these interactions into mathematical equations is detailed. Since a minimization problem occurs in the mathematical formulation, some insights in optimization theory are developed. Finally, once the model is mathematically formulated, its geometric interpretation is proposed. This interpretation gives groundwork to the techniques that detect the discontinuity points.

1.1 Atmospheric aerosol particles

An atmospheric aerosol is an ensemble of solid, liquid or mixed-phase particles suspended in air [53, 96]. Each particle consists of an aggregate of atoms and/or molecules bonded together. An aerosol particle is a single particle within an aerosol, but often loosely referred to as simply aerosol.

Particle emission originates from natural and anthropogenic sources [96, 99, 102]. Natural sources include wind uplift of sea spray, soil dust, pollen, spores and bacteria, volcanic outgassing, natural biomass fires, and lightning. Anthropogenic sources include fossil-fuel combustion, biofuel burning, biomass burning, and wind uplift of soil dust over eroded land. Globally, particle emission rates from natural sources exceed those from anthropogenic sources. In urban areas, the reverse is true.

Aerosol particles have always been present on the Earth since its formation. However their increasing amount, essentially due to human activities, leads to deep impacts on the planet [78]. Aerosols affect health, air quality, cloud formation, meteorology and climate. Submicrometer particles (those smaller than $1\mu m$ in diameter) affect human health by penetrating to the deepest part of human lungs [13, 77]. Aerosol particles with a diameter of $0.2 - 1\mu m$ that contain sulfate, nitrate, and organic carbon scatter light efficiently. Aerosol particles smaller than $1\mu m$ that contain black carbon absorb light efficiently. Aerosol absorption and scattering affect (1) radiative energy fluxes, which affect temperatures, and (2) photolysis, which affects the composition of the atmosphere. Aerosol particles also serve as nuclei for the formation of cloud drops [88]. In fact, without aerosol particles, clouds would rarely form in the atmosphere. Actually, the large number of particles in the atmosphere leads to clouds that are more reflective and last longer. Finally, aerosol particles serve as sites on which chemical reactions take place and as sites for trace gases to condense upon or dissolve with. Therefore the study of these particles is crucial in order to regulate and decrease their impacts.

The atmospheric aerosols are divided into two categories: the organic aerosols, which only contain atoms of carbon, hydrogen, oxygen and nitrogen; and the inorganic aerosols which can be made of sulfate, nitrate, sodium, trace metals, carbonaceous material, etc. All these aerosols may be composed of several liquid phases and the inorganic compounds can react and lead to the formation of solid phases in addition [53, 96]. Phases are sometimes confused with states of matter but there are significant differences. Gas, liquid and solid are known as the states of matter, but each of solid and liquid states may exist in one or more forms. The term *phase* is required to describe the various forms. A phase is a region of matter with uniform chemical composition and physical properties. For example, salad dressing may separate into an oil-phase and a water-rich phase; there exist then 2 phases in the dressing, both of which are in the liquid state. The repartition of the aerosol between different phases is called the *phase equilibrium* or *phase repartition* of the aerosol [5, 69].

Aerosols are subjected to an array of processes that modify their size, composition, and phase repartition. The subject of this thesis is concerned by the time evolution of a system composed by organic aerosol particles and gases when the processes of evaporation and condensation are taken into consideration together with the phase equilibrium process inside the particle. Evaporation occurs when a liquid molecule on a particle surface changes state to a gas and diffuses away from the surface. Condensation occurs when a gas diffuses to and sticks to the surface of a particle and changes state to a liquid.

In the next section a description of the gas-aerosol system is presented as well as the governing equations that describe the evaporation and condensation. For realistic and chemically interesting simulations let us consider N identical aerosol particles in the system. These particles interact with the gas identically and at the same time. For this reason only the time evolution of one particle is depicted on the forthcoming graphs and numerical results.

1.2 The gas-aerosol system

The modeling of the evolution of identical organic aerosol particles embedded in a gas of same chemical composition is studied. The particles are supposed to initially exist and do not disappear. Let us assume on the one hand that the temperature and the pressure are constant, and on the other hand that the system formed by the aerosols and gas is closed, namely matter is neither created nor destroyed. Since the system is closed a mass transfer between the aerosol particles and gas takes place through evaporation or condensation until the equilibrium between the concentrations of the gas far from the particles, and at the particle surface is reached.

Schematically the system is represented in Figure 1.1 when only one particle is in the system. The aerosol particles are supposed to be spherical and surrounded by gas molecules. Any chemical component existing in the gas state is also assumed to be present inside the particles. For example if the aerosols are composed of water (H_2O) and hexacosanol ($C_{26}H_{54}O$), then these two chemical components are in a liquid state in the particles and there are molecules of water and molecules of hexacosanol in the gas state surrounding the aerosols.

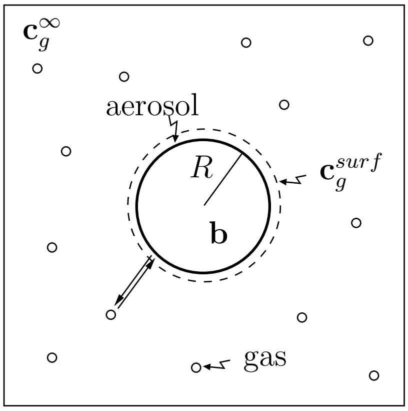


Figure 1.1: Gas-aerosol system when $N = 1$.

Let us introduce some notations:

- s, the number of different chemical components existing in the system,
- $\mathbf{b}$, the composition-vector of the s chemical components present in a particle $[mol]$,
- N, the number of particles of composition $\mathbf{b}$ per unit of volume $[m^{-3}]$,
- $\mathbf{c}_g^\infty$, the vapor-phase concentration-vector of the s chemical components present in the gas far from the particle, i.e. far enough from the particles in order to be considered at equilibrium $[mol/m^3]$,
- $\mathbf{c}_g^{surf}$, the vapor-phase concentration-vector of the s chemical components present in the gas at the particle surface $[mol/m^3]$,
- R, the radius of each aerosol particle $[m]$.

The variables $\mathbf{b}$, $\mathbf{c}_g^\infty$, $\mathbf{c}_g^{surf}$ and R are functions of time, whereas s and N are constant.

The evaporation and condensation processes induce mass transfer between gas and aerosols. The values of $\mathbf{b}$, $\mathbf{c}_g^\infty$, $\mathbf{c}_g^{surf}$ and R vary because of this mass flux. The modeling of the flux is done through a system of ordinary differential equations (ODE). The other point of interest is the determination of the phase equilibrium of each aerosol particle. This equilibrium state is characterized by the global minimum of the Gibbs free energy of each particle. In the next sections the coupling between these two problems is established. The expression of the mass flux depends on the liquid phases computed by the phase equilibrium problem, whereas the minimization problem is solved for a composition-vector **b** that is solution of the ODE problem. Let us begin with the description of the mass flux.

1.3 Mass transfer

The mass transfer between the particles and the surrounding gas is modeled by the following system of ordinary differential equations (ODE):

$$\begin{aligned} \frac{d}{dt}\mathbf{c}_g^\infty(t) &= -N\,\mathbf{j}(\mathbf{c}_g^\infty(t), \mathbf{c}_g^{surf}(t), R(t)), && (1.3.1a)\\ \frac{d}{dt}\mathbf{b}(t) &= \mathbf{j}(\mathbf{c}_g^\infty(t), \mathbf{c}_g^{surf}(t), R(t)), && (1.3.1b)\\ R(t) &= \left(\frac{3}{4\pi}\sum_{i=1}^{s}\frac{m_{c,i}\,b_i(t)}{\rho_i}\right)^{\frac{1}{3}}, && (1.3.1c) \end{aligned}$$

where for $i = 1, \ldots, s$, ρ_i is the density of the chemical components i, $m_{c,i}$ is the atomic mass of the chemical component i, and **j** represents the molecular flux between the gas and each aerosol particle.

The equation (1.3.1c) comes from the expression of the volume V of a particle. Since the aerosol particle is supposed to be spherical, the volume is given by

$$V = \frac{4}{3}\pi R^3.$$

On the other hand the volume V can be considered as the addition of the volume V_i of each component present in the aerosol, i.e.

$$V = \sum_{i=1}^{s} V_i.$$

This volume V_i can be expressed by the density ρ_i of the chemical component i

$$V_i = \frac{m_{c,i}b_i}{\rho_i}, \quad i = 1, \ldots, s.$$

Combining these last 3 equations the relation (1.3.1c) is obtained.

Concerning the mass flux **j**, multiple definitions exist. Particle growth or evaporation depends on the direction of the net flux of vapor molecules relative to a particle, namely $\mathbf{c}_g^\infty - \mathbf{c}_g^{surf}$ [96]. If $c_{g,i}^\infty > c_{g,i}^{surf}$, the flow of gas molecules of species i goes inside the particle and if $c_{g,i}^{\infty} < c_{g,i}^{surf}$, it goes outside the particle. Different formulations of the correction term in front of $\mathbf{c}_g^{\infty} - \mathbf{c}_g^{surf}$ have been suggested in atmospheric models [112]. Works of Fuchs [40], Fuchs and Sutugin [41], Dahneke [27], and Wexler and Seinfeld [107] can be cited as the major references. Let us follow the definition proposed by Wexler and Seinfeld, i.e.

$$j_i\left(c_{g,i}^{\infty}(t), c_{g,i}^{surf}(t), R(t)\right) = \frac{4\pi R(t) D_i}{\frac{\lambda_{air}}{\alpha_i R(t)} + 1}\left(c_{g,i}^{\infty}(t) - c_{g,i}^{surf}(t)\right), \qquad \text{for } i = 1, \ldots, s; \tag{1.3.2}$$

where D_i is the diffusion coefficient of the species i, λ_{air} is the mean free path of the air and α_i is the accommodation coefficient of the chemical species i. Let us define the following matrix $\mathbf{H}$

$$\mathbf{H} = \operatorname{diag}\left(4\pi R D_i \frac{1}{\frac{\lambda}{\alpha_i R} + 1}\right)_{i=1,\ldots,s}.$$

The expression of the flux $\mathbf{j}$ becomes

$$\mathbf{j}(\mathbf{c}_g^{\infty}, \mathbf{c}_g^{surf}, R) = \mathbf{H}(R)\left(\mathbf{c}_g^{\infty} - \mathbf{c}_g^{surf}\right), \tag{1.3.3}$$

which represents the number of moles of gas exchanged between each aerosol particle and the surrounding gas per unit of time.

Remark 1.3.1. *The matrix* $\mathbf{H}$ *uniquely depends on the radius* R. *All the other parameters are chemical features of the components present in the aerosol and all are independent of the time. Thus the matrix is written as* $\mathbf{H}(R)$.

Remark 1.3.2. *In equation (1.3.3) the Kelvin effect is neglected. Actually equation (1.3.3) is correct when the surface between the particle aerosol and the gas molecules is flat. As soon as the surface is curved, the formulation of the flux must be modified as follows [53, 96]*

$$\mathbf{j}(\mathbf{c}_g^{\infty}, \mathbf{c}_g^{surf}, R) = \mathbf{H}(R)\left(\mathbf{c}_g^{\infty} - \boldsymbol{\eta}(\mathbf{b}, R)\,\mathbf{c}_g^{surf}\right). \tag{1.3.4}$$

The corrective term $\boldsymbol{\eta}(\mathbf{b}, R)$ *is the* $s \times s$ *diagonal matrix defined by*

$$\boldsymbol{\eta}(\mathbf{b}, R) = \operatorname{diag}\left(\exp\left(\frac{2\sigma(\mathbf{b}) m_{w,i}}{\mathcal{R}_c T \rho_i R}\right)\right)_{i=1,\ldots,s},$$

where $\sigma(\mathbf{b})$ *is the surface tension coefficient when the composition of the particle is given by* $\mathbf{b}$, $m_{w,i}$ *is the molecular weight of the chemical species* i, T *is the temperature of the system and* $\mathcal{R}_c$ *is the ideal gas constant (*$\mathcal{R}_c = 8.20574587 \cdot 10^{-5}\,[m^3 atm K^{-1} mol^{-1}]$*).*

At small particle sizes the Kelvin effect can be significant. However this effect becomes negligible for particle diameters larger than $20\,nm$. *For a* $20\,nm$ *diameter particle the effect represents only a* 5% *correction to* $\mathbf{c}_g^{surf}$ *[11, 96]. Consequently the Kelvin effect is not included in our model.*

In the differential system (1.3.1) the equations govern the time evolution of the variables $\mathbf{c}_g^{\infty}$, $\mathbf{b}$ and R, but an expression controlling the variable $\mathbf{c}_g^{surf}$ is missing. Let us leave aside the calculation of $\mathbf{c}_g^{surf}$ for a moment and study the determination of the phase equilibrium of an aerosol particle.

1.4 Phase equilibrium problem

Organic aerosol particles can be separated into several liquid phases as explained in Section 1.1. For each composition vector $\mathbf{b}$ of the aerosol, the partitioning of organics between different liquid phases is determined by minimizing the Gibbs free energy of the particle [5, 107]. This optimization problem is called the *phase equilibrium problem* (PEP). For this section we consider an aerosol particle with a fixed composition vector $\mathbf{b}$ (i.e. neither evaporation nor condensation occurs with the surrounding media) and define explicitly the PEP.

1.4.1 Liquid-liquid separation

Before starting with the formulation of the PEP, let us have a look at the example depicted in Table 1.1 and coming from [5] in order to better understand the notion of *liquid phase* and the forthcoming notations. In this example the particle is made of n-Butyl-Acetate ($C_6H_{12}O_2$) and Water (H_2O), and the composition-vector is defined by $\mathbf{b}^T = (0.5,\ 0.5)$. In other words the particle is composed of 0.5 mole of n-Butyl-Acetate and 0.5 mole of Water. Amundson et al. in [5] found that for a temperature of 298 K and a pressure of 1 atm the aerosol is separated into 2 liquid phases. The first liquid phase contains 0.545 mole of a mixture composed of 8.3 % of n-Butyl-Acetate and 91.7 % of Water. The second liquid phase contains 0.455 mole, made of 99.9 % of n-Butyl-Acetate and 0.01 % of Water.

Chemical components	**b**	**Liquid phase I**	**Liquid phase II**
n-Butyl-Acetate ($C_6H_{12}O_2$)	0.50	0.083	0.999
Water (H_2O)	0.50	0.917	0.001
Number of moles	1	0.545	0.455

Table 1.1: Phase equilibrium for an aerosol composed of n-Butyl-Acetate and Water at temperature 298 K and pressure 1 atm.

Let us denote by $\mathbf{x}_\alpha$, $\alpha = 1, 2$, the mole-fraction vector associated to the liquid phase α and by y_α, $\alpha = 1, 2$, the total number of moles in liquid phase α. For the example in Table 1.1 one has

$$\mathbf{x}_1 = \begin{pmatrix} 0.083 \\ 0.917 \end{pmatrix},\ \mathbf{x}_2 = \begin{pmatrix} 0.999 \\ 0.001 \end{pmatrix},\ y_1 = 0.545 \text{ and } y_2 = 0.455.$$

It ensues that the total organic mass is conserved through the relation

$$\sum_{\alpha=1}^{2} y_\alpha \mathbf{x}_\alpha = \mathbf{b}.$$

In this example the number of liquid phases present in the aerosol is equal to 2. Depending on the value of $\mathbf{b}$, the parameters characterizing the interactions between the chemical species, the temperature and the pressure in the system, this number changes. In fact its value always remains between 1 and s at constant temperature and pressure (Gibbs-Duhem relation [96]). Let us denote by p the maximal number of liquid phases in the aerosol. Thus the following inequalities hold: $1 \leq p \leq s$.

If a liquid phase α, $1 \leq \alpha \leq p$, is not present at equilibrium in the particle, then the corresponding total number of moles y_α is equal to 0. In the other case, if the liquid phase is present in the aerosol, then the value of y_α is positive, i.e. $y_\alpha > 0$. Consequently the following relation holds

$$y_\alpha \geq 0, \quad \forall\, \alpha = 1, \ldots, p.$$

Concerning the variable $\mathbf{x}_\alpha$, the definition of a mole-fraction vector implies that the sum of the components of $\mathbf{x}_\alpha$ is equal to 1. Moreover we make an assumption here that all phases contain all chemicals. On a mathematical point of view it yields

$$\mathbf{e}^T \mathbf{x}_\alpha = 1 \text{ and } \mathbf{x}_\alpha > 0, \quad \forall\, \alpha = 1, \ldots, p; \tag{1.4.1}$$

where $\mathbf{e}^T = (1, \ldots, 1)$ and the relation $\mathbf{x}_\alpha > 0$ is a shortened expression for $x_{\alpha,i} > 0$, $\forall\, i = 1, \ldots, s$, $\forall\, \alpha = 1, \ldots, p$.

The relation (1.4.1) is clearly satisfied by both liquid phases of the example in Table 1.1. In the case $y_\alpha = 0$, the corresponding mole-fraction vector $\mathbf{x}_\alpha$ does not take any particular physical value since the liquid phase is not present in the particle. Thus this case has no influence on the mass balance and the following relation still holds

$$\sum_{\alpha=1}^{p} y_\alpha \mathbf{x}_\alpha = \mathbf{b}.$$

1.4.2 Formulation of the phase equilibrium problem

With all the notations and relations given in the previous subsection, the phase equilibrium problem can now be defined. The determination of the number and composition of the liquid phases present in the aerosol particle of composition $\mathbf{b}$ is given by the global minimum of the following constrained minimization problem [5, 28, 96]

$$\begin{aligned} \min_{\{y_\alpha, \mathbf{x}_\alpha\}_{\alpha=1}^{p}} \quad & \sum_{\alpha=1}^{p} y_\alpha\, \bar{g}(\mathbf{x}_\alpha) \\ \text{s.t.} \quad & \sum_{\alpha=1}^{p} y_\alpha \mathbf{x}_\alpha = \mathbf{b}, \\ & y_\alpha \geq 0, \quad \mathbf{e}^T \mathbf{x}_\alpha = 1, \quad \mathbf{x}_\alpha > 0, \quad \alpha = 1, \ldots, p; \end{aligned} \tag{1.4.2}$$

where $\bar{g}$ is the molar Gibbs free energy.

In (1.4.2) the objective function gives the total Gibbs free energy of the particle whereas the constraints ensure the conservation of the mass balance and the characteristics of the variables y_α and $\mathbf{x}_\alpha$, $\alpha = 1, \ldots, p$.

Remark 1.4.1. *If in problem (1.4.2) there exist two phases* α *and* β *such that* $y_\alpha > 0$, $y_\beta > 0$ *and* $\mathbf{x}_\alpha = \mathbf{x}_\beta$, *it means that the phases* α *and* β *are similar. Then we combine both phases in one unique phase by setting* $y_\alpha = y_\alpha + y_\beta$ *and* $y_\beta = 0$. *This trick allows to assume that all liquid phases (and therefore* $\mathbf{x}_\alpha$*) present in the aerosol have distinct compositions.*

The molar Gibbs free energy function $\bar{g}$ is a homogeneous function of degree one that is usually defined by

$$\bar{g}(\mathbf{x}) = \mathbf{x}^T \boldsymbol{\mu}_l(\mathbf{x}), \tag{1.4.3}$$

where $\boldsymbol{\mu}_l(\mathbf{x})$ is the *chemical potential* vector for the mole-fraction vector $\mathbf{x}$ defined by

$$\boldsymbol{\mu}_l(\mathbf{x}) \;=\; \boldsymbol{\mu}_l^* + \mathcal{R}_c T \ln(\mathbf{a}_l(\mathbf{x})), \tag{1.4.4}$$

where $\mathbf{a}_l$ is the vector describing the *activity* of the species i, $i = 1, \ldots, s$ and $\boldsymbol{\mu}_l^*$ is the chemical potential at the hypothetical state for which $a_{l,i}(\mathbf{x})$, $i = 1, \ldots, s$, tends to 1 [96] (which corresponds to pure mixture). As one can observe in the relations (1.4.3) and (1.4.4), the function $\bar{g}$ depends on the chemical species constituting the aerosol particle. Hence the graph of $\bar{g}$ depends on the chemical species. Nevertheless the properties of $\bar{g}$ mentioned below hold for any chemical composition of the aerosol.

The function $\bar{g}$ is continuous on $\mathbb{R}^s_+$, belongs to $\mathcal{C}^\infty(\mathbb{R}^s_{++})$ with $\mathbb{R}_{++}$ denoting the set of real positive numbers, and is such that

$$\lim_{x_i \to 0} \frac{\partial \bar{g}}{\partial x_i}(x_1, \ldots, x_s) = -\infty, \qquad \forall i = 1, \ldots, s;$$

that is, the values of $\bar{g}$ approach finite limits as any given mole fraction tends to zero, and these limiting values are approached with negatively infinite slope.

From the definition of $\bar{g}$ and the fact that the temperature and the pressure are constant, the *Gibbs-Duhem equation* reads [96]

$$\mathbf{x}^T \boldsymbol{\nabla} \boldsymbol{\mu}_l(\mathbf{x}) = 0, \quad \forall \mathbf{x} \in \mathbb{R}^s_{++} \text{ s.t. } \mathbf{e}^T \mathbf{x} = 1.$$

This equation implies the 3 following relations $\forall \mathbf{x} \in \mathbb{R}^s_{++}$ satisfying $\mathbf{e}^T \mathbf{x} = 1$

$$\boldsymbol{\nabla}\, \bar{g}(\mathbf{x}) \;=\; \boldsymbol{\mu}_l(\mathbf{x}), \tag{1.4.5}$$

$$\bar{g}(\mathbf{x}) \;=\; \mathbf{x}^T \boldsymbol{\nabla} \bar{g}(\mathbf{x}), \tag{1.4.6}$$

$$\boldsymbol{\nabla}^2 \bar{g}(\mathbf{x})\, \mathbf{x} \;=\; \mathbf{0}. \tag{1.4.7}$$

Relationship (1.4.7) means that the pair $(0, \mathbf{x})$ is an eigen-pair for the matrix $\boldsymbol{\nabla}^2 \bar{g}(\mathbf{x})$.

The definitions of $\bar{g}$ and of the chemical potential for the nonideal solution $\mathbf{x}_\alpha$, $\alpha = 1, \ldots, p$ allow to transform the expression of the objective function in (1.4.2). Combining the relations (1.4.3) and (1.4.4) the objective function of (1.4.2) becomes

$$\begin{aligned}\sum_{\alpha=1}^{p} y_\alpha \bar{g}(\mathbf{x}_\alpha) &= \sum_{\alpha=1}^{p} y_\alpha\, \mathbf{x}_\alpha^T \boldsymbol{\mu}_l(\mathbf{x}_\alpha)\\ &= \sum_{\alpha=1}^{p} y_\alpha\, \mathbf{x}_\alpha^T\, \boldsymbol{\mu}_l^* + \mathcal{R}_c T \sum_{\alpha=1}^{p} y_\alpha\, \mathbf{x}_\alpha^T\, \ln(\mathbf{a}_l(\mathbf{x}_\alpha))\\ &= \left(\sum_{\alpha=1}^{p} y_\alpha\, \mathbf{x}_\alpha\right)^T \boldsymbol{\mu}_l^* + \mathcal{R}_c T \sum_{\alpha=1}^{p} y_\alpha\, \mathbf{x}_\alpha^T\, \ln(\mathbf{a}_l(\mathbf{x}_\alpha)),\end{aligned}$$

since $\boldsymbol{\mu}_l^*$ is a constant vector independent of $\mathbf{x}_\alpha$, $\alpha = 1, \ldots, p$.

With the mass balance equation this relation becomes

$$\sum_{\alpha=1}^{p} y_\alpha \bar{g}(\mathbf{x}_\alpha) = \mathbf{b}^T \boldsymbol{\mu}_l^* + \mathcal{R}_c T \sum_{\alpha=1}^{p} y_\alpha\, \mathbf{x}_\alpha^T\, \ln(\mathbf{a}_l(\mathbf{x}_\alpha)).$$

Since $\mathbf{b}^T \boldsymbol{\mu}_l^*$ and $\mathcal{R}_c T$ are constants, to minimize $\sum_{\alpha=1}^{p} y_\alpha \bar{g}(\mathbf{x}_\alpha)$ is equivalent to minimize $\sum_{\alpha=1}^{p} y_\alpha\, \mathbf{x}_\alpha^T\, \ln(\mathbf{a}_l(\mathbf{x}_\alpha))$. Then let us define the *normalized Gibbs free energy* by

$$g(\mathbf{x}) = \mathbf{x}^T \ln(\mathbf{a}_l(\mathbf{x})).$$

The relationship between the molar Gibbs free energy and its normalized version is given by

$$\bar{g}(\mathbf{x}) \;=\; \mathbf{x}^T \boldsymbol{\mu}_l^* \;+\; \mathcal{R}_c T\, g(\mathbf{x}), \quad \forall \mathbf{x} \in \mathbb{R}^s_{++}.$$

Consequently the function g inherits the properties of $\bar{g}$, in particular

$$g(\mathbf{x}) \;=\; \mathbf{x}^T \boldsymbol{\nabla} g(\mathbf{x}), \quad \forall \mathbf{x} \in \mathbb{R}^s_{++} \text{ s.t. } \mathbf{e}^T\mathbf{x} = 1. \tag{1.4.8}$$

The PEP is then equivalent to solve

$$\begin{aligned}\min_{\{y_\alpha, \mathbf{x}_\alpha\}_{\alpha=1}^p} \quad & \sum_{\alpha=1}^{p} y_\alpha\, g(\mathbf{x}_\alpha)\\ \text{s.t.} \quad & \sum_{\alpha=1}^{p} y_\alpha \mathbf{x}_\alpha = \mathbf{b},\\ & y_\alpha \geq 0, \quad \mathbf{e}^T\mathbf{x}_\alpha = 1, \quad \mathbf{x}_\alpha > 0, \quad \alpha = 1, \ldots, p.\end{aligned} \tag{1.4.9}$$

In the sequel, this formulation of the phase equilibrium is considered. Problem (1.4.9) is a nonconvex nonlinear constrained optimization problem with equality and inequality constraints. The molar Gibbs free energy g depends on interaction parameters between the chemical species composing the particle [39, 51]. The determination of these parameters is a difficult task and still an ongoing research [24]. The semi-empirical thermodynamic model UNIFAC is a well-established method for estimating the activity coefficients that define g . In this thesis, the molar Gibbs free energy is obtained using the UNIFAC model [39].

In the following subsection we present an overview of the theory of constrained optimization and its first-order optimality conditions.

1.4.3 First-order optimality conditions

A general formulation for the constrained optimization problem (1.4.9) is given by [75]

$$\min_{\mathbf{y}\in\mathbb{R}^n} f(\mathbf{y}) \quad \text{subject to} \begin{cases} c_i(\mathbf{y}) = 0, & i \in \mathscr{E}, \\ c_i(\mathbf{y}) \geq 0, & i \in \mathscr{I}, \end{cases} \tag{1.4.10}$$

where the functions f and c_i are all smooth, real-valued functions on a subset of $\mathbb{R}^n$, and $\mathscr{E}$ and $\mathscr{I}$ are two sets of indices.

In this subsection optimality conditions that characterize the solutions of constrained optimization problem (1.4.10) are presented. First let us define the feasible set Ω as the set of points that satisfy the constraints; that is

$$\Omega = \{\mathbf{y} \in \mathbb{R}^n \,|\, c_i(\mathbf{y}) = 0, \forall i \in \mathscr{E}; \, c_i(\mathbf{y}) \geq 0, \forall i \in \mathscr{I}\}. \tag{1.4.11}$$

A point $\mathbf{y}$ is said to be *feasible* if it belongs to the set Ω. So the problem (1.4.10) can be rewritten more compactly as

$$\min_{\mathbf{y}\in\Omega} f(\mathbf{y}).$$

At a feasible point $\mathbf{y}$, the inequality constraint $i \in \mathscr{I}$ is said to be *active* if $c_i(\mathbf{y}) = 0$ and *inactive* if the strict inequality $c_i(\mathbf{y}) > 0$ is satisfied. Thus let us define the *active set* $\mathscr{A}(\mathbf{y})$ as

Definition 1.4.1. *The* active set $\mathscr{A}(\mathbf{y})$ *at any feasible point* $\mathbf{y}$ *consists of the equality constraint indices from* $\mathscr{E}$ *together with the indices of the inequality constraints* i *for which* $c_i(\mathbf{y}) = 0$*; that is*

$$\mathscr{A}(\mathbf{y}) = \mathscr{E} \cup \{i \in \mathscr{I} \,|\, c_i(\mathbf{y}) = 0\}.$$

An important condition that is assumed to hold in the majority of optimization algorithms is the so-called *linear independence constraint qualification (LICQ).* The LICQ states that at any feasible point $\mathbf{y}$, the gradients of all the active constraints are linearly independent. The main purpose of the LICQ is to ensure that the set of constraints is well-defined, in a way that there are no redundant constraints or no constraints defined such that their gradients are always equal to zero.

Definition 1.4.2. *Given the point* $\mathbf{y}$ *and the active set* $\mathscr{A}(\mathbf{y})$*, the linear independence constraint qualification (LICQ) holds if the set of active constraint gradients* $\{\boldsymbol{\nabla} c_i(\mathbf{y}), i \in \mathscr{A}(\mathbf{y})\}$ *is linearly independent.*

Let us state now the first-order necessary conditions for $\mathbf{y}^*$ to be a local minimizer of (1.4.10). As a preliminary to stating the necessary conditions, let us define the *Lagrangian* function for the general problem (1.4.10)

$$\mathcal{L}(\mathbf{y}, \boldsymbol{\lambda}) = f(\mathbf{y}) - \sum_{i\in\mathscr{E}\cup\mathscr{I}} \lambda_i \, c_i(\mathbf{y}). \tag{1.4.12}$$

The vector $\boldsymbol{\lambda} = (\lambda_i)_{i\in\mathscr{E}\cup\mathscr{I}}$ is called the Lagrange multiplier vector.

The necessary conditions defined in the following theorem are called *first-order conditions* or *Karush-Kuhn-Tucker (KKT) conditions.* These conditions describe the properties of the gradients of the objective and constraint functions.

Theorem 1.4.1. *Suppose that* $\mathbf{y}^*$ *is a local solution of (1.4.10), that the functions* f *and* c_i *are continuously differentiable, and that the LICQ holds at* $\mathbf{y}^*$*. Then there is a Lagrange multiplier vector* $\boldsymbol{\lambda}^*$*, with components* λ_i^*, $i \in \mathscr{E} \cup \mathscr{I}$*, such that the following conditions are satisfied at* $(\mathbf{y}^*, \boldsymbol{\lambda}^*)$

$$\begin{aligned} \boldsymbol{\nabla}_{\mathbf{y}} \mathcal{L}(\mathbf{y}^*, \boldsymbol{\lambda}^*) &= \mathbf{0}, \\ c_i(\mathbf{y}^*) &= 0, \quad \forall\, i \in \mathscr{E}, \\ c_i(\mathbf{y}^*) &\geq 0, \quad \forall\, i \in \mathscr{I}, \\ \lambda_i^* &\geq 0, \quad \forall\, i \in \mathscr{I}, \\ \lambda_i^* c_i(\mathbf{y}^*) &= 0, \quad \forall\, i \in \mathscr{E} \cup \mathscr{I}. \end{aligned} \tag{1.4.13}$$

The point $\mathbf{y}^*$ *is called a* KKT-point.

The first condition of (1.4.13) states that $(\mathbf{y}^*, \boldsymbol{\lambda}^*)$ is a stationary point of the Lagrangian $\mathcal{L}$, the second and third groups of equations ensure that $\mathbf{y}^*$ is a feasible point of (1.4.10), the fourth group enforces the nonnegativity of the components of the Lagrange multiplier $\boldsymbol{\lambda}^*$, whereas the last set of equations are complementarity conditions which imply that either constraint i is active or $\lambda_i^* = 0$, or possibly both. In particular, the Lagrange multipliers corresponding to inactive inequality constraints are zero. Thus the terms for indices $i \notin \mathscr{A}(\mathbf{y}^*)$ can be omitted from the stationarity condition and this condition rewrites

$$0 \;=\; \boldsymbol{\nabla}_{\mathbf{y}} \mathcal{L}(\mathbf{y}^*, \boldsymbol{\lambda}^*) \;=\; \boldsymbol{\nabla} f(\mathbf{y}^*) \;-\; \sum_{i \in \mathscr{A}(\mathbf{y}^*)} \lambda_i^* \boldsymbol{\nabla} c_i(\mathbf{y}^*).$$

For a given problem (1.4.10) and solution point $\mathbf{y}^*$, there may be many vectors $\boldsymbol{\lambda}^*$ for which the KKT conditions (1.4.13) are satisfied. When the LICQ holds, however, the optimal $\boldsymbol{\lambda}^*$ is unique.

Let us examine if the LICQ holds for problem (1.4.9). First let us define the vector $\mathbf{y}$ in the case of problem (1.4.9):

$$\mathbf{y}^T \;=\; (\mathbf{x}_1^T, \ldots, \mathbf{x}_p^T, y_1, \ldots, y_p) \in \mathbb{R}^{sp}_{++} \times \mathbb{R}^p_+.$$

The vector $\mathbf{y}$ is of dimension $n \;=\; sp \,+\, p$. The equality and inequality constraints are then written as

$$\begin{aligned} c_i(\mathbf{y}) \;&=\; \sum_{\alpha=1}^{p} y_\alpha x_{\alpha,i} \,-\, b_i, \quad i = 1, \ldots, s; \\ c_{s+i}(\mathbf{y}) \;&=\; \mathbf{e}^T \mathbf{x}_i \,-\, 1, \quad i = 1, \ldots, p; \\ c_{s+p+i}(\mathbf{y}) \;&=\; y_i, \quad i = 1, \ldots, p. \end{aligned}$$

Lemma 1.4.2. *Assume that the vectors* $\mathbf{x}_\alpha$, $\alpha = 1, \ldots, p$*, are linearly independent. Then, the LICQ holds for (1.4.9) at any feasible point* $\mathbf{y}^T \;=\; (\mathbf{x}_1^T, \ldots, \mathbf{x}_p^T, y_1, \ldots, y_p)$.

Proof. Let $\mathbf{y}$ be a feasible point of (1.4.9). Without loss of generality let us suppose that the first $\tilde{p}$ constraints are active and the last $p-\tilde{p}$ constraints are inactive, namely:

$$y_\alpha \;=\; 0,\ \forall \alpha = 1,\dots,\tilde{p}, \qquad \text{and} \qquad y_\alpha \;>\; 0,\ \forall \alpha = \tilde{p}+1,\dots,p.$$

Then the matrix of active constraint gradients is given by

$$(\boldsymbol{\nabla} c_1(\mathbf{y}),\dots,\boldsymbol{\nabla} c_{s+p+\tilde{p}}(\mathbf{y})) \;=\; \begin{pmatrix} y_1\,\mathbf{I}_s & \mathbf{e} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ y_2\,\mathbf{I}_s & \mathbf{0} & \mathbf{e} & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ y_p\,\mathbf{I}_s & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{e} & \mathbf{0} \\ \mathbf{x}_1^T & 0 & 0 & \cdots & 0 & \mathbf{e}_1^T \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{x}_{\tilde{p}}^T & 0 & 0 & \cdots & 0 & \mathbf{e}_{\tilde{p}}^T \\ \mathbf{x}_{\tilde{p}+1}^T & 0 & 0 & \cdots & 0 & \mathbf{0}^T \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{x}_p^T & 0 & 0 & \cdots & 0 & \mathbf{0}^T \end{pmatrix},$$

where $\mathbf{e}^T \;=\; (1,\dots,1)$, $\mathbf{I}_s$ is the identity matrix of dimension $s\times s$, $\mathbf{e}_i^T$, $i=1,\dots,\tilde{p}$, are the standard basis vectors defined by $e_{ij} \;=\; \delta_{ij}$ (the Kronecker symbol), and $\mathbf{0}$ designs either a vector or a matrix whose elements are all equal to 0. Since the mole fraction vectors $\mathbf{x}_\alpha$, $\alpha=\tilde{p}+1,\dots,p$ are linearly independent, the column rank of the above matrix is equal to $s+p+\tilde{p}$. Consequently the set of active constraint gradients is linearly independent and the LICQ holds. □

Suppose that the point $(\mathbf{x}_1^T,\dots,\mathbf{x}_p^T,y_1,\dots,y_p)$ is a local solution of (1.4.9) and that the constraints functions c_i, $i=1,\dots,s{+}2p$ are continuously differentiable in the neighborhood of $(\mathbf{x}_1^T,\dots,\mathbf{x}_p^T,y_1,\dots,y_p)$, then by the Theorem 1.4.1 there exist Lagrange multipliers $\boldsymbol{\lambda}\in\mathbb{R}^s$, ζ_α, $\theta_\alpha\in\mathbb{R}$, $\alpha=1,\dots,p$, such that the following KKT conditions are satisfied

$$\begin{aligned} y_\alpha(\boldsymbol{\nabla} g(\mathbf{x}_\alpha)+\boldsymbol{\lambda}) \;+\; \zeta_\alpha\mathbf{e} \;&=\; \mathbf{0}, \quad \forall\,\alpha=1,\dots,p, && (1.4.14a)\\ g(\mathbf{x}_\alpha) \;+\; \boldsymbol{\lambda}^T\mathbf{x}_\alpha \;-\; \theta_\alpha \;&=\; 0, \quad \forall\,\alpha=1,\dots,p, && (1.4.14b)\\ \mathbf{b}-\sum_{\alpha=1}^{p} y_\alpha\,\mathbf{x}_\alpha \;&=\; \mathbf{0}, && (1.4.14c)\\ 1-\mathbf{e}^T\mathbf{x}_\alpha \;&=\; 0, \quad \forall\,\alpha=1,\dots,p, && (1.4.14d)\\ y_\alpha \;&\geq\; 0, \quad \forall\,\alpha=1,\dots,p, && (1.4.14e)\\ \theta_\alpha \;&\geq\; 0, \quad \forall\,\alpha=1,\dots,p, && (1.4.14f)\\ \theta_\alpha y_\alpha \;&=\; 0, \quad \forall\,\alpha=1,\dots,p, && (1.4.14g) \end{aligned}$$

where the complementary conditions for the Lagrange multipliers $\boldsymbol{\lambda}$ and ζ_α, $\alpha=1,\dots,p$, are omitted.

Let us define the sets of indices $\mathcal{A} = \{\alpha \in \{1,\ldots,p\} \,|\, y_\alpha = 0\}$ and $\mathcal{I} = \{\alpha \in \{1,\ldots,p\} \,|\, y_\alpha > 0\}$. The set $\mathcal{A}$ represents the set of indices of the active inequality constraints and is a subset of $\mathscr{A}$, and $\mathcal{I}$ is the set of inactive constraints. Let $p^{\mathcal{A}}$, resp. $p^{\mathcal{I}}$, be the cardinal of $\mathcal{A}$, resp. $\mathcal{I}$, such that $p^{\mathcal{A}} + p^{\mathcal{I}} = p$.

Remark 1.4.2. *In the sequel an exponent $\mathcal{I}$, resp. $\mathcal{A}$, is added to the variables y_α and $\mathbf{x}_\alpha$ to specify that $\alpha \in \mathcal{I}$, resp. $\mathcal{A}$. For instance, the expression $y_\alpha^{\mathcal{I}}$ stands for all y_α with $\alpha \in \mathcal{I}$. Moreover the notation $\alpha = 1,\ldots,p^{\mathcal{I}}$ is considered equivalent to $\forall \alpha \in \mathcal{I}$.*

If $y_\alpha > 0$, the complementary equation (1.4.14g) implies $\theta_\alpha = 0$. Hence for all liquid phases present in the particle the second stationarity equation (1.4.14b) becomes

$$g(\mathbf{x}_\alpha) + \boldsymbol{\lambda}^T \mathbf{x}_\alpha = 0.$$

Furthermore if the first stationarity equation (1.4.14a) is multiplied by $\mathbf{x}_\alpha$, and if the property (1.4.8) of g and the above equation are used, the following successive relations hold

$$\begin{aligned} \mathbf{0} &= y_\alpha \left(\boldsymbol{\nabla}^T g(\mathbf{x}_\alpha)\mathbf{x}_\alpha + \boldsymbol{\lambda}^T \mathbf{x}_\alpha \right) + \zeta_\alpha \, \mathbf{e}^T \mathbf{x}_\alpha \\ &= y_\alpha \left(g(\mathbf{x}_\alpha) + \boldsymbol{\lambda}^T \mathbf{x}_\alpha \right) + \zeta_\alpha \, \mathbf{e}^T \mathbf{x}_\alpha \\ &= \zeta_\alpha \, \mathbf{e}^T \mathbf{x}_\alpha. \end{aligned}$$

Since $\mathbf{e}^T \mathbf{x}_\alpha = 1$, $\forall \alpha = 1,\ldots,p$, it follows

$$\zeta_\alpha = 0, \; \forall \alpha \in \mathcal{I}.$$

This equality in (1.4.14a) gives

$$y_\alpha \left(\boldsymbol{\nabla} g(\mathbf{x}_\alpha) + \boldsymbol{\lambda} \right) = 0, \; \forall \alpha \in \mathcal{I}.$$

In conclusion the following relation is established for all liquid phases present in the aerosol

$$\boldsymbol{\nabla} g(\mathbf{x}_\alpha) = -\boldsymbol{\lambda}, \; \forall \alpha \in \mathcal{I}. \tag{1.4.15}$$

And it follows immediately

$$\boldsymbol{\nabla} g(\mathbf{x}_\alpha) = \boldsymbol{\nabla} g(\mathbf{x}_\beta), \; \forall \alpha, \beta \in \mathcal{I}. \tag{1.4.16}$$

Relations (1.4.15) and (1.4.16) mean that all liquid phases present in the aerosol have their gradients equal to the opposite of the Lagrange multiplier $\boldsymbol{\lambda}$ when they are at the KKT-point.

Before summarizing the equations that constitute the model, let us come back to the missing relation in (1.3.1), namely the computation of the gas-concentration vector $\mathbf{c}_g^{surf}$.

1.5 Computation of the gas-concentration vector $\mathbf{c}_g^{surf}$

Let us study the variable $\mathbf{c}_g^{surf}$ more precisely and begin this subsection with some chemistry arguments. Since the atmosphere can be considered as an ideal gas mixture with negligible error [88, 96] the following relation holds

$$p_i V \; = \; n_i \mathcal{R}_c T, \qquad \forall i =, 1, \ldots, s;$$

where p_i is the partial pressure of compound i and n_i is the number of moles of compound i, $i = 1, \ldots, s$. This relation is called the ideal gas law. Thanks to this relation an expression for $\mathbf{c}_g^{surf}$ can be given

$$\mathbf{c}_g^{surf} = \frac{1}{\mathcal{R}_c T} \, \mathbf{p}_g^{surf}, \tag{1.5.1}$$

where $\mathbf{p}_g^{surf}$ is the pressure of the gas at the particle surface.

For the determination of $\mathbf{p}_g^{surf}$ let us consider the chemical potential of the gaseous and liquid states. Since no reaction occurs in the gas-aerosol system, the gas-liquid equilibrium relation expresses the equality between the chemical potentials of the gaseous state and all the liquid phases present in the aerosol [32, 79], namely

$$\boldsymbol{\mu}_g^{surf} \; = \; \boldsymbol{\mu}_l(\mathbf{x}_1^{\mathcal{I}}) \; = \; \ldots \; = \; \boldsymbol{\mu}_l(\mathbf{x}_{p^{\mathcal{I}}}^{\mathcal{I}}),$$

where $\boldsymbol{\mu}_g^{surf}$ and $\boldsymbol{\mu}_l$ are respectively the chemical potential vector of size s for the gaseous and liquid states. Since the gas is supposed to be ideal, the chemical potential $\boldsymbol{\mu}_g^{surf}$ is equal to

$$\boldsymbol{\mu}_g^{surf} \; = \; \boldsymbol{\mu}_g^0 + \mathcal{R}_c T \ln(\mathbf{p}_g^{surf}),$$

where $\boldsymbol{\mu}_g^0$ is the standard chemical potential vector defined at a pressure of $1 \, atm$.

The chemical potential vector $\boldsymbol{\mu}_l$ in a non ideal solution (as it is usually the case with atmospheric aerosols) is recalled

$$\boldsymbol{\mu}_l(\mathbf{x}_\alpha) \; = \; \boldsymbol{\mu}_l^* + \mathcal{R}_c T \ln(\mathbf{a}_l(\mathbf{x}_\alpha)).$$

From the previous subsection the following relations hold

$$\begin{aligned} \boldsymbol{\nabla} g(\mathbf{x}_\alpha^{\mathcal{I}}) &= -\boldsymbol{\lambda}, \quad \forall \alpha \in \mathcal{I}, \\ \boldsymbol{\nabla} g(\mathbf{x}_\alpha) &= \ln(\mathbf{a}_l(\mathbf{x}_\alpha)), \, \forall \alpha = 1, \ldots, p. \end{aligned}$$

Then

$$\ln(\mathbf{a}_l(\mathbf{x}_\alpha^{\mathcal{I}})) \; = \; -\boldsymbol{\lambda}, \, \forall \alpha \in \mathcal{I},$$

and consequently

$$\boldsymbol{\mu}_l(\mathbf{x}_1^{\mathcal{I}}) \; = \; \ldots \; = \; \boldsymbol{\mu}_l(\mathbf{x}_{p^{\mathcal{I}}}^{\mathcal{I}}) \; = \; \boldsymbol{\mu}_l^* - \mathcal{R}_c T \boldsymbol{\lambda}.$$

The condition for the equilibrium between gas and the aerosol is shortened to

$$\boldsymbol{\mu}_g^{surf} \; = \; \boldsymbol{\mu}_l^* - \mathcal{R}_c T \boldsymbol{\lambda}.$$

Thus the gas-liquid equilibrium becomes

$$\ln(\mathbf{p}_g^{surf}) = -\boldsymbol{\lambda} + \frac{1}{\mathcal{R}_c T}\,(\boldsymbol{\mu}_l^* - \boldsymbol{\mu}_g^0). \tag{1.5.2}$$

The values of $\boldsymbol{\mu}_g^0$ and $\boldsymbol{\mu}_l^*$ are often not available and difficult to measure. For that reason let us transform this expression further. Consider the relation (1.5.2) and suppose that there is only one component in the system and that this component is pure. Then the activity and the pressure become equal to

$$\begin{aligned} \mathbf{a}_l &= \mathbf{1} \Rightarrow \boldsymbol{\lambda} = -\ln(\mathbf{a}_l) = \mathbf{0}, \\ \mathbf{p}_g^{surf} &= \mathbf{p}_g^o, \end{aligned}$$

where $\mathbf{p}_g^o$ is the vapor pressure of the gas. In this case the expression (1.5.2) is transformed in

$$\ln(\mathbf{p}_g^o) = \frac{1}{\mathcal{R}_c T}\,(\boldsymbol{\mu}_l^* - \boldsymbol{\mu}_g^0). \tag{1.5.3}$$

Thus an equivalent expression for the term $\frac{1}{\mathcal{R}_c T}\,(\boldsymbol{\mu}_l^* - \boldsymbol{\mu}_g^0)$ is obtained. The advantage of this expression is that $\mathbf{p}_g^o$ is known and given in traditional chemical tables. Thanks to (1.5.3) the relation for the pressure at the surface particle becomes

$$\ln(\mathbf{p}_g^{surf}) = -\boldsymbol{\lambda} + \ln(\mathbf{p}_g^o); \tag{1.5.4}$$

and the new expression of the gas concentration-vector at the surface is given by

$$\mathbf{c}_g^{surf} = \frac{1}{\mathcal{R}_c T}\,\exp\left(-\boldsymbol{\lambda} + \ln(\mathbf{p}_g^o)\right). \tag{1.5.5}$$

The variable $\mathbf{c}_g^{surf}$ can also be expressed in term of $\mathbf{x}_\alpha^{\mathcal{I}}$ by writting

$$\mathbf{c}_g^{surf} = \frac{1}{\mathcal{R}_c T}\,\exp\left(\boldsymbol{\nabla} g(\mathbf{x}_\alpha^{\mathcal{I}}) + \ln(\mathbf{p}_g^o)\right), \tag{1.5.6}$$

where any inactive constraint can be considered to define the value $\boldsymbol{\nabla} g(\mathbf{x}_\alpha^{\mathcal{I}})$.

1.6 Formulation of the coupled model

The purpose of this thesis is the modeling and computation of the gas-particle partitioning and phase equilibrium for organic aerosol particles. In Section 1.3 the ordinary differential equations for modeling the mass transfer between the aerosols and the surrounding gas were formulated as

$$\begin{aligned} \frac{d}{dt}\mathbf{c}_g^\infty(t) &= -N\,\mathbf{j}\left(\mathbf{c}_g^\infty(t), \mathbf{c}_g^{surf}(t), R(t)\right), \\ \frac{d}{dt}\mathbf{b}(t) &= \mathbf{j}\left(\mathbf{c}_g^\infty(t), \mathbf{c}_g^{surf}(t), R(t)\right), \\ R(t) &= \left(\frac{3}{4\pi}\sum_{i=1}^{s}\frac{m_{c,i}b_i(t)}{\rho_i}\right)^{\frac{1}{3}}, \end{aligned} \tag{1.6.1}$$

with the expressions for the flux $\mathbf{j}$ and the gas concentration $\mathbf{c}_g^{surf}$

$$\begin{aligned}\mathbf{j}(\mathbf{c}_g^{\infty}(t), \mathbf{c}_g^{surf}(t), R(t)) &= \mathbf{H}(R(t))\left(\mathbf{c}_g^{\infty}(t) - \mathbf{c}_g^{surf}(t)\right),\\ \mathbf{c}_g^{surf}(t) &= \frac{1}{\mathcal{R}_c T}\exp\left(\boldsymbol{\nabla} g(\mathbf{x}_\alpha^{\mathcal{I}}(t)) + \ln(\mathbf{p}_g^o)\right).\end{aligned}$$

Moreover the determination of the number and composition of the liquid phases present in each aerosol was presented in Section 1.4 and consists in solving the following minimization problem

$$\begin{aligned}\{\mathbf{x}_\alpha(t), y_\alpha(t)\}_{\alpha=1}^p &= \underset{\{\bar{\mathbf{x}}_\alpha, \bar{y}_\alpha\}_{\alpha=1}^p}{\operatorname{argmin}} \sum_{\alpha=1}^p \bar{y}_\alpha\, g(\bar{\mathbf{x}}_\alpha) \\ &\qquad \text{s.t.} \ \sum_{\alpha=1}^p \bar{y}_\alpha \bar{\mathbf{x}}_\alpha = \mathbf{b}(t), \\ &\qquad \mathbf{e}^T\bar{\mathbf{x}}_\alpha = 1,\ \bar{\mathbf{x}}_\alpha > 0,\ \bar{y}_\alpha \geq 0,\ \alpha = 1, \ldots, p.\end{aligned} \tag{1.6.2}$$

The formulation is slightly different from (1.4.9). The motivation is to better represent $\{\mathbf{x}_\alpha, y_\alpha\}_{\alpha=1}^p$ as the minimizer of the optimization problem, i.e. the points that realize the global minimum of the total Gibbs free energy of the particle.

Hence the coupling between the ordinary differential system (1.6.1) and the optimization problem (1.6.2) is now clear. In (1.6.1) the expression of the mass flux depends on $\boldsymbol{\nabla} g(\mathbf{x}_\alpha^{\mathcal{I}})$ whereas $\mathbf{b}$ appears in the equality constraint on the mass balance of (1.6.2).

Equations from (1.6.1) and (1.6.2) form the complete system that models the time evolution of the gas-aerosol system. However this set of equations can be written on a reduced form. If the second differential equation of (1.6.1) is multiplied by N and the result is added to the first differential equation, then the below relation is obtained

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{c}_g^{\infty} + N\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{b} = \mathbf{0},$$

which is equivalent to

$$\frac{\mathrm{d}}{\mathrm{d}t}(\mathbf{c}_g^{\infty} + N\,\mathbf{b}) = \mathbf{0}.$$

Therefore the quantity $\mathbf{c}_g^{\infty} + N\,\mathbf{b}$ is constant, since the gas-aerosol system is supposed to be closed. Let us denote by $\mathbf{b}^{\text{tot}}$ the total concentration-vector of the system. This vector is thus constant and equal to

$$\mathbf{c}_g^{\infty} + N\,\mathbf{b} = \mathbf{b}^{\text{tot}}. \tag{1.6.3}$$

Consequently once the value of $\mathbf{b}$ is known, the value of $\mathbf{c}_g^{\infty}$ is immediately deduced. The first differential equation in (1.6.1) can then be omitted from the set of equations. Moreover the flux $\mathbf{j}$ can be rewritten as

$$\mathbf{j}(\mathbf{c}_g^{\infty}, \mathbf{c}_g^{surf}, R) = \mathbf{j}(\mathbf{b}, \mathbf{x}_\alpha^{\mathcal{I}}, R) = \mathbf{H}(R)\left(\mathbf{b}^{\text{tot}} - N\,\mathbf{b} - \frac{1}{\mathcal{R}_c T}\exp(\boldsymbol{\nabla} g(\mathbf{x}_\alpha^{\mathcal{I}}) + \ln(\mathbf{p}_g^o))\right),$$

and the variable $\mathbf{c}_g^{surf}$ may also be removed from the set of equations.

Finally the final formulation of the model describing the evolution of the aerosol particles embedded in a gas phase is given by: find $\mathbf{b}, \mathbf{x}_\alpha : (0,T) \to \mathbb{R}_+^s$ and $R, y_\alpha : (0,T) \to \mathbb{R}_+$, $\alpha = 1, \dots, p$ satisfying

$$\begin{aligned}
\frac{d}{dt}\mathbf{b}(t) &= \mathbf{H}(R(t))\left(\mathbf{b}^{\text{tot}} - N\mathbf{b}(t) - \frac{1}{\mathcal{R}_c T}\exp(\boldsymbol{\nabla} g(\mathbf{x}_\alpha^{\mathcal{I}}(t)) + \ln(\mathbf{p}_g^o))\right), \\
R(t) &= \frac{3}{4\pi}\left(\sum_{i=1}^{s}\frac{m_{c,i}b_i(t)}{\rho_i}\right)^{\frac{1}{3}}, \\
\{\mathbf{x}_\alpha(t), y_\alpha(t)\}_{\alpha=1}^p &= \underset{\{\bar{\mathbf{x}}_\alpha, \bar{y}_\alpha\}_{\alpha=1}^p}{\operatorname{argmin}} \sum_{\alpha=1}^{p} \bar{y}_\alpha\, g(\bar{\mathbf{x}}_\alpha) \\
&\qquad \text{s.t.} \sum_{\alpha=1}^{p}\bar{y}_\alpha\bar{\mathbf{x}}_\alpha = \mathbf{b}(t), \\
&\qquad \mathbf{e}^T\bar{\mathbf{x}}_\alpha = 1,\ \bar{\mathbf{x}}_\alpha > 0,\ \bar{y}_\alpha \geq 0,\ \alpha = 1, \dots, p
\end{aligned} \tag{1.6.4}$$

where T is the final time of integration and the initial condition is given by $\mathbf{b}(0) = \mathbf{b}_0$ with $\mathbf{b}_0$, a given initial composition-vector.

The value of $\mathbf{c}_g^\infty$ and $\mathbf{c}_g^{surf}$ are then computed on the following manner

$$\begin{aligned}
\mathbf{c}_g^\infty &= \mathbf{b}^{\text{tot}} - N\,\mathbf{b}, \\
\mathbf{c}_g^{surf} &= \frac{1}{\mathcal{R}_c T}\exp(\boldsymbol{\nabla} g(\mathbf{x}_\alpha^{\mathcal{I}}) + \ln(\mathbf{p}_g^o)).
\end{aligned}$$

Remark 1.6.1. *In (1.6.4) the variable R could also be removed from the system. However in that case the expression of the flux becomes more complicated because of the definition of* $\mathbf{H}$. *For more readability in (1.6.4) we have chosen to keep the variable R.*

The system (1.6.4) couples ordinary differential equations and a minimization problem with mixed constraints. The ODE are nonlinear. Due to the possible wide range of time for evaporation and condensation among the different chemical species in the gas-aerosol system, the ODE form a stiff system [60, 83, 110]. The objective function in the minimization problem is nonconvex and nonlinear, and one of the equality constraint is nonlinear. Hence the numerical method that solves the system (1.6.4) has to handle stiff and nonlinear ODE coupled with a nonconvex, nonlinear minimization problem with mixed constraints. Moreover the variables y_α and $\mathbf{x}_\alpha$ lose regularity when one variable $y_\alpha(t) > 0$ vanishes (*activation of an inequality constraint*) or, reciprocally, when one variable $y_\alpha(t) = 0$ becomes strictly positive (*deactivation of an inequality constraint*). The numerical method has also to accurately detect and compute the times of activation and deactivation of constraints in order to guarantee the correctness and accuracy of the solution.

Two different approaches to solve (1.6.4) are proposed in this thesis. The first approach consists of a time splitting between the differential equations and the optimization problem.

Since an efficient method that solves the minimization of the Gibbs free energy of the particle has been developed in [4, 5], the aim of the first approach is to adapt the method of Amundson et al. and include it in a dynamical system. With this approach the chemical and physical characteristics of the problem are conserved.

The second approach leaves the physics and chemistry aside and exploits the mathematical properties of (1.6.4). If the number of active inequality constraints is fixed, the considered system can be associated to a system of differential algebraic equations (DAE), by replacing the minimization problem by its first order optimality conditions (Theorem 1.4.1). In that case, since the computation of the *global minimum of energy* is required, uniqueness is lost and the solutions may bifurcate between branches of global optima, local optima or saddle-points. Efficient techniques to solve DAE systems relying on implicit Runge-Kutta methods have been developed in [7, 49, 50, 61]. The second approach follows the works of E. Hairer and G. Wanner in [50].

Both techniques are presented in the sequel, with numerical results and some theoretical considerations. Before their description let us give a geometrical interpretation of (1.6.4) that is helpful to understand both numerical methods.

1.7 Geometric interpretation

A geometric interpretation of (1.6.4) is useful to understand the dynamics of the system and design efficient numerical techniques. First let us consider the optimization problem solely with a fixed point $\mathbf{b}$ and observe that if $\{y_\alpha, \mathbf{x}_\alpha\}_{\alpha=1}^p$ is solution of the minimization problem for $\mathbf{b}$, then for any $c > 0$, $\{cy_\alpha, \mathbf{x}_\alpha\}_{\alpha=1}^p$ is the solution of the minimization problem for the point $c\mathbf{b}$. Therefore, without loss of generality, it is assumed that $\mathbf{e}^T\mathbf{b} = 1$ in this section. The hereafter interpretation follows [5] and starts with the projection of the optimization problem on a reduced space of lower dimension.

Let Δ'_s be defined by $\Delta'_s = \{\mathbf{x} \in \mathbb{R}^s | \mathbf{e}^T\mathbf{x} = 1, \mathbf{x} \geq 0\}$ and for $r = s - 1$ denote $\Delta_r = \{\mathbf{z} \in \mathbb{R}^r | \mathbf{e}^T\mathbf{z} \leq 1, \mathbf{z} \geq 0\}$. The unit simplex Δ_r can be identified with Δ'_s via the mapping

$$\begin{aligned} \Pi\,:\,\Delta_r &\rightarrow \Delta'_s \\ \mathbf{z} &\mapsto \mathbf{x} = \mathbf{e}_s + Z_e\mathbf{z}, \end{aligned}$$

where $\mathbf{e}_s$ is the canonical basis vector and $Z_e^T = (\mathbf{I}_r, -\mathbf{e})$ with $\mathbf{I}_r$ the $r \times r$ identity matrix. Let $f = g\,o\,\Pi$. Then f belongs to the space E given by

$$E = \{f \in \mathcal{C}^\infty(\mathrm{int}\Delta_r) \,|\, f \in \mathcal{C}^0(\Delta_r),\ \partial f(\mathbf{z}) = \emptyset \text{ for } \mathbf{z} \in \partial\Delta_r\},$$

where $\partial f(\mathbf{z})$ represents the subdifferential of f at $\mathbf{z}$ [109].

Let P be the projection from $\mathbb{R}^s$ to $\mathbb{R}^r$ defined by $P(x_1, \ldots, x_r, x_s) = (x_1, \ldots, x_r)$, and denote $\mathbf{z}_\alpha = P\mathbf{x}_\alpha$ for $\alpha = 1, \ldots, p$, and $\mathbf{d} = P\mathbf{b}$. The minimization problem in (1.6.4) is equivalent after projection to

$$\begin{aligned} \min_{\{y_\alpha,\mathbf{z}_\alpha\}_{\alpha=1}^p} \quad & \sum_{\alpha=1}^p y_\alpha f(\mathbf{z}_\alpha), \\ \text{s.t.} \quad & \sum_{\alpha=1}^p y_\alpha \mathbf{z}_\alpha = \mathbf{d}, \\ & \sum_{\alpha=1}^p y_\alpha = 1, \\ & y_\alpha \geq 0, \quad \alpha = 1,\dots,p. \end{aligned} \tag{1.7.1}$$

Since the domain of f is Δ_r, the condition $\mathbf{z}_\alpha \in \Delta_r$ does not need to be included as constraint in (1.7.1).

Let us recall that the convex envelope of a set of points $Z \subset \mathbb{R}^r$ is the minimal convex set containing Z and the convex envelope of the real function f is the largest convex extended real-valued function majorized by f on Δ_r. In convex geometry Carathéodory's theorem is a classical result that states if a point lies in a convex envelope [93]. This theorem reads

Theorem 1.7.1. *Let Z be any set included in $\mathbb{R}^r$ and let $C = \mathrm{conv} Z$, the convex envelope of Z. Then $\mathbf{z} \in C$ if and only if $\mathbf{z}$ can be expressed as a convex combination of $r+1$ of the points and directions in Z (not necessarily distinct).*

A direct consequence of the Carathéodory's theorem implies that the problem (1.7.1) is equivalent to the determination of the convex envelope of f at point $\mathbf{d}$ [5]. This result is stated in the following theorem.

Theorem 1.7.2. *For every $\mathbf{d} \in \Delta_r$, the minimum of (1.7.1) is $\mathrm{conv}\, f(\mathbf{d})$, the value of the convex envelope of f at $\mathbf{d}$. Moreover, one has*

$$\mathrm{conv} f(\mathbf{d}) = \sum_{\alpha=1}^p y_\alpha f(\mathbf{z}_\alpha)$$

for some convex combination $\mathbf{d} = \sum_{\alpha=1}^p y_\alpha \mathbf{z}_\alpha$, $\sum_{\alpha=1}^p y_\alpha = 1$, $y_\alpha \geq 0$, $\alpha = 1,\dots,p$. The point $(\mathbf{z}_1^T,\dots,\mathbf{z}_p^T, y_1,\dots,y_p) \in \mathbb{R}^{sp}$ is called a phase splitting *of $\mathbf{d}$.*

A phase splitting is called *stable* if $y_\alpha > 0$ for all $\alpha = 1,\dots,p$ and $\mathbf{z}_\alpha$ are distinct. Note that any phase splitting can be transformed into a stable phase splitting by considering only the subset $\{\mathbf{z}_\alpha : y_\alpha > 0\}$ or the point $(\mathbf{z}_1^{\mathcal{I},T},\dots,\mathbf{z}_{p^{\mathcal{I}}}^{\mathcal{I},T}, y_1^{\mathcal{I}},\dots,y_{p^{\mathcal{I}}}^{\mathcal{I}})$.

The following result states the existence and uniqueness of the stable phase splitting for a given $\mathbf{d}$ and characterizes the geometrical structure of $\mathrm{conv}\, f(\mathbf{d})$.

Theorem 1.7.3. *There exists a residual set R of E such that for any function $f \in R$, every $\mathbf{d} \in \Delta_r$ has a unique stable phase splitting. More precisely, there exists a unique $(p^{\mathcal{I}}-1)$-simplex $\sum(\mathbf{d}) = \mathrm{conv}(\mathbf{z}_1^{\mathcal{I}},\dots,\mathbf{z}_{p^{\mathcal{I}}}^{\mathcal{I}})$ with $p^{\mathcal{I}} \leq s$ such that $\mathrm{conv}\, f(\mathbf{d}) = \sum_{\alpha\in\mathcal{I}} y_\alpha^{\mathcal{I}} f(\mathbf{z}_\alpha^{\mathcal{I}})$ with the barycentric representation $\mathbf{d} = \sum_{\alpha\in\mathcal{I}} y_\alpha^{\mathcal{I}} \mathbf{z}_\alpha^{\mathcal{I}}$, $\sum_{\alpha\in\mathcal{I}} y_\alpha^{\mathcal{I}} = 1$ and $y_\alpha^{\mathcal{I}} > 0$, $\forall \alpha \in \mathcal{I}$.*

The proof of this result can be found in [89]. From now on the function f is assumed to belong to R. For a given $\mathbf{d} \in \mathrm{int}\Delta_r$, the $(p^{\mathcal{I}}-1)$-simplex $\sum(\mathbf{d})$ is called the *phase simplex of* $\mathbf{d}$.

In phase equilibrium theory the Gibbs tangent plane criterion is an important tool to determine the correctness of the phase repartition. The *Gibbs tangent plane criterion* [37, 68] states that a $(p^{\mathcal{I}}-1)$-simplex $\sum(\mathbf{d}) = \mathrm{conv}(z_1^{\mathcal{I}}, \ldots, z_{p^{\mathcal{I}}}^{\mathcal{I}})$ is a phase simplex if and only if there exist multipliers $\boldsymbol{\eta} \in \mathbb{R}^r$ and $\gamma \in \mathbb{R}$ such that

$$\begin{aligned} \boldsymbol{\nabla} f(\mathbf{z}_\alpha^{\mathcal{I}}) + \boldsymbol{\eta} &= \mathbf{0}, \qquad \forall \alpha \in \mathcal{I}, \quad (1.7.2)\\ f(\mathbf{z}_\alpha^{\mathcal{I}}) + \boldsymbol{\eta}^T \mathbf{z}_\alpha^{\mathcal{I}} + \gamma &= 0, \qquad \forall \alpha \in \mathcal{I}, \quad (1.7.3)\\ f(\mathbf{z}) + \boldsymbol{\eta}^T \mathbf{z} + \gamma &\geq 0, \qquad \forall \mathbf{z} \in \Delta_r. \quad (1.7.4) \end{aligned}$$

Geometrically, the above relations stipulate that the affine hyperplane tangent to the graph of f at $(\mathbf{z}_\alpha^{\mathcal{I}}, f(\mathbf{z}_\alpha^{\mathcal{I}}))$, $\forall \alpha \in \mathcal{I}$ lies entirely below the graph of f. This hyperplane is called the *supporting tangent plane.*

Let us consider a last definition. A point $\mathbf{d} \in \mathrm{int}\ \Delta_r$ is said to be a *single-phase point* if and only if $\mathrm{conv}\, f(\mathbf{d}) = f(\mathbf{d})$. The following result states that the vertices of a phase simplex are single-phase points.

Theorem 1.7.4. *Consider* $\mathbf{d} \in$ *int* Δ_r *and* $\Sigma(\mathbf{d}) = \mathrm{conv}(\mathbf{z}_1^{\mathcal{I}}, \ldots, \mathbf{z}_{p^{\mathcal{I}}}^{\mathcal{I}})$ *the phase simplex of* $\mathbf{d}$. *Then for all* $\alpha \in \mathcal{I}$, $\mathbf{z}_\alpha^{\mathcal{I}} \in \mathrm{int}\Delta_r$ *and* $\mathrm{conv}\, f(\mathbf{z}_\alpha^{\mathcal{I}}) = f(\mathbf{z}_\alpha^{\mathcal{I}})$.

Finally the Corollary 1.7.5 states that the *Gibbs tangent plane criterion* (1.7.4) is equivalent to the *single-phase point criterion* to determine whether a “tangent” simplex is a phase simplex.

Corollary 1.7.5. *Let* $\Sigma = \mathrm{conv}(\mathbf{z}_1^{\mathcal{I}}, \ldots, \mathbf{z}_{p^{\mathcal{I}}}^{\mathcal{I}})$ *be a* $(p^{\mathcal{I}}-1)$*-simplex. Assume that the vertices of* Σ *are single-phase points. If there exist multipliers* $\boldsymbol{\eta} \in \mathbb{R}^r$ *and* $\gamma \in \mathbb{R}$ *satisfying conditions (1.7.2) and (1.7.3), then* Σ *is a phase simplex.*

Following these definitions and theorems, let us interpret the phase equilibrium problem geometrically. The energy function of g (and therefore of f) depends on the chemical components present in the aerosol and the temperature and the pressure present in the system. Nevertheless, for organic aerosols the graph of f is at the most composed of $r+1$ convex regions lying in the neighborhood of the vertices of Δ_r. Let us consider first the case of an aerosol made of 2 chemical components. Thereby $s = 2$, $r = 1$ and Δ_r is the interval $[0, 1]$. A generic representation of f is given in Figure 1.2 (namely the case when the maximum number of convex areas is reached). For the points $\mathbf{d}$ considered on the left and right graphs, the value of the convex envelope of f at $\mathbf{d}$ is equal to the value of f at $\mathbf{d}$ and $\mathrm{conv}\, f(\mathbf{d}) = f(\mathbf{z}_\alpha)$. This implies that the stable phase splitting of $\mathbf{d}$ is given by $(\mathbf{z}^T, y)$ with $p^{\mathcal{I}} = 1$, $\mathbf{z} = \mathbf{d}$ and $y = 1$, and that $\mathbf{d}$ is a single-phase point.

On the central graph of Figure 1.2 the convex envelope of f considered at points $\mathbf{d}$ is no longer superposed with f but follows the segment given by $[f(\mathbf{z}_1), f(\mathbf{z}_2)]$. Hence the minimum of (1.7.1) is given by $\mathrm{conv}\, f(\mathbf{d}) = y_1 f(\mathbf{z}_1) + y_2 f(\mathbf{z}_2)$, the stable phase splitting

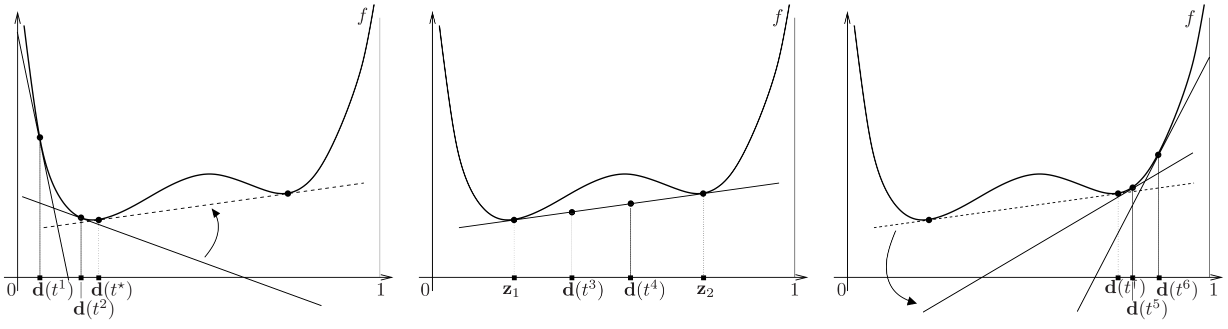


Figure 1.2: Geometric representation of the dynamic computation of the convex envelope. For a sequence of times $t^1 < t^2 < t^\star < t^3 < t^4 < t^\dagger < t^5 < t^6$, the vector $\mathbf{d}(t)$ moves from left to right. The supporting tangent plane follows the tangential slope at point $\mathbf{d}(t)$. Deactivation occurs at time $t^\star$ when the tangent plane (dashed line) touches the graph of f; Activation occurs at time $t^\dagger$ when the tangent plane (dashed line) gets released from the graph of f.

of $\mathbf{d}$ is $(\mathbf{z}_1^T, \mathbf{z}_2^T, y_1, y_2)$ with $p^{\mathcal{I}} = 2$ and $y_1 + y_2 = 1$, and the phase simplex of $\mathbf{d}$ is equal to $\operatorname{conv}(\mathbf{z}_1, \mathbf{z}_2)$ where the vertices $\mathbf{z}_1$ and $\mathbf{z}_2$ are single-phase points.

Each single-phase point is associated to a convex region of f. We denote by $\Delta_{r,\alpha}$ the part of Δ_r that corresponds to the convex region of f associated to $\mathbf{z}_\alpha$, and by $\Delta'_{s,\alpha}$ the image of $\Delta_{r,\alpha}$ through Π. Without loss of generality the single phase $\mathbf{z}_1$ is affiliated to the convex region situated on the left, and $\mathbf{z}_2$ to the one on the right. Hence the phase splitting is defined by $(\mathbf{z}_1^T, y_1)$ on the left graph of Figure 1.2 and by $(\mathbf{z}_2^T, y_2)$ on the right graph.

In Figure 1.2 the supporting tangent plane is drawn for all considered $\mathbf{d}$. It can be observed that every hyperplane lies below the graph of f as the Gibbs tangent plane criterion stated. When $\mathbf{d}$ is a single-phase point, the tangent plane is in contact with f at the point $(\mathbf{d}^T, f(\mathbf{d}))$ solely. When the phase simplex of $\mathbf{d}$ is given by $\operatorname{conv}(\mathbf{z}_1, \mathbf{z}_2)$ the tangent plane touches f at $(\mathbf{z}_1^T, f(\mathbf{z}_1))$ and $(\mathbf{z}_2^T, f(\mathbf{z}_2))$.

The domain Δ_r is then split into 3 areas according to the number of inactive inequality constraints. The separated domain is called a *phase diagram.* For the example in Figure 1.2 the interval $[0, 1]$ is separated as follow

$$[0,1] \;=\; [0, z_1] \cup]z_1, z_2[\cup [z_2, 1].$$

The digit 1 is associated to the areas $[0, z_1]$ and $[z_2, 1]$, and the digit 2 is given to $]z_1, z_2[$ in order to represent the number of inactive inequality constraints in each area. The areas with digit 1 are called *area 1* and the area with digit 2 is called *area 2.*

Let us consider the case where $\mathbf{b}$ (and therefore $\mathbf{d}$) evolves in time. The points $\mathbf{b}(t)$ are no longer supposed to be normalized. Conforming to the previous theory the points $\frac{1}{\mathbf{e}^T\mathbf{b}(t)}\mathbf{b}(t)$ lie in Δ'_s and the points $\mathbf{d}(t)$ represent the projection of $\frac{1}{\mathbf{e}^T\mathbf{b}(t)}\mathbf{b}(t)$ onto the simplex Δ_r. The time evolution of $\mathbf{b}$ is governed by the differential equation of (1.6.4) and requires the time-dependent computation of the stable phase simplex $\Sigma(\mathbf{d}(t))$. The activation/deactivation of constraints therefore corresponds to a change of dimension of the corresponding phase simplex $\Sigma(\mathbf{d}(t))$. In particular, the deactivation of a constraint can be interpreted as a *new tangential contact between the supporting tangent plane and the graph of the function f.*

Figure 1.2 shows the motion of the supporting tangent plane in one dimension of space, when the point $\mathbf{b}$ goes from left to right. When the tangent plane becomes in contact with the right convex region, one constraint is deactivated and the phase simplex' size increases by one ($p^{\mathcal{I}} = 1$ becomes $p^{\mathcal{I}} = 2$). Reciprocally, when the tangent plane leaves contact with the graph of f, the phase simplex' size decreases by one ($p^{\mathcal{I}} = 2$ becomes $p^{\mathcal{I}} = 1$ again).

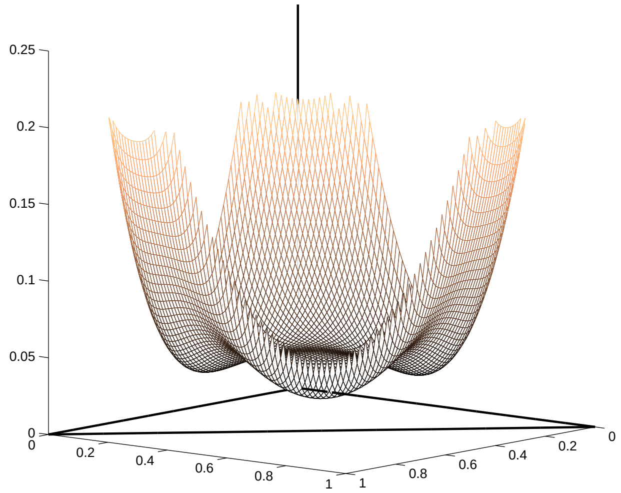


Figure 1.3: A generic representation of the reduced Gibbs free energy of a particle made of 3 chemical components.

Let us also study an aerosol particle made of 3 chemical components. In this case $r = 2$ and Δ_2 is the unit triangle defined by

$$\Delta_2 = \{\mathbf{z} \in \mathbb{R}^2 \,|\, \mathbf{e}^T\mathbf{z} \le 1, \mathbf{z} \ge 0\}.$$

In [56] Jiang et al. propose a generic form of the reduced energy f used in the chemical engineering literature. In the case $r = 2$ this form reads for $\mathbf{z} = (z_1, z_2) \in \Delta_2$

$$\begin{aligned} f(z_1, z_2) &= 0.76\, z_1 + 0.77\, z_2 + 0.78\,(1 - z_1 - z_2) \\ &\quad + z_1 \ln(z_1) + z_2 \ln(z_2) + (1 - z_1 - z_2)\ln(1 - z_1 - z_2) \\ &\quad + 10 z_1 z_2 \ln(1 - z_1 - z_2). \end{aligned}$$

The reduced energy f is depicted in Figure 1.3. For more visibility the graph is truncated at 0.2. The energy function f contains 3 distinct convex areas. Hence the maximal number of contact points between the supporting tangent plane and the surface is equal to 3 and the phase diagram of the particle is separated between areas where 1, 2 or 3 inequality constraints are inactive. The phase diagram is illustrated in Figure 1.4. As for the case $r = 1$, the areas with one inactive inequality constraint are situated in the neighborhood of the corners of Δ_2 and the indices of the phases are attributed as follow

- phase 1: bottom right corner,
- phase 2: top left corner,
- phase 3: bottom left corner.

In the interior of Δ_2 lays the area where all inequality constraints are inactive. The 3 remaining areas in Δ_2 are for the regions with 2 inactive constraints.

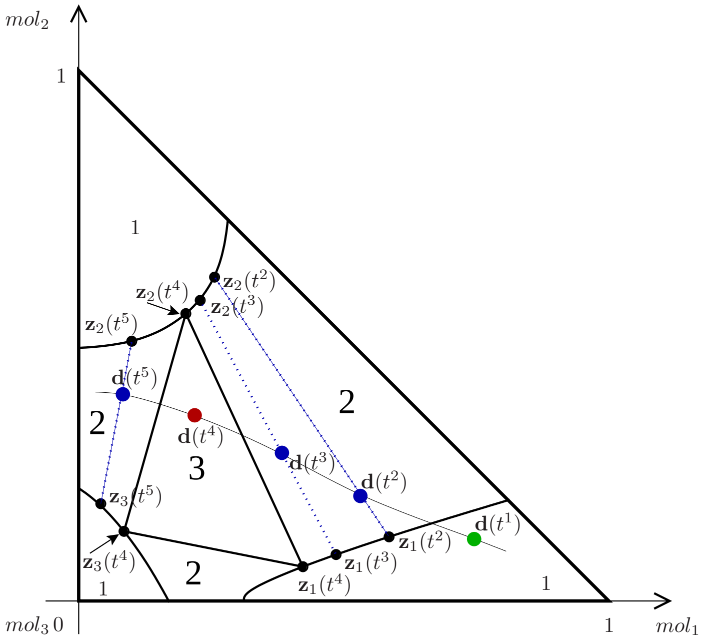


Figure 1.4: Trajectory of the composition-vector $\mathbf{d}$ on the phase diagram for a sequence of times $t^1 < t^2 < t^3 < t^4 < t^5$ when the aerosol is made of 3 chemical components.

On the phase diagram of Figure 1.4 the motion of the composition-vector $\mathbf{d}$ is shown for a sequence of times $t^1 < t^2 < t^3 < t^4 < t^5$. The vector $\mathbf{d}(t^i)$, $i = 1, \ldots, 5$ is represented by a green circle if the number of inactive constraints at this point is equal to 1, by a blue circle if this number is equal to 2, and by a red circle if it is equal to 3. The first point $\mathbf{d}(t^1)$ is in area 1. Consequently $\mathbf{d}(t^1)$ is a single-phase point collinear to $\mathbf{z}_1(t^1)$ and is represented by a green circle. For the points $\mathbf{d}(t^2)$ the phase simplex is given by

$$\Sigma(\mathbf{d}(t^2)) \;=\; \mathrm{conv}(\mathbf{z}_1(t^2), \mathbf{z}_2(t^2)).$$

In other words the inequality constraints $\alpha = 1$ and $\alpha = 2$ are inactive, and the minimum of energy for the particle at $\mathbf{d}(t^2)$ follows the segment defined by $[f(\mathbf{z}_1(t^2))\ f(\mathbf{z}_2(t^2))]$ and is given by $y_1(t^2)f(\mathbf{z}_1(t^2)) + y_2(t^2)f(\mathbf{z}_2(t^2))$. The vector $\mathbf{d}(t^3)$ is also situated in the area 2 and its phase simplex is given by

$$\Sigma(\mathbf{d}(t^3)) = \mathrm{conv}(\mathbf{z}_1(t^3), \mathbf{z}_2(t^3)),$$

which is a 1-simplex again. Unlike the case $r = 1$ the single-phase points $\mathbf{z}_1$ and $\mathbf{z}_2$ moves when $\mathbf{d}$ evolves in area 2. The trajectory of $\mathbf{z}_1$ and $\mathbf{z}_2$ follows the frontier between the area 2 and the area 1.

Once $\mathbf{d}$ is in area 3 the phase simplex is given by $\mathrm{conv}(\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3)$. The points $\mathbf{z}_1, \mathbf{z}_2$ and $\mathbf{z}_3$ remain fixed as long as $\mathbf{d}$ evolves in area 3 because they define the unique hyperplane that is in contact with f at 3 points. Only the barycentric coordinates y_1, y_2 and y_3 evolves such that the relation $\mathbf{d} = \sum_{\alpha=1}^{3} y_\alpha \mathbf{z}_\alpha$ holds as $\mathbf{d}$ moves.

The last considered value of $\mathbf{d}$ is in the area 2 situated on the left side of the phase diagram Δ_2. The phase simplex of $\mathbf{d}(t^5)$ is a segment but unlike the points $\mathbf{d}(t^2)$ and $\mathbf{d}(t^3)$ the segment is generated by the single-phase $\mathbf{z}_2(t^5)$ and $\mathbf{z}_3(t^5)$ and one has

$$\Sigma(\mathbf{d}(t^5)) = \mathrm{conv}(\mathbf{z}_2(t^5), \mathbf{z}_3(t^5)).$$

Hence the representation of the time evolution of $\mathbf{d}$ on the phase diagram allows to know the evolution of the phase equilibrium defined at each $\mathbf{d}$.

The size of the convex regions of the energy function g on the phase diagrams covers many orders of magnitude (see [5]), with normalized values ranging from 10^0 to 10^{-16}. The phase diagram presented in Figure 1.5 is a real example from atmospheric chemistry. For instance the point situated at the frontier between the areas 1, 2 and 3 on the bottom left part of Δ_2 (the point $\mathbf{z}_3(t^4)$ in Figure 1.4) has the coordinates

$$\mathbf{z}_3^T = (1.05 \cdot 10^{-3}, \quad 1.17 \cdot 10^{-15}).$$

Moreover a zoomed-in view of the bottom right corner of Δ_2 shows how the number of inactive constraints may rapidly vary in a very small region. Hence the method of resolution for finding the phase simplex points (or contact points between the graph and the supporting tangent plane) has to handle several scales.

The energy functions f presented in this section have a number of convex areas that is equal to s. This representation stands as the general situation. Let us qualify the associated phase diagram *classical*. In reality the number of convex areas takes a value comprised between 1 and s. In Figure 1.6 two such examples are depicted. For both examples the energy function has 2 convex areas. Hence no area with 3 inactive inequality constraints can exist. On the left example 2, different areas are present: a large one with one inactive constraint and a small one with 2 inactive constraints. In this example the vectors $\mathbf{z}_1$, $\mathbf{z}_2$ and $\mathbf{z}_3$ share the same area $\Delta_{2,1} = \Delta_{2,2} = \Delta_{2,3}$, but are still associated to a corner of Δ_2. The example on the right is the opposite case with a large area with 2 inactive constraints and 2 small areas with one inactive constraint. The area $\Delta_{2,2}$ is equal to $\Delta_{2,3}$. Let us call such phase diagrams *not classical*.

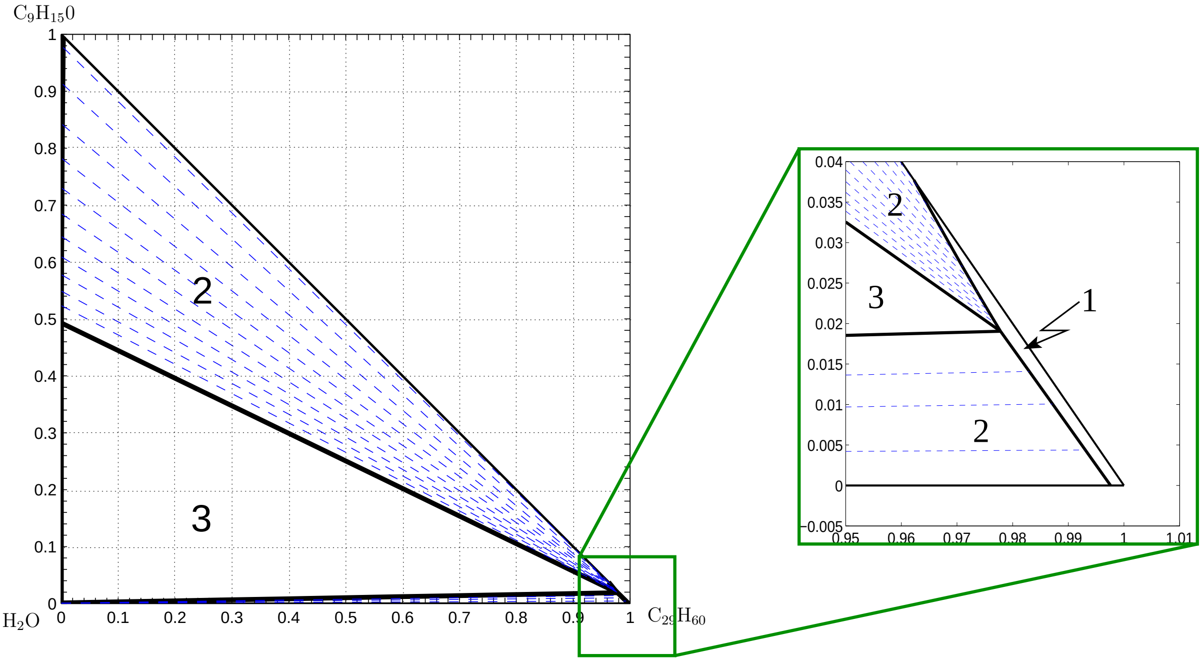


Figure 1.5: Phase diagram of the ternary system pinonic acid ($C_{10}H_{16}O_3$)/nonacosane ($C_{29}H_{60}$)/water (H_2O) at temperature $298K$ and pressure $1atm$.

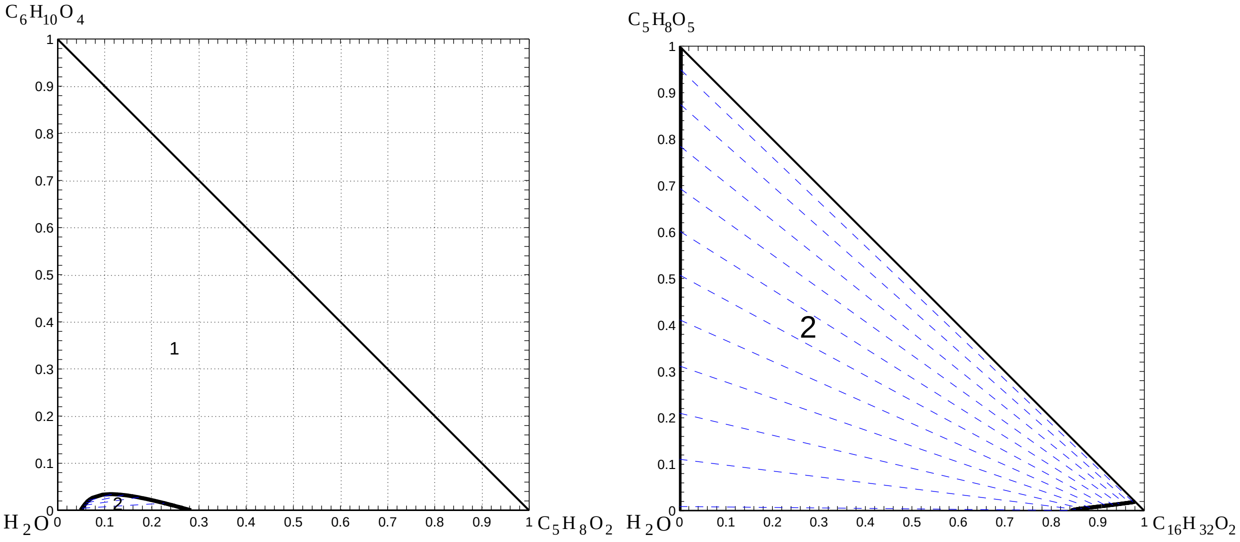


Figure 1.6: 2 phases diagrams of ternary system at temperature $298K$ and pressure $1atm$. On the left: adipic acid ($C_6H_{10}O_4$)/glutaraldehyde ($C_5H_8O_2$)/water. On the right: 2-hydroxy-glutaric acid ($C_5H_8O_5$)/palmitic acid ($C_{16}H_{32}O_2$)/water.

Remark 1.7.1. *Even if we work with* g *and the variables* $\mathbf{x}_\alpha$ *and* $\mathbf{b}$*, it is more convenient to represent to projections* f*,* $\mathbf{z}_\alpha$ *and* $\mathbf{d}$*. For that reason the figures in the remainder of this thesis always illustrate* f *and the projected variables* $\mathbf{z}_\alpha$ *and* $\mathbf{d}$*, but the notations* g*,* $\mathbf{x}_\alpha$ *and* $\mathbf{b}$ *are kept in the text and on the forthcoming figures.*

Chapter 2

Solving the phase equilibrium problem

In the previous chapter the modeling of the gas-aerosol system and its geometrical interpretation have been presented. Before starting the numerical resolution of (1.6.4), and since the PEP is an integral part of both resolution methods, let us describe further the characterizations of this optimization problem.

Amundson et al. in [4, 5] have studied the PEP for organic particles when the composition-vector **b** is a fixed input. In their works the PEP is described and its mathematical characterizations are presented. Moreover Amundson et al. propose an efficient technique to solve the PEP, based on a primal-dual interior-point method. Since this thesis is the continuity of their works, this chapter is dedicated to introduce their results. Some characteristics of the PEP and some insights on constrained optimization theory have already been given in the previous chapter. Here let us summarize the mathematical characteristics of the local minima of the PEP and present the primal-dual interior-point method of Amundson et al. [4, 5].

2.1 Mathematical characterizations of the local minima

Let us consider the reduced minimization problem (1.7.1), rewritten below

$$\begin{aligned} \min_{\{y_\alpha, \mathbf{z}_\alpha\}_{\alpha=1}^p} \quad & \sum_{\alpha=1}^p y_\alpha f(\mathbf{z}_\alpha), \\ \text{s.t.} \quad & \sum_{\alpha=1}^p y_\alpha \mathbf{z}_\alpha = \mathbf{d}, \quad \sum_{\alpha=1}^p y_\alpha = 1, \\ & y_\alpha \geq 0, \quad \alpha = 1, \dots, p. \end{aligned} \tag{2.1.1}$$

In the Section 1.7 the convex envelope of f was studied to get information about the global minimizer of (2.1.1). In this section, results that characterize the local minima of the problem and distinguish them from the global minimum, are presented.

In order to study the local minima, the Kuhn-Tucker theory introduced in Section 1.4.3 is used. As for the PEP (1.4.2), the LICQ holds for (2.1.1) at any feasible point. The proof of this result follows the proof of the Lemma 1.4.2. Furthermore the objective function and the constraints of (2.1.1) are continuously differentiable. Then the Theorem 1.4.1 that gives the KKT conditions, can be applied.

Theorem 2.1.1. *Let* $\mathbf{y}^T = (\mathbf{z}_1^T, \ldots, \mathbf{z}_p^T, y_1, \ldots, y_p)$ *be a local minimizer of (2.1.1) with* $\mathbf{d} \in \Delta_r$*. Then there exist unique Lagrange multipliers* $\boldsymbol{\eta} \in \mathbb{R}^r$*,* $\gamma \in \mathbb{R}$ *and* $\theta_\alpha \in \mathbb{R}$*,* $\alpha = 1, \ldots, p$*, such that*

$$
\begin{aligned}
y_\alpha(\boldsymbol{\nabla} f(\mathbf{z}_\alpha) + \boldsymbol{\eta}) &= \mathbf{0}, \quad \forall \alpha = 1, \ldots, p, && \text{(2.1.2a)}\\
f(\mathbf{z}_\alpha) + \boldsymbol{\eta}^T \mathbf{z}_\alpha + \gamma - \theta_\alpha &= 0, \quad \forall \alpha = 1, \ldots, p, && \text{(2.1.2b)}\\
\mathbf{d} - \sum_{\alpha=1}^{p} y_\alpha \mathbf{z}_\alpha &= \mathbf{0}, && \text{(2.1.2c)}\\
\sum_{\alpha=1}^{p} y_\alpha - 1 &= 0, && \text{(2.1.2d)}\\
y_\alpha &\geq 0, \quad \forall \alpha = 1, \ldots, p, && \text{(2.1.2e)}\\
\theta_\alpha &\geq 0, \quad \forall \alpha = 1, \ldots, p, && \text{(2.1.2f)}\\
\theta_\alpha y_\alpha &= 0, \quad \forall \alpha = 1, \ldots, p. && \text{(2.1.2g)}
\end{aligned}
$$

The Lagrange multiplier $\boldsymbol{\eta}$ is associated to the first equality constraint whereas γ is associated to the equality constraint $\sum_{\alpha=1}^{p} y_\alpha - 1 = 0$. The Lagrange multipliers θ_α are relative to the inequality constraints $y_\alpha \geq 0$, $\alpha = 1, \ldots, p$. In optimization theory, Lagrange multipliers are also called *dual* variables, whereas y_α and $\mathbf{x}_\alpha$, $\alpha = 1, \ldots, p$ are referred to as *primal* variables.

A solution of the KKT system (2.1.2) is called a KKT point of (2.1.1) and is generally non unique. The hyperplane associated with the KKT point $\mathbf{y}$ is defined by

$$
\mathcal{H}_{\mathbf{y}} = \{(\mathbf{z}^T, x_s) \in \mathbb{R}^s \mid \boldsymbol{\nabla} f(\mathbf{z}_\alpha)^T (\mathbf{z} - \mathbf{z}_\alpha) - x_s + f(\mathbf{z}_\alpha) = 0\},
$$

where $\mathbf{x}^T = (\mathbf{z}^T, x_s)$ defines the points lying in $\mathbb{R}^s$, and $\mathbf{z}_\alpha$ is any point of $\mathbf{y}$ such that $y_\alpha > 0$.

As for the optimization problem (1.4.2), let us denote by $\mathcal{I}$, resp. $\mathcal{A}$, the set of indices of inactive inequality constraints, resp. active. Then the index $\alpha \in \mathcal{I}$ if and only if $y_\alpha > 0$, and one has from (2.1.2a) and (2.1.2g) for all $\alpha \in \mathcal{I}$:

$$
\begin{aligned}
\boldsymbol{\nabla} f(\mathbf{z}_\alpha) &= -\boldsymbol{\eta}\\
\theta_\alpha &= 0.
\end{aligned}
$$

Hence (2.1.2b) becomes

$$
f(\mathbf{z}_\alpha) + \boldsymbol{\eta}^T \mathbf{z}_\alpha = -\gamma, \quad \forall \alpha \in \mathcal{I},
$$

and the equality in the definition of the hyperplane $\mathcal{H}_{\mathbf{y}}$ reads

$$-\boldsymbol{\eta}^T\mathbf{z} - x_s - \gamma \,=\, 0.$$

Equivalently if the hyperplane is characterized through the function h defined by $h(\mathbf{z}) = -\boldsymbol{\eta}^T\mathbf{z} - \gamma$, its definition reads

$$\mathcal{H}_{\mathbf{y}} \,=\, \{(\mathbf{z}^T, x_s) \in \mathbb{R}^s \,|\, h(\mathbf{z}) \,=\, x_s\}. \tag{2.1.3}$$

In this definition $\mathbf{y}$ is implicitly present through its dual variables η and γ. Moreover from this definition a point $(\mathbf{z}^T, x_s) \in \mathbb{R}^s$ is situated below the hyperplane defined at $\mathbf{y}$ if $h(\mathbf{z}) < x_s$ and above if $h(\mathbf{z}) > x_s$.

Remark 2.1.1. *Similarly to Section 1.7, a KKT point* $\mathbf{y}^T = (\mathbf{z}_1^T, \dots, \mathbf{z}_p^T, y_1, \dots, y_p)$ *of (2.1.1) can be quite improper. Some of the* $\mathbf{z}_\alpha$ *can not be present in the local minimizer, namely those for which* $y_\alpha = 0$*, and the* $\mathbf{z}_\alpha$ *need not be distinct. It is easy to remedy this by eliminating all the indices* α *such that* $y_\alpha = 0$*, and adding all the* $y_\alpha > 0$ *corresponding to the same* $\mathbf{z}_\alpha$*. Therefore, one needs only to consider a KKT point* $(\mathbf{z}_1^{\mathcal{I},T}, \dots, \mathbf{z}_{p^{\mathcal{I}}}^{\mathcal{I},T}, y_1^{\mathcal{I}}, \dots, y_{p^{\mathcal{I}}}^{\mathcal{I}})$ *of (2.1.1) that is* stable *in the sense that all the* $\mathbf{z}_\alpha$ *are present in the local minimizer and distinct.*

In the sequel let us denote the stable KKT point $(\mathbf{z}_1^{\mathcal{I},T}, \dots, \mathbf{z}_{p^{\mathcal{I}}}^{\mathcal{I},T}, y_1^{\mathcal{I}}, \dots, y_{p^{\mathcal{I}}}^{\mathcal{I}})$ in the shortened form $(y_\alpha, \mathbf{z}_\alpha)_{\alpha\in\mathcal{I}}$. Let us now establish the distinction between a global and the local minima of (2.1.1). From the multijet theory, as applied in [89], the assumption that f belongs to the residual set R implies the following corollary:

Corollary 2.1.2. *Let* $\mathbf{y} = (y_\alpha, \mathbf{z}_\alpha)_{\alpha\in\{1,\dots,p\}}$ *be a stable KKT point of (1.7.1) with* $\mathbf{d} \in \mathrm{int}\Delta_r$ *such that* $y_\alpha > 0$ *and* $\mathbf{z}_\alpha$ *are distinct. Then the set* $\Sigma = \mathrm{conv}(\mathbf{z}_1, \dots, \mathbf{z}_p)$ *is a* $(p-1)$*-simplex with* $p \le s$*.*

From Theorem 1.7.3, one deduces that for a global minimizer of (2.1.1), denoted by $\mathbf{y}^\dagger = (y_\alpha^\dagger, \mathbf{z}_\alpha^\dagger)_{\alpha\in\mathcal{I}^\dagger}$, the set $\Sigma^\dagger = \mathrm{conv}(\mathbf{z}_1^\dagger, \dots, \mathbf{z}_{p^{\mathcal{I}^\dagger}}^\dagger)$ is the phase simplex of $\mathbf{d}$. From Corollaries 1.7.5 and 2.1.2, the $\mathbf{z}_\alpha$, $\alpha \in \mathcal{I}$, of a KKT point of problem (2.1.1) form a tangent simplex and the single-phase point criterion can be applied to determine if this KKT point is a global minimum of (2.1.1). The aim of the following theorem is to make this statement precise.

Theorem 2.1.3. *Consider* $\mathbf{y}^\dagger = (y_\alpha^\dagger, \mathbf{z}_\alpha^\dagger)_{\alpha\in\mathcal{I}}$ *a feasible point of (2.1.1). The point* $\mathbf{y}^\dagger$ *is a global minimum of (2.1.1) if and only if* $\mathbf{y}^\dagger$ *is a KKT point of (2.1.1) and* $\mathbf{z}_\alpha^\dagger$ *are single-phase points for all* $\alpha \in \mathcal{I}$*.*

Therefore a global criterion to determine whether a KKT point is a global minimum is established. From Theorem 1.7.3, this criterion is equivalent to require that the hyperplane associated with the KKT point lies below the graph of f on Δ_r.

Theorem 2.1.4. *Consider* $\mathbf{y}^\dagger = (y_\alpha^\dagger, \mathbf{z}_\alpha^\dagger)_{\alpha\in\mathcal{I}}$ *a feasible point of (2.1.1). The point* $\mathbf{y}^\dagger$ *is a global minimum of (2.1.1) if and only if* $\mathbf{y}^\dagger$ *is a KKT point of (2.1.1) and*

$$f(\mathbf{z}) \geq h^\dagger(\mathbf{z}), \quad \forall \mathbf{z} \in \Delta_r, \tag{2.1.4}$$

where $h^\dagger$ *is the function associated to the hyperplane defined at the KKT point* $\mathbf{y}^\dagger$.

The characteristics of local and global minima of (2.1.1) are thus established. Now let us go back to the original PEP and study what are the implications of the Theorems 2.1.1-2.1.4 on (1.4.2). The point $(y_\alpha, \mathbf{x}_\alpha)_{\alpha\in\mathcal{I}}$ is a local minimizer of (1.4.2) if and only if $(y_\alpha, \mathbf{z}_\alpha = P\mathbf{x}_\alpha)_{\alpha\in\mathcal{I}}$ is a local minimizer of (2.1.1). Moreover $\mathbf{d}$ is in $\mathrm{int}\Delta_r$ if and only if $\mathbf{b}$ is in $\mathrm{rint}\Delta'_s$, the relative interior of Δ'_s [12]. Then if $(y_\alpha, \mathbf{x}_\alpha)_{\alpha\in\mathcal{I}}$ is a local minimizer of (1.4.2) and $\mathbf{b}$ is in $\mathrm{rint}\Delta'_s$, $y_\alpha > 0$ implies that $\mathbf{x}_\alpha$ belongs to $\mathrm{rint}\Delta'_s$, so that g is differentiable at $\mathbf{x}_\alpha$. Since the LICQ holds at $(y_\alpha, \mathbf{x}_\alpha)_{\alpha=1,\ldots,p}$, one has then

$$y_\alpha \left(\boldsymbol{\nabla} g(\mathbf{x}_\alpha) + \boldsymbol{\lambda}\right) + \zeta_\alpha \mathbf{e} = \mathbf{0}, \qquad \alpha = 1,\ldots,p, \tag{2.1.5a}$$

$$g(\mathbf{x}_\alpha) + \boldsymbol{\lambda}^T \mathbf{x}_\alpha - \theta_\alpha = 0, \qquad \alpha = 1,\ldots,p, \tag{2.1.5b}$$

$$\mathbf{b} - \sum_{\alpha=1}^{p} y_\alpha \mathbf{x}_\alpha = \mathbf{0}, \tag{2.1.5c}$$

$$1 - \mathbf{e}^T\mathbf{x}_\alpha = 0, \quad \mathbf{x}_\alpha > 0, \qquad \alpha = 1,\ldots,p, \tag{2.1.5d}$$

$$\theta_\alpha y_\alpha = 0, \quad \theta_\alpha \geq 0, \quad y_\alpha \geq 0, \qquad \alpha = 1,\ldots,p, \tag{2.1.5e}$$

where ζ_α is the Lagrange multiplier associated to the equality constraints $\mathbf{e}^T\mathbf{x}_\alpha - 1 = 0$ and $\boldsymbol{\lambda}$ is related to the multipliers $\boldsymbol{\eta}$ and γ in (2.1.2) via

$$\boldsymbol{\eta} = \mathbf{Z}_{\mathbf{e}}^T \boldsymbol{\lambda}, \quad \gamma = \lambda_s, \tag{2.1.6}$$

with λ_s the s^{th} component of $\boldsymbol{\lambda}$.

Note that the tangent plane criterion (2.1.4) stated in Theorem 2.1.4 is a global condition. There is no rigorous approach to determine whether the tangent plane arising from a KKT point lies below the molar Gibbs free energy surface for all feasible compositions $\mathbf{z}$ in Δ_r. Therefore, one has to rely on local criteria, even if this increases the odds of finding local minima. One such local criteria is related to a local phase stability test, which states that, if a postulated KKT point $\mathbf{y}^\dagger = (y_\alpha^\dagger, \mathbf{z}_\alpha^\dagger)_{\alpha\in\mathcal{I}}$ is thermodynamically stable with respect to perturbations in any or all of the phases, then

$$f(\mathbf{z}) - h^\dagger(\mathbf{z}) \approx (\mathbf{z} - \mathbf{z}_\alpha^\dagger)^T \boldsymbol{\nabla}^2 f(\mathbf{z}_\alpha^\dagger)(\mathbf{z} - \mathbf{z}_\alpha^\dagger) \geq 0, \quad \forall \mathbf{z} \in \mathcal{B}_\epsilon(\mathbf{z}_\alpha^\dagger), \tag{2.1.7}$$

where $\mathcal{B}_\epsilon(\mathbf{z}_\alpha^\dagger)$ is a neighborhood of $\mathbf{z}_\alpha^\dagger$ in Δ_r. Relationship (2.1.7) is equivalent to $\boldsymbol{\nabla}^2 f(\mathbf{z}_\alpha^\dagger) \geq 0$. From now on, let us assume that the Hessian matrix of f, $\boldsymbol{\nabla}^2 f$, is positive definite at the phases $\mathbf{z}_\alpha^\dagger$ of $\mathbf{y}^\dagger$, i.e.,

$$\boldsymbol{\nabla}^2 f(\mathbf{z}_\alpha^\dagger) \geq 0. \tag{2.1.8}$$

Relation (2.1.8) is also called the *meta-stability conditions* for problem (2.1.1).

The above mathematical characterizations are, however, not directly applicable for computation because finding a solution that satisfies the KKT system (2.1.2), or equivalently (2.1.5), is a difficult problem. The difficulty is mainly caused by the combinatorial aspect of the KKT system (2.1.5), or more precisely by the complementary conditions: $\theta_\alpha \geq 0$, $y_\alpha \geq 0$ and $\theta_\alpha y_\alpha = 0$. Indeed one could attempt to guess the *optimal active set* $\mathcal{A}$. Based on this guess, one could transform (2.1.5) into a system of nonlinear equations, which is much more computationally tractable. Unfortunately, the set of all possible active sets grows exponentially with p, the number of phases considered.

Moreover, not all the solutions of the KKT system (2.1.5) are solutions of the optimization problem in (1.4.2); some of them could be, for example, maximizers, saddle points, or unstable local minimizers. Therefore this type of approach can only be practical if initiated by a correct guess of the inactive and active sets, $\mathcal{I}$ and $\mathcal{A}$. This question is addressed in the next section.

2.2 Solution of the phase equilibrium problem

2.2.1 Active set method

The accurate identification of active constraints is important [34, 101]. Such an identification, by removing the difficult combinatorial aspect of the optimization problem in (1.4.2), reduces the mixed constrained minimization problem to an equality constrained problem which is much easier to deal with, and allows a faster convergence to the phase equilibrium. An active set identification procedure is presented here that correctly detects active constraints in a neighborhood of a KKT point. In order to identify accurately the active constraints, one needs to have a pair of primal and dual variables, *e.g.*, $(y_1, \dots, y_p, \mathbf{x}_1, \dots, \mathbf{x}_p; \boldsymbol{\lambda}, \zeta_1, \dots, \zeta_p, \theta_1, \dots, \theta_p)$, that is close to a KKT point. The primal-dual interior-point algorithm presented later will produce such a sequence of primal and dual variables. In the sequel, the pair $(y_1, \dots, y_p, \mathbf{x}_1, \dots, \mathbf{x}_p; \boldsymbol{\lambda}, \zeta_1, \dots, \zeta_p, \theta_1, \dots, \theta_p)$ is shortened in $(y_\alpha, \mathbf{x}_\alpha; \boldsymbol{\lambda}, \zeta_\alpha, \theta_\alpha)$.

The inequality constraints defined in the PEP are only about the variables y_α, $\alpha = 1, \dots, p$. Hence the detection of the active inequality constraints in the neighborhood of a KKT point simply consists in observing the value of y_α, $\alpha \in \mathcal{I}$. If this value is lower than a prescribed threshold, then the index of the corresponding constraint has to be removed from the set $\mathcal{I}$. Formally if the pair $(y_\alpha, \mathbf{x}_\alpha; \boldsymbol{\lambda}, \zeta_\alpha, \theta_\alpha)$ is sufficiently close to a KKT point, then the set $\mathcal{I}$ of inactive constraints actually present at the equilibrium can be obtained by removing the activating constraints from the system [34, 75, 101]

$$\mathcal{I} = \{1, \dots, p\} \backslash \{\alpha \,|\, 0 < y_\alpha < \epsilon_y\} \tag{2.2.1}$$

where ϵ_y is a given threshold. In numerical experiments ϵ_y is set to 10^{-8} when $\mathbf{b}$ is normalized (i.e. $\mathbf{e}^T\mathbf{b} = 1$). The exact solution of (1.4.2) can be computed based on the inactive constraints from the following reduced KKT system of equations:

$$
\begin{array}{rcll}
y_\alpha\left(\boldsymbol{\nabla} g(\mathbf{x}_\alpha)+\boldsymbol{\lambda}\right)+\zeta_\alpha\mathbf{e} & = & \mathbf{0}, & \alpha\in\mathcal{I},\\
g(\mathbf{x}_\alpha)+\boldsymbol{\lambda}^T\mathbf{x}_\alpha & = & 0, & \alpha\in\mathcal{I},\\
\displaystyle\sum_{\alpha\in\mathcal{I}} y_\alpha\mathbf{x}_\alpha & = & \mathbf{b}, & \\
\mathbf{e}^T\mathbf{x}_\alpha=1,\quad \mathbf{x}_\alpha>0,\quad y_\alpha & > & 0, & \alpha\in\mathcal{I}.
\end{array}
\tag{2.2.2}
$$

This identification procedure permits to couple the interior-point method presented in the next subsection with an active set method for the activation/deactivation of the inequality constraints. The procedure is the following: the inactive set $\mathcal{I}$ is initialized to $\{1,\dots,p\}$. Then an iterative sequence is built, that alternates between solving (2.2.2) and updating the inactive set.

Remark 2.2.1. *The system (2.2.2) only includes the indices $\alpha\in\mathcal{I}$. Consequently the variables $y_\alpha,\mathbf{x}_\alpha,\theta_\alpha$ and ζ_α, $\alpha\in\mathcal{A}$ are not updated and their values at the previous iteration are used to define the new iterate $(y_\alpha^+,\mathbf{x}_\alpha^+;\boldsymbol{\lambda}^+,\zeta_\alpha^+,\theta_\alpha^+)$.*

Let $\mathscr{P}^{\mathcal{I}}$ denote the index set of constraints $\alpha\in\mathcal{I}$ that satisfy $0<y_\alpha^+<\epsilon_y$. The set $\mathscr{P}^{\mathcal{I}}$ is the set of constraints that have to be removed from the set of inactive constraints at the next iteration.

At the next iteration, it is also possible that constraints $\alpha\in\mathcal{A}$ have to be added to the inactive set $\mathcal{I}$. The condition to deactivate a constraint $\alpha\in\mathcal{A}$ follows the definition of the dual variable θ_α that stipulates $\theta_\alpha^+\geq 0$. Following the remark 2.2.1, θ_α^+ is not updated. However an update can be made by using the relation (2.1.5b) and one gets

$$\theta_\alpha^+=g(\mathbf{x}_\alpha^+)+(\boldsymbol{\lambda}^+)^T\mathbf{x}_\alpha^+. \tag{2.2.3}$$

Note that in this expression the variable $\boldsymbol{\lambda}^+$ is solution of (2.2.2) whereas $\mathbf{x}_\alpha^+$ is the same as at the previous iteration since $\alpha\in\mathcal{A}$. So if the dual variable is such that $\theta_\alpha^+<0$, then the index α have to be added to $\mathcal{I}$. Let $\mathscr{P}^{\mathcal{A}}$ denote the index set of constraints $\alpha\in\mathcal{A}$, that satisfy $\theta_\alpha^+<0$, the new inactive set $\mathcal{I}^+$ is then given at the next iteration by

$$\mathcal{I}^+=\left(\mathcal{I}\cup\mathscr{P}^{\mathcal{A}}\right)\backslash\mathscr{P}^{\mathcal{I}}. \tag{2.2.4}$$

The KKT equations (2.2.2) are then updated and another iteration is carried out.

Remark 2.2.2. *The scalar product $\boldsymbol{\lambda}^T\mathbf{x}_\alpha$ may be rewritten as $\boldsymbol{\eta}^T\mathbf{z}_\alpha+\gamma$ thanks to the relation between $\boldsymbol{\lambda}$ and $(\boldsymbol{\eta},\gamma)$ and the relation $\mathbf{e}^T\mathbf{x}_\alpha=1$. Moreover by definition of the projection, we have $g(\mathbf{x}_\alpha)=f(\mathbf{z}_\alpha)$. Then the relation (2.2.3) is equal to*

$$\theta_\alpha^+ \;=\; f(\mathbf{z}_\alpha^+)+(\boldsymbol{\eta}^+)^T\mathbf{z}_\alpha^+ +\gamma^+ \;=\; f(\mathbf{z}_\alpha^+)-h(\mathbf{z}_\alpha^+).$$

Thus $\theta_\alpha^+>0$ indicates that the point $(\mathbf{z}_\alpha^{+,T},f(\mathbf{z}_\alpha^+))$ is situated above the tangent hyperplane defined at $(y_\alpha^+,\mathbf{x}_\alpha^+)$. The Gibbs tangent criterion is violated if $\theta_\alpha^+<0$ and the constraint α has to be added to $\mathcal{I}$ in order to converge to the phase equilibrium.

The primal-dual interior-point method and the details of its coupling with the active set identification procedure are addressed in the next subsection.

2.2.2 Primal-dual interior-point method

Interior-point methods have proved to be successful for nonlinear optimization and are currently the most powerful algorithms for large-scale nonlinear programming [75]. Interior-point method can be seen as barrier methods and one approach is to soften the non-negativity constraints $y_\alpha \geq 0$ by adding slack variables s_α, $\alpha = 1, \ldots, p$ and incorporating them into a logarithmic barrier term in the objective function. The optimization problem (1.4.2) is transformed into the following barrier problem:

$$\begin{array}{lll}\min\limits_{\{y_\alpha,\mathbf{x}_\alpha,s_\alpha\}_{\alpha=1}^p} & B_\nu(y_\alpha,\mathbf{x}_\alpha) = \sum\limits_{\alpha=1}^p y_\alpha g(\mathbf{x}_\alpha) - \nu \sum\limits_{\alpha=1}^p \ln s_\alpha, & \\ \text{s. t.} & \sum\limits_{\alpha=1}^p y_\alpha \mathbf{x}_\alpha = \mathbf{b}, & \\ & \mathbf{e}^T\mathbf{x}_\alpha = 1, \quad \mathbf{x}_\alpha > 0, \quad \alpha = 1,\ldots,p, & \\ & y_\alpha - s_\alpha = 0, \quad s_\alpha > 0, \quad \alpha = 1,\ldots,p, & \end{array} \tag{2.2.5}$$

where ν is a positive parameter.

Problem (2.2.5) is not equivalent to (1.4.2), but contains only equality constraints, and is much simpler to solve than (1.4.2). In the primal-dual interior-point algorithm of Amundson et al., (2.2.5) is approximately solved by applying one Newton iteration to its KKT system of equations, then decreasing the parameter ν, and repeating the process. This leads to a sequence of iterates that converges to a solution of (1.4.2) as $\nu \to 0$ under certain assumptions, as mentioned in the next lemma, which is an application of Theorem 8 in [36] with box constraints $y_\alpha \geq 0$.

Lemma 2.2.1. *Since the objective function and constraints of the problem (1.4.2) are continuous, the solution to the penalized problem (2.2.5) converges to the solution to the initial problem (1.4.2), when the penalty parameter ν tends to zero.*

This convergent sequence is used for a finite termination of the algorithm by applying the active phase identification procedure outlined in the previous section. Once the active constraints are identified and removed from the iterations, the exact solution of the PEP can be obtained by setting $\nu = 0$ and computing in a final step an equilibrium point only on the inactive constraints.

Let us now consider the problem of finding an approximate solution of problem (2.2.5) for a fixed value of the parameter ν. Denoting the Lagrange multipliers for $y_\alpha - s_\alpha = 0$ again by θ_α, $\alpha = 1, \ldots, p$, the KKT conditions for the barrier problem take the form:

$$\begin{array}{rcll} y_\alpha\left(\boldsymbol{\nabla} g(\mathbf{x}_\alpha) + \boldsymbol{\lambda}\right) + \zeta_\alpha \mathbf{e} & = & \mathbf{0}, & \alpha = 1,\ldots,p, \\ g(\mathbf{x}_\alpha) + \boldsymbol{\lambda}^T\mathbf{x}_\alpha - \theta_\alpha & = & 0, & \alpha = 1,\ldots,p, \\ \sum\limits_{\alpha=1}^p y_\alpha \mathbf{x}_\alpha & = & \mathbf{b}, & \\ \mathbf{e}^T\mathbf{x}_\alpha = 1, \quad \mathbf{x}_\alpha & > & 0, & \alpha = 1,\ldots,p, \end{array}$$

$$\begin{aligned} y_\alpha - s_\alpha = 0, \quad s_\alpha &> 0, \quad \alpha = 1, \ldots, p, \\ s_\alpha \theta_\alpha - \nu = 0, \quad \theta_\alpha &> 0, \quad \alpha = 1, \ldots, p, \end{aligned}$$

where, the last two sets of equations can be combined by eliminating the slacks s_α, yielding to the reduced system

$$\begin{aligned} y_\alpha \left(\nabla g(\mathbf{x}_\alpha) + \boldsymbol{\lambda} \right) + \zeta_\alpha \mathbf{e} &= \mathbf{0}, \quad \alpha = 1, \ldots, p, & \text{(2.2.6a)} \\ g(\mathbf{x}_\alpha) + \boldsymbol{\lambda}^T \mathbf{x}_\alpha - \theta_\alpha &= 0, \quad \alpha = 1, \ldots, p, & \text{(2.2.6b)} \\ \sum_{\alpha=1}^{p} y_\alpha \mathbf{x}_\alpha &= \mathbf{b}, & \text{(2.2.6c)} \\ \mathbf{e}^T \mathbf{x}_\alpha = 1, \quad \mathbf{x}_\alpha &> 0, \quad \alpha = 1, \ldots, p, & \text{(2.2.6d)} \\ y_\alpha \theta_\alpha - \nu = 0, \quad y_\alpha > 0, \quad \theta_\alpha &> 0, \quad \alpha = 1, \ldots, p. & \text{(2.2.6e)} \end{aligned}$$

Note that the above KKT system (2.2.6) contains only equations, and can be viewed as a perturbation of the original KKT system (2.1.5) where the complementary slackness conditions are approximated by a set of equations that is controlled by ν. Note also that the KKT system (2.2.6) produces a sequence of primal and dual variables $(y_\alpha, \mathbf{x}_\alpha; \boldsymbol{\lambda}, \zeta_\alpha, \theta_\alpha)$ that converges to a solution of (2.1.5) as $\nu \to 0$; the convergent sequence is used in the active phase identification procedure (2.2.1) for a finite termination of the algorithm.

Let us ignore (for the moment) the fact that y_α and θ_α must be positive. System (2.2.6) is solved with the Newton method. Let us denote by $p_{y_\alpha}, \mathbf{p}_{\mathbf{x}_\alpha}, \mathbf{p}_{\boldsymbol{\lambda}}, p_{\zeta_\alpha}$ and p_{θ_α} respectively the increments of the variables $y_\alpha, \mathbf{x}_\alpha, \boldsymbol{\lambda}, \zeta_\alpha$ and θ_α, $\alpha = 1, \ldots, p$. A Newton iteration gives the following system

$$\begin{aligned} y_\alpha \nabla^2 g(\mathbf{x}_\alpha) \mathbf{p}_{\mathbf{x}_\alpha} + (\nabla g(\mathbf{x}_\alpha) + \boldsymbol{\lambda}) p_{y_\alpha} & \\ + y_\alpha \mathbf{p}_{\boldsymbol{\lambda}} + p_{\zeta_\alpha} \mathbf{e} &= -y_\alpha \nabla g(\mathbf{x}_\alpha) - y_\alpha \boldsymbol{\lambda} - \zeta_\alpha \mathbf{e}, \\ & \qquad \alpha = 1, \ldots, p, \\ (\nabla g(\mathbf{x}_\alpha) + \boldsymbol{\lambda})^T \mathbf{p}_{\mathbf{x}_\alpha} + \mathbf{x}_\alpha^T \mathbf{p}_{\boldsymbol{\lambda}} - p_{\theta_\alpha} &= -g(\mathbf{x}_\alpha) - \mathbf{x}_\alpha^T \boldsymbol{\lambda} + \theta_\alpha, \\ & \qquad \alpha = 1, \ldots, p, \\ \sum_{\alpha=1}^{p} y_\alpha \mathbf{p}_{\mathbf{x}_\alpha} + \sum_{\alpha=1}^{p} p_{y_\alpha} \mathbf{x}_\alpha &= \mathbf{b} - \sum_{\alpha=1}^{p} y_\alpha \mathbf{x}_\alpha, \\ \mathbf{e}^T \mathbf{p}_{\mathbf{x}_\alpha} &= 1 - \mathbf{e}^T \mathbf{x}_\alpha, \quad \alpha = 1, \ldots, p, \\ \theta_\alpha y_\alpha^{-1} p_{y_\alpha} + p_{\theta_\alpha} &= \nu y_\alpha^{-1} - \theta_\alpha, \quad \alpha = 1, \ldots, p, \end{aligned}$$

which is further simplified by eliminating p_{θ_α} from the second set of equations via the relations from the last set of equations

$$p_{\theta_\alpha} = \nu y_\alpha^{-1} - \theta_\alpha - \theta_\alpha y_\alpha^{-1} p_{y_\alpha},$$

giving

$$
\begin{aligned}
y_\alpha \boldsymbol{\nabla}^2 g(\mathbf{x}_\alpha)\mathbf{p}_{\mathbf{x}_\alpha} + (\boldsymbol{\nabla} g(\mathbf{x}_\alpha) + \boldsymbol{\lambda}) p_{y_\alpha} & & \qquad (2.2.7)\\
+ y_\alpha \mathbf{p}_{\boldsymbol{\lambda}} + p_{\zeta_\alpha}\mathbf{e} &= -y_\alpha \boldsymbol{\nabla} g(\mathbf{x}_\alpha) - y_\alpha \boldsymbol{\lambda} - \zeta_\alpha \mathbf{e}, \\
& \qquad \alpha = 1, \ldots, p, \\
(\boldsymbol{\nabla} g(\mathbf{x}_\alpha) + \boldsymbol{\lambda})^T \mathbf{p}_{\mathbf{x}_\alpha} + \mathbf{x}_\alpha^T \mathbf{p}_{\boldsymbol{\lambda}} & \\
+ \theta_\alpha y_\alpha^{-1} p_{y_\alpha} &= -g(\mathbf{x}_\alpha) - \mathbf{x}_\alpha^T \boldsymbol{\lambda} + \nu y_\alpha^{-1}, \qquad (2.2.8)\\
& \qquad \alpha = 1, \ldots, p, \\
\sum_{\alpha=1}^{p} y_\alpha \mathbf{p}_{\mathbf{x}_\alpha} + \sum_{\alpha=1}^{p} \mathbf{x}_\alpha \, p_{y_\alpha} &= \mathbf{b} - \sum_{\alpha=1}^{p} y_\alpha \mathbf{x}_\alpha. \qquad (2.2.9)\\
\mathbf{e}^T \mathbf{p}_{\mathbf{x}_\alpha} &= 1 - \mathbf{e}^T \mathbf{x}_\alpha, \qquad (2.2.10)\\
& \qquad \alpha = 1, \ldots, p.
\end{aligned}
$$

This system is written on the reduced matrix form

$$
\begin{pmatrix}
y_\alpha \boldsymbol{\nabla}^2 g(\mathbf{x}_\alpha) & \boldsymbol{\nabla} g(\mathbf{x}_\alpha) + \boldsymbol{\lambda} & y_\alpha \mathbf{I} & \mathbf{e} \\
(\boldsymbol{\nabla} g(\mathbf{x}_\alpha) + \boldsymbol{\lambda})^T & \frac{\theta_\alpha}{y_\alpha}\mathbf{I} & \mathbf{x}_\alpha^T & 0 \\
y_\alpha \mathbf{I} & \mathbf{x}_\alpha & \mathbf{0} & \mathbf{0} \\
\mathbf{e}^T & 0 & \mathbf{0} & 0
\end{pmatrix}
\begin{pmatrix}
\mathbf{p}_{\mathbf{x}_\alpha} \\ p_{y_\alpha} \\ \mathbf{p}_{\boldsymbol{\lambda}} \\ p_{\zeta_\alpha}
\end{pmatrix}
=
\begin{pmatrix}
-y_\alpha \boldsymbol{\nabla} g(\mathbf{x}_\alpha) - y_\alpha \boldsymbol{\lambda} - \zeta_\alpha \mathbf{e} \\
-g(\mathbf{x}_\alpha) - \mathbf{x}_\alpha^T \boldsymbol{\lambda} + \nu y_\alpha^{-1} \\
\mathbf{b} - \sum_{\alpha=1}^{p} y_\alpha \mathbf{x}_\alpha \\
1 - \mathbf{e}^T \mathbf{x}_\alpha
\end{pmatrix}.
\qquad (2.2.11)
$$

One can observe that the matrix is symmetric. Since the pair $(0, \mathbf{x})$ is an eigen-pair for the matrix $\boldsymbol{\nabla}^2 g(\mathbf{x})$ (relation (1.4.7)), $\boldsymbol{\nabla}^2 g(\mathbf{x})$ is not invertible and techniques based on directly computing the Schur complement or its inverse cannot be applied to solve the linear system. Hence a technique of deflating $\boldsymbol{\nabla}^2 g(\mathbf{x}_\alpha)$ is applied to transform the linear system (2.2.7)-(2.2.10) so that the singularity no longer poses a difficulty. More precisely, the idea is to project the system (2.2.7)-(2.2.10) onto the null-space of $\mathbf{e}^T$ so that the corresponding reduced Hessian $\mathbf{Z}_{\mathbf{e}}^T \boldsymbol{\nabla}^2 g(\mathbf{x}_\alpha) \mathbf{Z}_{\mathbf{e}}$ is not singular. The reduced Hessian $\mathbf{Z}_{\mathbf{e}}^T \boldsymbol{\nabla}^2 g(\mathbf{x}_\alpha) \mathbf{Z}_{\mathbf{e}}$ is positive definite in a neighborhood of a stable equilibrium. Then, the reduced system with the positive definite Hessian allows us to apply a Schur complement method for its solution.

The null-space matrix of $\mathbf{e}^T$ is given by

$$
\mathbf{Z}_{\mathbf{e}} = \begin{pmatrix} \mathbf{I}_r \\ -\mathbf{e}^T \end{pmatrix}.
$$

Let us define

$$
\boldsymbol{\nabla}_{\mathbf{z}_\alpha} f(\mathbf{z}_\alpha) = \mathbf{Z}_{\mathbf{e}}^T \boldsymbol{\nabla} g(\mathbf{x}_\alpha), \quad \boldsymbol{\nabla}^2_{\mathbf{z}_\alpha} f(\mathbf{z}_\alpha) = \mathbf{Z}_{\mathbf{e}}^T \boldsymbol{\nabla}^2 g(\mathbf{x}_\alpha) \mathbf{Z}_{\mathbf{e}}, \quad \boldsymbol{\eta} = \mathbf{Z}_{\mathbf{e}}^T \boldsymbol{\lambda},
$$

where $\boldsymbol{\nabla}_{\mathbf{z}_\alpha} f$ and $\boldsymbol{\nabla}^2_{\mathbf{z}_\alpha} f$ are the reduced gradient and reduced Hessian of g respectively. Let us recall the relation between $\boldsymbol{\lambda}$ and the reduced variables $\boldsymbol{\eta}$ and γ

$$
\boldsymbol{\lambda} = \begin{pmatrix} \boldsymbol{\eta} + \gamma\, \mathbf{e} \\ \gamma \end{pmatrix} \iff \left(\boldsymbol{\eta} = \boldsymbol{\lambda}_{1:r} - \gamma\, \mathbf{e} \text{ and } \gamma = \lambda_s \right), \qquad (2.2.12)
$$

where $\boldsymbol{\lambda}_{1:r}$ stands for the subvector of $\boldsymbol{\lambda}$ made of the first r components.

The linear system (2.2.7)-(2.2.10) is then equivalent to:

$$\begin{pmatrix} y_\alpha \boldsymbol{\nabla}^2 f(\mathbf{z}_\alpha) & \boldsymbol{\nabla} f(\mathbf{z}_\alpha) + \boldsymbol{\eta} & y_\alpha \mathbf{I} & \mathbf{0} \\ (\boldsymbol{\nabla} f(\mathbf{z}_\alpha) + \boldsymbol{\eta})^T & \frac{\theta_\alpha}{y_\alpha}\mathbf{I} & \mathbf{z}_\alpha^T & \mathbf{e}^T\mathbf{x}_\alpha \\ y_\alpha \mathbf{I} & \mathbf{z}_\alpha & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{e}^T\mathbf{x}_\alpha & \mathbf{0} & 0 \end{pmatrix} \begin{pmatrix} \mathbf{p}_{\mathbf{z}_\alpha} \\ p_{y_\alpha} \\ \mathbf{p}_{\boldsymbol{\eta}} \\ p_\gamma \end{pmatrix} = \begin{pmatrix} \mathbf{b}_{\mathbf{z}_\alpha} \\ b_{y_\alpha} \\ \mathbf{b}_{\boldsymbol{\eta}} \\ b_\gamma \end{pmatrix} \tag{2.2.13}$$

where

$$\begin{aligned} \mathbf{b}_{\mathbf{z}_\alpha} &= -y_\alpha \boldsymbol{\nabla}_{\mathbf{z}_\alpha} f - y_\alpha \boldsymbol{\eta} - y_\alpha \mathbf{Z}_{\mathbf{e}}^T \boldsymbol{\nabla}^2 g(\mathbf{x}_\alpha)\mathbf{p}^0_{\mathbf{x}_\alpha} \\ &= -y_\alpha \boldsymbol{\nabla}_{\mathbf{z}_\alpha} f - y_\alpha \boldsymbol{\eta} - y_\alpha (1 - \mathbf{e}^T\mathbf{x}_\alpha)\left(\partial^2_{1:r,s} g(\mathbf{x}_\alpha) - \partial^2_{s,s} g(\mathbf{x}_\alpha)\,\mathbf{e}\right), \\ b_{y_\alpha} &= -g(\mathbf{x}_\alpha) - \mathbf{x}_\alpha^T \boldsymbol{\lambda} + \nu y_\alpha^{-1} - (\boldsymbol{\nabla} g(\mathbf{x}_\alpha) + \boldsymbol{\lambda})^T \mathbf{p}^0_{\mathbf{x}_\alpha} \\ &= -g(\mathbf{x}_\alpha) - \mathbf{x}_\alpha^T \boldsymbol{\lambda} + \nu y_\alpha^{-1} - (1 - \mathbf{e}^T\mathbf{x}_\alpha)\left(\partial_{n_s} g(\mathbf{x}_\alpha) + \gamma\right), \\ \mathbf{b}_{\boldsymbol{\eta}} &= \mathbf{d} - \textstyle\sum_{\alpha=1}^p y_\alpha \mathbf{z}_\alpha, \\ b_\gamma &= \mathbf{e}^T\mathbf{b} - \textstyle\sum_{\alpha=1}^p y_\alpha, \end{aligned}$$

and $\mathbf{p}^0_{\mathbf{x}_\alpha} = \begin{pmatrix} \mathbf{0} \\ 1 - \mathbf{e}^T\mathbf{x}_\alpha \end{pmatrix}$ which is a particular solution of (2.2.10). Details on the numerical resolution of (2.2.13) can be found in [4, 5]. Then the increments $\mathbf{p}_{\mathbf{x}_\alpha}$ and $\mathbf{p}_{\boldsymbol{\lambda}}$ are given by

$$\begin{aligned} \mathbf{p}_{\mathbf{x}_\alpha} &= \mathbf{p}^0_{\mathbf{x}_\alpha} + \mathbf{Z}_{\mathbf{e}} \mathbf{p}_{\mathbf{z}_\alpha}, \\ \mathbf{p}_{\boldsymbol{\lambda}} &= \begin{pmatrix} \mathbf{p}_{\boldsymbol{\eta}} + p_\gamma \mathbf{e} \\ p_\gamma \end{pmatrix}. \end{aligned}$$

The new estimate of the solution of the KKT system (2.2.6) is proceeded by

$$y_\alpha^+ = y_\alpha + \tau p_{y_\alpha}, \qquad \alpha = 1, \ldots, p, \tag{2.2.14a}$$

$$\mathbf{x}_\alpha^+ = \mathbf{x}_\alpha + \tau \mathbf{p}_{\mathbf{x}_\alpha}, \qquad \alpha = 1, \ldots, p, \tag{2.2.14b}$$

$$\boldsymbol{\lambda}^+ = \boldsymbol{\lambda} + \tau \mathbf{p}_{\boldsymbol{\lambda}}, \tag{2.2.14c}$$

$$\zeta_\alpha^+ = \zeta_\alpha + \tau p_{\zeta_\alpha}, \qquad \alpha = 1, \ldots, p, \tag{2.2.14d}$$

$$\theta_\alpha^+ = \theta_\alpha + \tau p_{\theta_\alpha}, \qquad \alpha = 1, \ldots, p. \tag{2.2.14e}$$

The step length τ is chosen to ensure that $y_\alpha^+ > 0$ and $\theta_\alpha^+ > 0$. In other words

$$\tau = \max\{0 < \sigma \leq 1 \,|\, y_\alpha + \sigma p_{y_\alpha} > 0,\ \theta_\alpha + \sigma p_{\theta_\alpha} > 0,\ \alpha = 1, \ldots, p\}.$$

The combination of the active set identification procedure with the Newton algorithm consists in working only with the indices $\alpha \in \mathcal{I}$ and updates the inactive set $\mathcal{I}$ at each step of the Newton method. This algorithm is summarized as follows

Algorithm 2.2.1 (Summary of active set/ interior-point/ Newton algorithm)**.** *Initialize* $y_1^0, \ldots, y_p^0$, $\mathbf{x}_1^0, \ldots, \mathbf{x}_p^0$, $\boldsymbol{\lambda}^0$, $\zeta_1^0, \ldots, \zeta_p^0$, $\theta_1^0, \ldots, \theta_p^0$, ν^0 *and* $\mathcal{I}^0$*. For* $j = 0, 1, 2, \ldots$*, execute the following steps until some stopping criterion is satisfied*

1. *Compute the Newton direction* $(p^j_{y_\alpha}, \mathbf{p}^j_{\mathbf{x}_\alpha}, p^j_{\zeta_\alpha}, \mathbf{p}^j_{\boldsymbol{\lambda}}, p^j_{\theta_\alpha})$ *of the Newton method associated with the inactive set* $\mathcal{I}^j$ *(i.e. with* $\alpha \in \mathcal{I}^j$*);*
2. *Compute the step length* τ *in (2.2.14) to ensure that* $y^{j+1}_\alpha > 0$ *and* $\theta^{j+1}_\alpha > 0$, $\alpha \in \mathcal{I}^j$*;*
3. *Update* $y^{j+1}_\alpha, \mathbf{x}^{j+1}_\alpha, \boldsymbol{\lambda}^{j+1}, \zeta^{j+1}_\alpha, \theta^{j+1}_\alpha$ *for* $\alpha \in \mathcal{I}^j$*;*
4. *Update the set of inactive constraints* $\mathcal{I}^{j+1}$ *with (2.2.4);*
5. *Compute the new parameter* ν^{j+1}*.*

Terminate with a resolution of the linear system with $\nu = 0$ *and the last inactive set* $\mathcal{I}^{j+1}$*.*

The stopping criterion at the j^{th} step is based on the increments of the linear system (2.2.13) and reads

$$\|(\mathbf{p}^{j,T}_{\mathbf{z}_\alpha}, p^j_{y_\alpha}, \mathbf{p}^{j,T}_{\boldsymbol{\eta}}, p^j_\gamma)\|_\infty \le tol_{ipm}, \text{ with } \alpha \in \mathcal{I}^j,$$

where tol_{ipm} is a given tolerance for the primal-dual interior-point method.

Strategy on the decrease of the parameter ν can be found in [5, 6, 29, 38] and is not discussed here. The forthcoming discussion is about the initialization of the procedure of the algorithm above-mentioned for the PEP.

2.2.3 Initialization procedure

Let $\mathbf{y}^\dagger = (y^\dagger_\alpha, \mathbf{z}^\dagger_\alpha)_{\alpha\in\mathcal{I}^\dagger}$ be the global minimizer of (2.1.1). In order to use a common terminology in the field of interior-point methods, we would like to obtain a central path $\{y^\nu_\alpha, \mathbf{z}^\nu_\alpha\}_{\alpha=1,\dots,p}$ generated by the interior-point method as described in the Algorithm 2.2.1 that converges to $\mathbf{y}^\dagger$. For the initialization of $\Sigma^0 := \mathrm{conv}(\mathbf{z}^0_1, \dots, \mathbf{z}^0_p)$, the points $\mathbf{z}^0_\alpha$ are initialized in the corners of the simplex Δ_r in order to cover all convex areas. Typically the initialization of the vector $\mathbf{z}^0_\alpha$ is given by

$$\begin{aligned} x^0_{\alpha,i} &= \begin{cases} \epsilon, & \text{if } i \neq \alpha, \\ 1 - r\epsilon, & \text{if } i = \alpha, \end{cases} \quad \text{for } i = 1, \dots, s; \\ \text{and } \mathbf{z}^0_\alpha &= P\mathbf{x}^0_\alpha, \end{aligned}$$

with $0 < \epsilon \ll 1$ given and $r = s - 1$.

Once the initial simplex Σ^0 is set, then y^0_α, $\alpha = 1, \dots, p$, are initialized as the barycentric coordinates of $\mathbf{d}$ in Σ^0: $\mathbf{d} = \sum_{\alpha=1}^p y^0_\alpha \mathbf{z}^0_\alpha$. If there exist $\alpha \in \{1, \dots, p\}$ such that $y^0_\alpha \le \varepsilon_y$, then the indices α are removed from $\mathcal{I}^0$, y^0_α is set to ε_y and the barycentric coordinates are recomputed such that $\mathbf{d} = \sum_{\alpha\in\mathcal{I}^0} y^0_\alpha \mathbf{z}^0_\alpha$.

Let us study the initialization of the dual variables $\boldsymbol{\lambda}^0$ and θ^0_α, $\alpha = 1, \dots, p$. Since the variables $\boldsymbol{\eta}$ and γ are employed in the resolution of the optimization instead of $\boldsymbol{\lambda}$, let us consider the initialization of $\boldsymbol{\eta}^0$ and γ^0. To that aim let us resume the first equation of the KKT conditions for the barrier problem (2.2.6)

$$y^0_\alpha \left(\boldsymbol{\nabla} g(\mathbf{x}^0_\alpha) + \boldsymbol{\lambda}^0\right) + \zeta^0_\alpha \mathbf{e} = \mathbf{0}, \quad \alpha = 1, \dots, p.$$

Then let us multiply this latter by $\mathbf{Z}_{\mathbf{e}}^T$. We obtain for all $\alpha = 1, \ldots, p$

$$\begin{aligned} \mathbf{0} &= y_\alpha^0 \left(\mathbf{Z}_{\mathbf{e}}^T \boldsymbol{\nabla} g(\mathbf{x}_\alpha^0) + \mathbf{Z}_{\mathbf{e}}^T \boldsymbol{\lambda}^0 \right) + \zeta_\alpha^0 \mathbf{Z}_{\mathbf{e}}^T \mathbf{e} \\ &= y_\alpha^0 \left(\boldsymbol{\nabla} f(\mathbf{z}_\alpha^0) + \boldsymbol{\eta}^0 \right). \end{aligned}$$

The variable $\boldsymbol{\eta}^0$ is set to minimize $\boldsymbol{\nabla} f(\mathbf{z}_\alpha^0) + \boldsymbol{\eta}^0$ in least squares sense with the weights y_α^0, $\alpha = 1, \ldots, p$. In other terms for given $\mathbf{z}_\alpha^0$ and y_α^0, $\alpha = 1, \ldots, p$, $\boldsymbol{\eta}^0$ is the minimizer of

$$\min_{\boldsymbol{\eta}} \sum_{\alpha=1}^{p} \frac{y_\alpha^0}{2} \|\boldsymbol{\nabla} f(\mathbf{z}_\alpha^0) + \boldsymbol{\eta}\|_2^2.$$

The minimum is obtain by finding the zero of the first order conditions, namely

$$\sum_{\alpha=1}^{p} y_\alpha^0 (\boldsymbol{\nabla} f(\mathbf{z}_\alpha^0) + \boldsymbol{\eta}) = \mathbf{0}.$$

The initialization for $\boldsymbol{\eta}^0$ is then established

$$\boldsymbol{\eta}^0 = \frac{1}{\sum_{\alpha=1}^{p} y_\alpha^0} \sum_{\alpha=1}^{p} y_\alpha^0 \boldsymbol{\nabla} f(\mathbf{z}_\alpha^0). \tag{2.2.15}$$

Concerning the initialization of γ let us resume this time the second KKT conditions of (2.2.6)

$$g(\mathbf{x}_\alpha) + \boldsymbol{\lambda}^T \mathbf{x}_\alpha - \theta_\alpha = 0, \quad \forall \alpha = 1, \ldots, p,$$

which is equivalent to

$$f(\mathbf{z}_\alpha) + \mathbf{z}_\alpha^T \boldsymbol{\eta} + \gamma - \theta_\alpha = 0, \quad \forall \alpha = 1, \ldots, p. \tag{2.2.16}$$

The initialization of γ^0 is as before set to minimize the above equation in least squares sense with the weights y_α^0, $\alpha = 1, \ldots, p$. Thus γ^0 is the root of

$$\sum_{\alpha=1}^{p} y_\alpha^0 \left(f(\mathbf{z}_\alpha^0) + \boldsymbol{\eta}^{0,T} \mathbf{z}_\alpha^0 + \gamma^0 - \theta_\alpha^0 \right) = 0,$$

for given y_α^0, $\mathbf{z}_\alpha^0$, $\alpha = 1, \ldots, p$, and $\boldsymbol{\eta}^0$.

One deduces the following expression for γ^0:

$$\gamma^0 = -\frac{\sum_{\alpha=1}^{p} y_\alpha^0 \left(f(\mathbf{z}_\alpha^0) + \boldsymbol{\eta}^{0,T} \mathbf{z}_\alpha^0 - \theta_\alpha^0 \right)}{\sum_{\alpha=1}^{p} y_\alpha^0}.$$

Using the complementarity relation $\nu = y_\alpha^0 \theta_\alpha^0$ for $\alpha = 1, \ldots, p$, the following relation is obtained

$$\sum_{\alpha=1}^{p} y_\alpha^0 \theta_\alpha^0 = p \, \nu^0$$

that allows to establish the initialization formula for γ^0

$$\gamma^0 = -\frac{\sum_{\alpha=1}^{p} y_\alpha^0 (f(\mathbf{z}_\alpha^0) + \boldsymbol{\eta}^{0,T}\mathbf{z}_\alpha^0) - p\,\nu^0}{\sum_{\alpha=1}^{p} y_\alpha^0}. \tag{2.2.17}$$

Since the dual variables ζ_α, $\alpha = 1, \ldots, p$, do not appear in the reduced linear system (2.2.13) there is no need to initialize them. Consequently the last dual variables to initialize is θ_α, $\alpha = 1, \ldots, p$. If the α^{th} inequality constraints is inactive, i.e. if $y_\alpha^0 > 0$, then the variable θ_α^0 is initialized by using the complementarity relation

$$\theta_\alpha^0 = \frac{\nu^0}{y_\alpha^0}, \quad \alpha \in \mathcal{I}^0.$$

When $\alpha \in \mathcal{A}^0$, the equation (2.2.16) gives immediately the relation

$$\theta_\alpha^0 = f(\mathbf{z}_\alpha^0) + \boldsymbol{\eta}^{0,T}\mathbf{z}_\alpha^0 + \gamma^0 \mathbf{e}^T \mathbf{x}_\alpha^0, \quad \alpha \in \mathcal{A}^0.$$

Finally the initial value of the barrier parameter ν^0 is empirically set to 0.001.

Chapter 3

An optimization-based numerical method

The purpose of this chapter is the numerical resolution of the system (1.6.4) in order to determine $\mathbf{b}(t)$, $R(t)$, $\mathbf{x}_\alpha(t)$ and $y_\alpha(t)$, $\alpha = 1, \ldots, p$ and $t \in [0, T)$. In addition to these variables, the system (1.6.4) contains implicitly the variable $\mathcal{I}$ and $\mathcal{A}$, the set of indices of the inactive, resp. active, inequality constraints. As the other variables, the sets $\mathcal{I}$ and $\mathcal{A}$ evolve in time. Consequently let us write $\mathcal{I}(t)$ and $\mathcal{A}(t)$ from now on.

Two different methods for the resolution of the system (1.6.4) are presented. The first method exploits the characteristics of the minimization problem to ensure the admissibility of the solution and advocates a time splitting algorithm that decouples differential and optimization operators. The second method views the system (1.6.4) as a differential algebraic system and uses implicit Runge-Kutta methods to solve it. Both methods follow the same strategy

- Solve the system (1.6.4) for a fixed number of inactive inequality constraints.

- At each time step of the resolution, check if an inequality constraint has to be activated or deactivated through detection criteria.

- If an activation/deactivation occurs, the resolution is stopped and the computation of the activation or deactivation time and points is started.

- Once this computation is achieved, the resolution of (1.6.4) with a different number of inactive inequality constraints restarts.

In this chapter the first numerical method is introduced. The second method is developed in the next chapter. Both presentations are structured as the above-mentioned strategy. Hence let us begin with the time splitting scheme for the resolution of (1.6.4) when the number of inactive constraints is fixed. The detection and computation of the activation and deactivation are considered in a second step. Some theoretical results are presented in a simplified case and numerical results conclude this chapter.

3.1 Numerical method for a fixed number of inactive constraints

In this first section, the sets $\mathcal{I}(t)$ and $\mathcal{A}(t)$ are supposed to be constant. The first method of resolution is inspired by the optimization techniques described in Chapter 2 and developed by Amundson et al. in [4, 5]. The main idea is to keep the optimization algorithm based on a primal-dual interior-point method and include it in a dynamic structure. For that reason a time splitting scheme is used [2].

3.1.1 A time splitting scheme

Let us recall the complete system to solve: find $\mathbf{b}, \mathbf{x}_\alpha : (0,T) \to \mathbb{R}^s_+$ and $R, y_\alpha : (0,T) \to \mathbb{R}_+$, $\alpha = 1,\dots,p$ satisfying

$$\begin{aligned}
\frac{d}{dt}\mathbf{b}(t) &= \mathbf{j}\left(\mathbf{b}(t), \mathbf{x}^{\mathcal{I}}_\alpha(t), R(t)\right), \qquad \mathbf{b}(0) = \mathbf{b}_0 \\
R(t) &= \left(\frac{3}{4\pi}\sum_{i=1}^{s}\frac{m_{c,i}b_i(t)}{\rho_i}\right)^{\frac{1}{3}}, \\
\{\mathbf{x}_\alpha(t), y_\alpha(t)\}_{\alpha=1}^p &= \underset{\{\bar{\mathbf{x}}_\alpha,\bar{y}_\alpha\}_{\alpha=1}^p}{\operatorname{argmin}} \sum_{\alpha=1}^{p}\bar{y}_\alpha\, g(\bar{\mathbf{x}}_\alpha) \\
&\qquad \text{s.t.} \sum_{\alpha=1}^{p}\bar{y}_\alpha\bar{\mathbf{x}}_\alpha = \mathbf{b}(t), \\
&\qquad\quad \mathbf{e}^T\bar{\mathbf{x}}_\alpha = 1,\ \bar{\mathbf{x}}_\alpha > 0,\ \bar{y}_\alpha \geq 0,\ \alpha = 1,\dots,p,
\end{aligned} \tag{3.1.1}$$

where $T > 0$ is the final time of integration, $\mathbf{b}_0 \in \mathbb{R}^s_+$ is a given initial composition-vector and the flux $\mathbf{j}$ is defined by

$$\mathbf{j}(\mathbf{b}(t), \mathbf{x}^{\mathcal{I}}_\alpha(t), R(t)) = \mathbf{H}(R(t))\left(\mathbf{b}^{\text{tot}} - N\mathbf{b}(t) - \frac{1}{\mathcal{R}_c T}\exp\left(\boldsymbol{\nabla} g(\mathbf{x}^{\mathcal{I}}_\alpha(t)) + \ln(\mathbf{p}^o_g)\right)\right).$$

Let us remind that the exponent $\mathcal{I}$ is added to specify that $\alpha \in \mathcal{I}$ is such that $y_\alpha > 0$ (inactive constraint).

In the time splitting scheme, the differential equations are solved with the Crank-Nicolson method. Let $h > 0$ be a fixed time step, $t^n = n\,h$, $n = 0,\dots,m$, the discretization of $[0,T]$ with $t^m = T$. The approximations of the variables $\mathbf{b}$, R, $\mathbf{x}_\alpha$, and y_α, $\alpha = 1,\dots,p$, at time t^n are respectively denoted by $\mathbf{b}^n$, R^n, $\mathbf{x}^n_\alpha$, and y^n_α, $\alpha = 1,\dots,p$. The set $\mathcal{I}(t^n)$ and $\mathcal{A}(t^n)$ respectively denote the set of inactive/active inequality constraints defined at t^n. For this section one has $\mathcal{I}(t^n) = \mathcal{I}(t^0)$ and $\mathcal{A}(t^n) = \mathcal{A}(t^0)$, $\forall n \in \{0,\dots,m\}$.

The differential equations discretized in time with the Crank-Nicolson scheme consist of

$$\frac{1}{h}(\mathbf{b}^{n+1} - \mathbf{b}^n) = \frac{1}{2}\mathbf{j}(\mathbf{b}^n, \mathbf{x}^{\mathcal{I},n}_\alpha, R^n) + \frac{1}{2}\mathbf{j}(\mathbf{b}^{n+1}, \mathbf{x}^{\mathcal{I},n+1}_\alpha, R^{n+1}). \tag{3.1.2}$$

With the definition of the flux $\mathbf{j}$ at time t^{n+1} the equation (3.1.2) leads to

$$\begin{aligned}\mathbf{b}^{n+1} &= \left(\mathbf{I} + \frac{Nh}{2}\mathbf{H}(R^{n+1})\right)^{-1} \left[\mathbf{b}^n + \frac{h}{2}\mathbf{j}(\mathbf{b}^n, \mathbf{x}_\alpha^{\mathcal{I},n}, R^n)\right. \\ &\quad \left. + \frac{h}{2}\mathbf{H}(R^{n+1})\left(\mathbf{b}^{\text{tot}} - \frac{1}{\mathcal{R}_c T}\exp\left(\boldsymbol{\nabla} g(\mathbf{x}_\alpha^{\mathcal{I},n+1}) + \ln(\mathbf{p}_g^o)\right)\right)\right]. \qquad (3.1.3)\end{aligned}$$

Since $\mathbf{H}(R^{n+1})$ is a diagonal matrix, the term $\mathbf{I} + \frac{Nh}{2}\mathbf{H}(R^{n+1})$ is a diagonal matrix and its inverse is also diagonal with diagonal components defined by $\frac{1}{1+\frac{Nh}{2}H_{ii}(R^{n+1})}$, for $i = 1, \ldots, s$.

Discretizing the whole system (3.1.1), one obtains the following system to solve at each time step: find $\mathbf{b}^{n+1}, \mathbf{x}_\alpha^{n+1} \in \mathbb{R}^s$ and $y_\alpha^{n+1}, R^{n+1} \in \mathbb{R}$ for $\alpha = 1, \ldots, p$ that satisfy

$$\begin{aligned}\mathbf{b}^{n+1} &= \left(\mathbf{I} + \frac{Nh}{2}\mathbf{H}(R^{n+1})\right)^{-1} \left[\mathbf{b}^n + \frac{h}{2}\mathbf{j}(\mathbf{b}^n, \mathbf{x}_\alpha^{\mathcal{I},n}, R^n)\right. \\ &\quad \left. + \frac{h}{2}\mathbf{H}(R^{n+1})\left(\mathbf{b}^{\text{tot}} - \frac{1}{\mathcal{R}_c T}\exp\left(\boldsymbol{\nabla} g(\mathbf{x}_\alpha^{\mathcal{I},n+1}) + \ln(\mathbf{p}_g^o)\right)\right)\right], \\ R^{n+1} &= \left(\frac{3}{4\pi}\sum_{i=1}^{s}\frac{m_{c,i} b_i^{n+1}}{\rho_i}\right)^{\frac{1}{3}}, \qquad (3.1.4) \\ \{\mathbf{x}_\alpha^{n+1}, y_\alpha^{n+1}\}_{\alpha=1}^p &= \underset{\{\bar{\mathbf{x}}_\alpha, \bar{y}_\alpha\}_{\alpha=1}^p}{\operatorname{argmin}} \sum_{\alpha=1}^p \bar{y}_\alpha\, g(\bar{\mathbf{x}}_\alpha) \\ &\qquad \text{s.t. } \sum_{\alpha=1}^p \bar{y}_\alpha \bar{\mathbf{x}}_\alpha = \mathbf{b}^{n+1}, \\ &\qquad \mathbf{e}^T\bar{\mathbf{x}}_\alpha = 1,\ \bar{\mathbf{x}}_\alpha > 0,\ \bar{y}_\alpha \geq 0,\ \alpha = 1, \ldots, p.\end{aligned}$$

The system (3.1.4) can be expressed as the fixed point of a function $\mathbf{G} : \mathbb{R}^{s+1+sp+p} \longrightarrow \mathbb{R}^{s+1+sp+p}$ of the variables $\mathbf{b}^{n+1}, R^{n+1}, \mathbf{x}_\alpha^{n+1}$ and y_α^{n+1}, $\alpha = 1, \ldots, p$, that is defined as the right hand side of (3.1.4):

$$\begin{pmatrix} \mathbf{b}^{n+1} \\ R^{n+1} \\ \mathbf{x}_1^{n+1} \\ \vdots \\ \mathbf{x}_p^{n+1} \\ y_1^{n+1} \\ \vdots \\ y_p^{n+1} \end{pmatrix} = \mathbf{G}(\mathbf{b}^{n+1}, R^{n+1}, \mathbf{x}_1^{n+1}, \ldots, \mathbf{x}_p^{n+1}, y_1^{n+1}, \ldots, y_p^{n+1}).$$

In order to obtain such a fixed point, a sequence of iterates $((\mathbf{b}^{n+1})_l, (R^{n+1})_l, (\mathbf{x}_1^{n+1})_l, \ldots, (\mathbf{x}_p^{n+1})_l, (y_1^{n+1})_l, \ldots, (y_p^{n+1})_l)$ is computed and the following iterative algorithm is constructed, based on the following facts *i)* the system is mass-conserving, and *ii)* $\mathbf{x}_\alpha^{n+1}$ and y_α^{n+1}, $\alpha = 1, \ldots, p$, are computed independently.

Algorithm 3.1.1 (Fixed-point algorithm)**.** *For the resolution of (3.1.4)*

- *initialize* $(\mathbf{b}^{n+1})_0 = \mathbf{b}^n$, $(R^{n+1})_0 = R^n$, $(\mathbf{x}_\alpha^{n+1})_0 = \mathbf{x}_\alpha^n$ *and* $(y_\alpha^{n+1})_0 = y_\alpha^n$ *for* $\alpha = 1, \dots, p$.
- *For* $l = 0, 1, 2, \dots$, *execute the following steps until some stopping criterion is satisfied or the maximum number of iterations is reached*
    1. *Compute the composition-vector* $(\mathbf{b}^{n+1})_{l+1}$ *with the formula (3.1.3), namely*

$$
\begin{aligned}
(\mathbf{b}^{n+1})_{l+1} \;=\; & \left(\mathbf{I} + \frac{Nh}{2}\mathbf{H}((R^{n+1})_l)\right)^{-1} \Bigg[\mathbf{b}^n + \frac{h}{2}\mathbf{j}(\mathbf{b}^n, \mathbf{x}_\alpha^{\mathcal{I},n}, R^n) \\
& + \frac{h}{2}\mathbf{H}((R^{n+1})_l)\left(\mathbf{b}^{tot} - \frac{1}{\mathcal{R}_c T}\exp\left(\boldsymbol{\nabla} g((\mathbf{x}_\alpha^{\mathcal{I},n+1})_l) + \ln(\mathbf{p}_g^o)\right)\right)\Bigg].
\end{aligned}
$$

    2. *Compute the radius*

$$
(R^{n+1})_{l+1} = \left(\frac{3}{4\pi}\sum_{i=1}^{s}\frac{m_{c,i}(b_i^{n+1})_{l+1}}{\rho_i}\right)^{\frac{1}{3}}.
$$

    3. *Solve the optimization problem with the composition-vector* $(\mathbf{b}^{n+1})_{l+1}$ *to compute* $(\mathbf{x}_\alpha^{n+1})_{l+1}$ *and* $(y_\alpha^{n+1})_{l+1}$ *for* $\alpha = 1, \dots, p$:

$$
\begin{aligned}
\{(\mathbf{x}_\alpha^{n+1})_{l+1}, (y_\alpha^{n+1})_{l+1}\}_{\alpha=1}^p \;=\; & \operatorname*{argmin}_{\{\bar{\mathbf{x}}_\alpha, \bar{y}_\alpha\}_{\alpha=1}^p} \sum_{\alpha=1}^{p} \bar{y}_\alpha\, g(\bar{\mathbf{x}}_\alpha) \\
& s.t. \quad \sum_{\alpha=1}^{p} \bar{y}_\alpha \bar{\mathbf{x}}_\alpha \;=\; (\mathbf{b}^{n+1})_{l+1}, \\
& \qquad\; \mathbf{e}^T\bar{\mathbf{x}}_\alpha \;=\; 1,\ \bar{\mathbf{x}}_\alpha \;>\; 0,\ \alpha \;=\; 1, \dots, p, \\
& \qquad\; \bar{y}_\alpha \;\geq\; 0, \qquad\qquad\quad \alpha \;=\; 1, \dots, p.
\end{aligned}
$$

- *Set* $\mathbf{b}^{n+1} = (\mathbf{b}^{n+1})_{l+1}$, $R^{n+1} = (R^{n+1})_{l+1}$, $\mathbf{x}_\alpha^{n+1} = (\mathbf{x}_\alpha^{n+1})_{l+1}$, *and* $y_\alpha^{n+1} = (y_\alpha^{n+1})_{l+1}$ *for* $\alpha = 1, \dots, p$.

The chosen stopping criterion is based on the relative error between two consecutive iterates. In other words the above algorithm is stopped when the following criterion is satisfied

$$
\|(\mathbf{b}^{n+1})_{l+1} - (\mathbf{b}^{n+1})_l\|_2 \;\leq\; tol\,\|(\mathbf{b}^{n+1})_{l+1}\|_2,
$$

where *tol* is a given tolerance and $\|\cdot\|_2$ is the Euclidean norm.

Note that once $\mathbf{b}^{n+1}$ and $\mathbf{x}_\alpha^{\mathcal{I},n+1}$ are computed, the gas concentration-vectors $\mathbf{c}_g^{\infty,n+1}$ and $\mathbf{c}_g^{surf,n+1}$ are given by

$$
\begin{aligned}
\mathbf{c}_g^{\infty,n+1} \;&=\; \mathbf{b}^{\text{tot}} - N\mathbf{b}^{n+1}, \\
\mathbf{c}_g^{surf,n+1} \;&=\; \frac{1}{\mathcal{R}_c T}\exp\left(\boldsymbol{\nabla} g(\mathbf{x}_\alpha^{\mathcal{I},n+1}) + \ln(\mathbf{p}_g^o)\right).
\end{aligned}
$$

Hence at each iteration $l+1$ of the fixed-point method one has to solve the optimization problem to determine $(\mathbf{x}_\alpha^{n+1})_{l+1}$ and $(y_\alpha^{n+1})_{l+1}$ for $\alpha = 1, \ldots, p$. The optimization problem is solved with the interior-point method described in Chapter 2. Since several minimization problems are solved in the Algorithm 3.1.1 and that all these optimization problems are defined for slightly different composition-vectors $(\mathbf{b}^{n+1})_{l+1}$, a strategy is elaborated to reduce the number of interior-point iterations, the key-point being the initialization of the interior-point method. Such techniques are called *warm-start* strategies in optimization theory and are introduced in the following subsection.

3.1.2 A warm-start strategy

Warm-start strategies are used when a sequence of closely related optimization problems are solved. These techniques are based on the expectation that if the change in the data of the problem is small enough, the change in the optimal solution is also small. In other words two closely related optimization problems should in general share similar characteristics. Hence taking advantage of the resolution of an original optimization problem, computational costs can be reduced for solving a closely related problem. The techniques that identify an advanced starting point for the solution of a nearby optimization problem using the information gained from the original one are referred to as *warm-start* strategies. When no such information is used, the new problem is solved from a so-called *cold-start*.

Several warm-start strategies for linear programming have been developed. Among others one can cite the works of Benson and Shanno [9], Gondzio [44], Gondzio and Grothey [45], John and Yıldırım [57] or Yıldırım and Wright [108]. The case of nonlinear programming is less studied. Benson and Shanno proposed some issues in [10]. All these strategies are elaborated techniques originating generally from the generic warm-start algorithm developed by Yıldırım and Wright in [108]. This algorithm can be summarized as follows. Suppose that an original optimization problem is solved with a primal-dual interior-point method and that iterates generated during the resolution are saved. First the last iterate is considered and an adjustment is computed. If the adjusted iterate is an acceptable starting point for the closely related optimization problem, it is considered as a successful warm-start. Otherwise the algorithm returns to the next most advanced iterate among the stored iterates and repeats the same procedure. If none of the stored iterates yields an acceptable warm-start, then the algorithm reverts to cold-start.

In our case, the cold-start strategy corresponds to the initialization procedure introduced in Subsection 2.2.3. Since an active set strategy is added to the interior-point method and that a rapid resolution method is required, only the last iterate is considered for the warm-start. If the resolution of the new minimization problem does not succeed with the warm-start strategy based on the last iteration, the cold-start is chosen for the initialization. Our warm-start strategy is not as elaborated as the one's proposed in [9, 44, 57] because no adjustment in the primal and dual variables is computed. The primal variables y_α and $\mathbf{x}_\alpha$, $\alpha = 1, \ldots, p$, and the dual ones $\boldsymbol{\lambda}$ and θ_α, $\alpha = 1, \ldots, p$ are initialized by taking their value at the last iterate of the previous optimization problem. The only adjustment is applied to the barrier parameter ν that is computed thanks to the complementarity relation and given by

$$\nu^0 = \frac{\sum_{\alpha=1}^{p} \theta_\alpha y_\alpha}{p}, \tag{3.1.5}$$

where θ_α and y_α, $\alpha = 1, \ldots, p$ are the solution of the original optimization problem.

To summarize, at the initial time step t^0 the first optimization problem to solve is initialized with a cold-start since no previous optimization problem exists. The barrier parameter ν and the primal and dual variables $\mathbf{x}_\alpha, y_\alpha, \eta, \gamma, \theta_\alpha$, $\alpha = 1, \ldots, p$ are thus initialized following the techniques presented in Subsection 2.2.3. For all the following minimization problems and as long as the number of inactive inequality constraints remains fixed, the warm-start strategy of equation (3.1.5) is employed.

The resolution of the optimization problem may not succeed. The first kind of failure is the non-convergence of the active set/ interior-point/ Newton Algorithm 2.2.1. The maximum number of iterations is reached whereas the discrepancy is still greater than a fixed tolerance. The second kind of failure is due to the trajectory of $\mathbf{x}_\alpha$, $\alpha \in \mathcal{I}$ on the phase diagram. As described in Section 1.7 the vectors $\mathbf{x}_\alpha$, $\alpha = 1, \ldots, p$ must remain in their respective convex region $\Delta^{'}_{s,\alpha}$ in order to ensure the admissibility of the solution. However at the boundary of these convex regions, the Hessian of the reduced function f is singular. Then if a vector $\mathbf{x}_\alpha$ comes closer to this boundary, the ill-conditioning of $\boldsymbol{\nabla}^2 f$ may lead to unstable Newton iterates and the resolution of the optimization problem fails.

If one of the failure cases occurs when warm-start techniques are used, the resolution is stopped and restarts with a cold-start strategy for the initialization procedure. In the case of a failure with the cold-start, the interior-point method can not be used to solve the optimization problem and the resolution method is stopped. According to [4, 5], this case does not exist in theory and is rare in practice.

The warm-start strategy is helpful for decreasing the computational time. However this strategy encourages the solution of the optimization problem to stay in phase simplices of the same dimension and consequently to enforce the global solution of the optimization problem to follow a minimizer of the Gibbs energy that may become a local minimizer. The detection of the change of dimension of phase simplices, or equivalently the detection of activation/deactivation of inequality constraints is adressed in Section 3.2. The technique relies on the active set strategy with the update of the set of inactive constraints $\mathcal{I}$, and on the Gibbs tangent plane criterion presented in Section 1.7.

3.2 Detection of the discontinuity times and points

Let us begin this section by recalling the definition of the activation and the deactivation of the inequality constraints $y_\alpha(t) \geq 0$, $\alpha = 1, \ldots, p$.

The activation of the inequality constraint $y_{\bar{\alpha}}(t) \geq 0$ corresponds to the transition from the state $y_{\bar{\alpha}}(t) > 0$ to the state $y_{\bar{\alpha}}(t) = 0$. The minimal time t^* such that the transition occurs is called the *activation time*. At this particular time t^* the variable $y_{\bar{\alpha}}$ is truncated to zero and its first derivative is discontinuous. It also induces a discontinuity in the first derivative of the variables y_α, $\alpha \in \mathcal{I}$ and of all the other variables in the optimization problem, namely $\mathbf{x}_\alpha$ and θ_α for $\alpha \in \mathcal{I}$, and $\boldsymbol{\lambda}$. Furthermore the loss of regularity of $\boldsymbol{\lambda}$ at t^* implies the loss of regularity of the gas concentration-vector $\mathbf{c}_g^{surf}$ at t^*.

Similarly the deactivation of the inequality constraint $y_{\bar{\alpha}}(t) \geq 0$ refers to the transition from the state $y_{\bar{\alpha}}(t) = 0$ to $y_{\bar{\alpha}}(t) > 0$. The minimal time $t^\dagger$ such that the transition occurs is called the *deactivation time* and as well as $y_{\bar{\alpha}}$, $\mathbf{x}_{\bar{\alpha}}$ and $\theta_{\bar{\alpha}}$, the variables y_α, $\mathbf{x}_\alpha$ and θ_α for $\alpha \in \mathcal{I}$, $\boldsymbol{\lambda}$ and $\mathbf{c}_g^{surf}$ lose their regularity at $t^\dagger$.

The times t^* and $t^\dagger$ are also called *discontinuity times*, and the above-mentioned variables, together with $\mathbf{b}$ evaluated at the discontinuity time t^* or $t^\dagger$ are called *discontinuity points*.

The discontinuity points have to be detected with accuracy [35, 42, 47, 48], although the time at which the discontinuities occurs is not known in advance. Moreover, this time is not explicitly described by an *event function*, since the variables y_α and $\mathbf{x}_\alpha$, $\alpha = 1, \ldots, p$, are the result of an underlying minimization problem for given $\mathbf{b}$.

Numerical methods for the tracking of discontinuity points usually consist of two steps: (i) the detection of the time interval $[t^n, t^{n+1}]$ that contains the event; (ii) the accurate computation of the event time and discontinuity points. The first step applied to (3.1.1) is detailed in this current section.

3.2.1 Criterion for the detection

Since no explicit function characterizes the activation or the deactivation, one needs to define criteria instead. An activation or a deactivation is expressed through the variables y_α, $\alpha = 1, \ldots, p$ as explained above. However thanks to the definition of the set of indices $\mathcal{I}(t)$ both events can also be expressed as follows

- an activation occurs in the time interval $[t^n, t^{n+1}]$ if
$$\operatorname{card}(\mathcal{I}(t^{n+1})) = \operatorname{card}(\mathcal{I}(t^n)) - 1,$$
- a deactivation occurs in the time interval $[t^n, t^{n+1}]$ if
$$\operatorname{card}(\mathcal{I}(t^{n+1})) = \operatorname{card}(\mathcal{I}(t^n)) + 1.$$

The active set identification procedure in the resolution of the optimization problem detailed in Section 2.2.1 allows to update the set of inactive constraints $\mathcal{I}(t)$. Let us suppose that we solve the optimization problem with the composition vector $(\mathbf{b}^{n+1})_{l+1}$ (step 3 of the Algorithm 3.1.1). The resolution performs the Algorithm 2.2.1. Let $\mathcal{I}_{l+1}^j(t^{n+1})$ denotes the inactive set at the j^{th} iterate of the Algorithm 2.2.1 for the resolution of the minimization problem defined for $(\mathbf{b}^{n+1})_{l+1}$, and $\mathcal{A}_{l+1}^j(t^{n+1})$ the corresponding active set. The inactive set is then updated as follows

$$\begin{aligned}\mathcal{I}_{l+1}^{j+1}(t^{n+1}) \;=\; \Big(\mathcal{I}_{l+1}^j(t^{n+1}) \cup \{\alpha \in \mathcal{A}_{l+1}^j(t^{n+1}) \,|\, \bar{\theta}_\alpha^{j+1} = g(\bar{\mathbf{x}}_\alpha^{j+1}) + \bar{\boldsymbol{\lambda}}^{j+1,T} \bar{\mathbf{x}}_\alpha^{j+1} \;<\; 0\}\Big) \\ \setminus \; \{\alpha \in \mathcal{I}_{l+1}^j(t^{n+1}) \,|\, 0 \;<\; \bar{y}_\alpha^{j+1} \;<\; \epsilon_y\}.\end{aligned}$$

If such an index α exists, then an event occurs in the time interval $[t^n,\, t^{n+1}]$. If $\alpha \in \mathcal{A}^j_{l+1}(t^{n+1})$, a deactivation is detected. If $\alpha \in \mathcal{I}^j_{l+1}(t^{n+1})$, an activation is detected. The Algorithm 3.1.1 is then stopped and the computation of the discontinuity time and points starts.

Thus the active set identification procedure constitutes a first class of criteria for the detection of either an activation or a deactivation of an inequality constraint. Note that in the resolution of the optimization problem the variable $\bar{\mathbf{x}}_\alpha$ is not updated as long as $\alpha \in \mathcal{A}$. Consequently in the active set identification procedure, $\bar{\theta}^{\mathcal{A}}_\alpha$ is always computed with the same $\bar{\mathbf{x}}^{\mathcal{A}}_\alpha$. If $\bar{\mathbf{x}}^{\mathcal{A}}_\alpha$ is not accurate the deactivation may be missed. Such a situation is illustrated in Figure 3.1 (left), where the Gibbs free energy g is represented by a black curve. The black straight lines are the supporting tangent plane at two consecutive points of the simulation $(\mathbf{b}^n, g(\mathbf{b}^n))$ and $(\mathbf{b}^{n+1}, g(\mathbf{b}^{n+1}))$ when the deactivation is missed.

As described in Section 1.7 the deactivation of a constraint occurs when the supporting tangent plane to the energy function g becomes tangent to a new point on the graph of the function. In the example considered in Figure 3.1, $s = 2$, $\Delta_r = [0, 1]$ and we suppose that $\mathbf{b}^n$ is the last correct single-phase point, situated on the left side of the phase diagram. Moreover let us assume that $\mathbf{b}^n$ moves to the right until the area where both inequality constraints are deactivated and the deactivation of the second constraint does not occur. In that case $\mathbf{b}^{n+1}$ remains a single-phase point and the corresponding tangent plane crosses the graph of the function. The mole-fraction vector associated to the active constraint is denoted by $\mathbf{x}^{\mathcal{A}}_2$ and is located on the right side of the phase diagram. From the first time-step, and because of the warm-start strategy, the variable $\mathbf{x}^{\mathcal{A}}_2$ is not updated. In Figure 3.1 (left) the point $(\mathbf{x}^{\mathcal{A}}_2, g(\mathbf{x}^{\mathcal{A}}_2))$ is situated above the supporting tangent plane and consequently the variable $\theta^{\mathcal{A}}_2$, which represents the distance between the supporting tangent plane and $(\mathbf{x}^{\mathcal{A}}_2, g(\mathbf{x}^{\mathcal{A}}_2))$ remains positive. Thus the deactivation is not detected and the index 2 remains in $\mathcal{A}$ (eventhough the tangent plane at $(\mathbf{b}^{n+1}, g(\mathbf{b}^{n+1}))$ crosses the graph of g). The deactivation is missed because of the poor approximation of $\mathbf{x}^{\mathcal{A}}_2$.

Consequently the criterion stemmed from the active set identification procedure is not sufficient to detect the deactivation. Let us find another criterion. The deactivation occurs when the supporting tangent plane crosses the curve g. Since the function g is known only point-wise, the intersection between the supporting tangent plane and the graph of g cannot be computed analytically. However, it is not necessary to compute this intersection, but only to find one point $(\mathbf{x}, g(\mathbf{x}))$ located below the tangent plane. Let us sign the distance between $(\mathbf{x}, g(\mathbf{x}))$ and the supporting tangent plane in such a way that the distance is said to be positive if $(\mathbf{x}, g(\mathbf{x}))$ lies above the tangent plane, equal to zero if $(\mathbf{x}, g(\mathbf{x}))$ is situated on the tangent plane and negative if $(\mathbf{x}, g(\mathbf{x}))$ is below the tangent plane. The points at negative distance are located in the convex areas associated to the active constraints $\Delta^{'}_{s,\alpha}$, $\alpha \in \mathcal{A}$. Since there is no condition on $\mathbf{x}^{\mathcal{A}}_\alpha$ except $\mathbf{e}^T\mathbf{x}^{\mathcal{A}}_\alpha - 1 = 0$ and that these points are not updated in the resolution of the optimization problems, we define $\mathbf{x}^{\mathcal{A}}_\alpha$ such that the point $(\mathbf{x}^{\mathcal{A}}_\alpha, g(\mathbf{x}^{\mathcal{A}}_\alpha))$ is situated at minimal distance from the supporting tangent plane. If $d^n(\mathbf{x})$ denotes the signed distance between $(\mathbf{x}, g(\mathbf{x}))$ and the supporting tangent plane at time t^n, then the criterion to detect the presence of the deactivation of an inequality

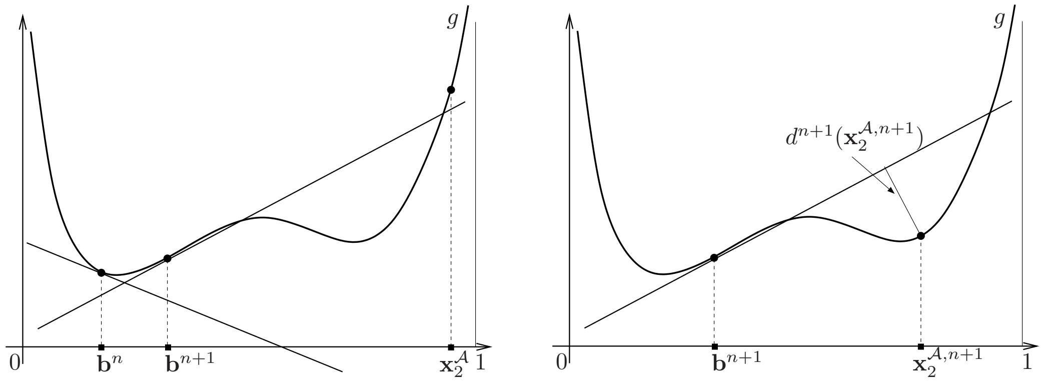

g
0
$\mathbf{b}^n$
$\mathbf{b}^{n+1}$
$\mathbf{x}_2^{\mathcal{A}}$
1
g
$d^{n+1}(\mathbf{x}_2^{\mathcal{A},n+1})$
0
$\mathbf{b}^{n+1}$
$\mathbf{x}_2^{\mathcal{A},n+1}$
1

Figure 3.1: Deactivation of an inequality constraint. Left: the point $\mathbf{x}_2^{\mathcal{A}}$ remains unchanged during the simulation and $(\mathbf{x}_2^{\mathcal{A}}, g(\mathbf{x}_2^{\mathcal{A}}))$ is situated above the supporting tangent plane. In that case the deactivation of the constraint is missed. Right: the variable $\mathbf{x}_2^{\mathcal{A}}$ is updated at each time step as the point at minimal distance to the supporting tangent plane. In that case the point $(\mathbf{x}_2^{\mathcal{A},n+1}, g(\mathbf{x}_2^{\mathcal{A},n+1}))$ is situated below the supporting tangent plane and the deactivation is detected in the time interval $[t^n, t^{n+1}]$.

constraint is to check at each time step t^{n+1} if

$$\exists\, \bar{\alpha} \in \mathcal{A}(t^{n+1}) \text{ such that } d^n(\mathbf{x}_{\bar{\alpha}}^n) > 0 \text{ and } d^{n+1}(\mathbf{x}_{\bar{\alpha}}^{n+1}) < 0, \tag{3.2.1}$$

where $\mathbf{x}_{\bar{\alpha}}^n$, $\mathbf{x}_{\bar{\alpha}}^{n+1} \in \Delta'_{s,\bar{\alpha}}$ are the points that respectively minimize $d^n(\cdot)$ and $d^{n+1}(\cdot)$ in the convex area $\Delta'_{s,\bar{\alpha}}$.

Remark 3.2.1. *The points at minimal distance to the tangent plane are not computed at each active set identification procedure of the optimization problems, but only at each time step. Since the difference between two successives iterates $\mathbf{b}_l^n$ in the fixed-point algorithm is usually not large, it is enough to compute the point at minimal distance $\mathbf{x}_\alpha^{\mathcal{A}}$ at each time step.*

The computation of the points that minimize the distance to the supporting tangent plane is explained in the following section. Before ending this section let us consider a last criterion for the detection of an activation or a deactivation of an inequality constraint.

As explained in the previous section, the primal-dual interior-point method may fail. If the warm-start strategy was used to initialized the interior-point method, the failure may indicate that an event has occurred in the considered time interval. Hence when the cold-start strategy is used in order to solve the interior-point method, one needs to compare the number of inactive inequality constraints with the one of the previous time step. If these 2 numbers are not equal, it indicates that either an activation (if the new number of inactive constraints is smaller) or a deactivation of a constraint occurs.

Let us summarize the criteria of detection described in this section. Suppose that the time interval is $[t^n, t^{n+1}]$. If one of the following criteria is satisfied

- there exist an iterate l in the fixed-point algorithm and an iterate j in the Algorithm 2.2.1 for the resolution of the optimization problem defined for $(\mathbf{b}^{n+1})_{l+1}$, such that
    - ▶ $\mathrm{card}(\mathcal{I}_{l+1}^{j+1}(t^{n+1})) = \mathrm{card}(\mathcal{I}_{l+1}^{j}(t^{n+1})) - 1$ (activation),
    - ▶ or $\mathrm{card}(\mathcal{I}_{l+1}^{j+1}(t^{n+1})) = \mathrm{card}(\mathcal{I}_{l+1}^{j}(t^{n+1})) + 1$ (deactivation);
- there exists an iterate l in the fixed-point algorithm for which the resolution of the optimization problem defined for $(\mathbf{b}^{n+1})_{l+1}$ failed with the warm-start strategy, succeed with the cold-start strategy, but whose number of inactive constraints stemmed from the resolution with the cold-start strategy, is
    - ▶ or less than $\mathrm{card}(\mathcal{I}(t^n))$ (activation),
    - ▶ greater than $\mathrm{card}(\mathcal{I}(t^n))$ (deactivation);
- there exists an index $\alpha \in \mathcal{A}(t^{n+1})$ such that the distance of the point $(\mathbf{x}_\alpha^{n+1}, g(\mathbf{x}_\alpha^{n+1}))$ to the supporting tangent plane is negative (deactivation);

then stop the numerical resolution of (3.1.1) and compute the discontinuity time and points.

Note that in the above-mentioned criteria, the last criterion is specific to the detection of a deactivation, whereas the others stand for both types of events.

3.2.2 Computation of the minimal distance criterion

Let us determine first the equation describing the supporting tangent plane as well as the distance between the plane and any points $(\mathbf{x}, g(\mathbf{x}))$, $\mathbf{x} \in \mathbb{R}^s_{++}$. As described in Section 1.7 the supporting tangent plane is the affine hyperplane tangent to the graph of g at the points $(\mathbf{x}_\alpha, g(\mathbf{x}_\alpha))$, $\alpha \in \mathcal{I}$. Since $\boldsymbol{\nabla} g(\mathbf{x}_\alpha) = \boldsymbol{\nabla} g(\mathbf{x}_\beta)$, $\forall \alpha, \beta \in \mathcal{I}$, the normal vector to the tangent plane is uniquely determined. Similarly to Chapter 2 the supporting tangent plane is then defined by the set of points $(\mathbf{x}, x_{s+1}) \in \mathbb{R}^s \times \mathbb{R}$ satisfying

$$\boldsymbol{\nabla} g(\mathbf{x}_\alpha)^T(\mathbf{x}_\alpha - \mathbf{x}) + x_{s+1} - g(\mathbf{x}_\alpha) \; = \; 0,$$

where $\mathbf{x}_\alpha$ is any of the points with $\alpha \in \mathcal{I}$.

Since $\boldsymbol{\nabla} g(\mathbf{x})^T\mathbf{x} \; = \; g(\mathbf{x})$, $\forall \mathbf{x} \in \mathbb{R}^s_{++}$ the definition of the hyperplane is reduced to

$$-\boldsymbol{\nabla} g(\mathbf{x}_\alpha)^T\mathbf{x} + x_{s+1} \; = \; 0.$$

The vector $\mathbf{x}_\alpha$ being solution of the PEP and α belonging to $\mathcal{I}$, the relation $\boldsymbol{\lambda} \; = \; -\boldsymbol{\nabla} g(\mathbf{x}_\alpha)$ holds (cf relation (1.4.15)) and the above equation becomes

$$\boldsymbol{\lambda}^T\mathbf{x} + x_{s+1} \; = \; 0.$$

The signed distance of any point $(\mathbf{x}, g(\mathbf{x}))$ to the tangent plane is thus given by

$$d(\mathbf{x}) \; = \; \frac{\boldsymbol{\lambda}^T\mathbf{x} + g(\mathbf{x})}{\|\mathbf{n}\|_2},$$

where $\mathbf{n} = \left(-\boldsymbol{\lambda}^T, -1\right)^T$ and $\|\mathbf{n}\|_2$ is the Euclidean norm of $\mathbf{n}$. Since the vector $\mathbf{n}$ is independent of $\mathbf{x}$, let us consider the *normalized* distance, again denoted by d, defined by

$$d(\mathbf{x}) = \boldsymbol{\lambda}^T\mathbf{x} + g(\mathbf{x}). \tag{3.2.2}$$

Hence at each time step t^{n+1} of the time discretization algorithm, and for all inactive constraints $\alpha \in \mathcal{A}$, the computation of the point $\mathbf{x}_\alpha^{\mathcal{A},n+1} \in \Delta'_{s,\alpha}$ is given by the resolution of the following minimization problem

$$\mathbf{x}_\alpha^{\mathcal{A},n+1} = \arg\min_{\mathbf{x}\in\Delta'_{s,\alpha}} d^{n+1}(x) = \arg\min_{\mathbf{x}\in\Delta'_{s,\alpha}} \boldsymbol{\lambda}^{n+1,T}\mathbf{x} + g(\mathbf{x}), \tag{3.2.3}$$

where $\boldsymbol{\lambda}^{n+1}$ stems from the resolution of the PEP defined for $\mathbf{b}^{n+1}$ in Algorithm 3.1.1 and d^{n+1} denotes the distance function d defined at time t^{n+1}.

The distance function d possesses several local minima. Each $\mathbf{x}_\alpha^{\mathcal{I}}$ realizes a local minimum that is such that $d(\mathbf{x}_\alpha^{\mathcal{I}}) = 0$. If no deactivation occurs, the points $\mathbf{x}_\alpha^{\mathcal{I}}$ achieve the global minimum since $d(\mathbf{x}_\alpha^{\mathcal{A}}) > 0$, $\forall\alpha \in \mathcal{A}$. The determination of $\mathbf{x}_\alpha^{\mathcal{A}}$ corresponds in fact to finding the point located in $\Delta'_{s,\alpha}$ that realizes the local minima of the distance function.

In the minimization problem (3.2.3) $\mathbf{x}$ is constrained to remain in the domain $\Delta'_{s,\alpha}$. This domain is not known explicitly and its size can vary widely. One way to characterize $\Delta'_{s,\alpha}$ is to impose a constraint to (3.2.3) that expresses the positive-definiteness of the Hessian matrix $\boldsymbol{\nabla}^2 g(\mathbf{x})$. Since it is difficult to handle such a constraint, another approach is considered by solving a new minimization problem where the sole condition on $\mathbf{x}$ is $\mathbf{e}^T\mathbf{x} - 1 = 0$ and the constraint $\mathbf{x} \in \Delta'_{s,\alpha}$ is imposed weakly via a variable step length during the resolution. This new problem is defined as follows

$$\begin{aligned} \mathbf{x}_\alpha^{\mathcal{A},n+1} = \quad & \arg\min_{\mathbf{x}} \quad \boldsymbol{\lambda}^{n+1,T}\mathbf{x} + g(\mathbf{x}), \\ & \text{s.t.} \qquad \mathbf{e}^T\mathbf{x} - 1 = 0. \end{aligned} \tag{3.2.4}$$

The KKT conditions relative to (3.2.4) lead to the nonlinear system:

$$\begin{aligned} \boldsymbol{\nabla} g(\mathbf{x}) + \boldsymbol{\lambda}^{n+1} + \zeta\mathbf{e} &= \mathbf{0}, \\ \mathbf{e}^T\mathbf{x} - 1 &= 0, \end{aligned} \tag{3.2.5}$$

where $\zeta \in \mathbb{R}$ is a Lagrange multiplier associated to the equality constraint $\mathbf{e}^T\mathbf{x} - 1 = 0$. The unknowns are $\mathbf{x}$ and ζ, and the size of (3.2.5) is $s + 1$, which is small by opposition to the optimization problem arising in the PEP. However the small nonlinear system (3.2.5) has to be solved at each time step and for all $\alpha \in \mathcal{A}$. The resolution method must not be time consuming numerically.

Problem (3.2.5) is solved with the Newton method and the corresponding Newton system reads

$$\begin{pmatrix} \boldsymbol{\nabla}^2 g(\mathbf{x}) & \mathbf{e} \\ \mathbf{e}^T & 0 \end{pmatrix} \begin{pmatrix} \mathbf{p}_\mathbf{x} \\ p_\zeta \end{pmatrix} = -\begin{pmatrix} \boldsymbol{\nabla} g(\mathbf{x}) + \boldsymbol{\lambda}^{n+1} + \zeta\mathbf{e} \\ \mathbf{e}^T\mathbf{x} - 1 \end{pmatrix}, \tag{3.2.6}$$

where $\mathbf{p}_\mathbf{x}$ and p_ζ are the increments corresponding to the variables $\mathbf{x}$ and ζ.

Lemma 3.2.1. *If* $\mathbf{x}$ *belongs to a convex region of* g*, (3.2.6) is solvable.*

Proof. If $\mathbf{x}$ remains in a convex region of g, $\boldsymbol{\nabla}^2 g(\mathbf{x})$ is symmetric positive definite and the inertia theorem (see *e.g.* [46]) allows to conclude that the matrix of (3.2.6) is invertible. ☐

Following Lemma 3.2.1, the numerical algorithm for the solution of (3.2.6) must pay attention to building a sequence of iterates that remains in the convex region $\Delta'_{s,\alpha}$. The initial guess of the Newton method is given either in a neighborhood of the vertices of the simplex Δ'_s (as the initial guesses of the interior-point method described in [5]), or by the converged iterate obtained in the convex region $\Delta'_{s,\alpha}$ at the previous time step (continuation approach [29]).

For each iterate $\mathbf{x}^i$ of the Newton sequence, the sequence is re-initialized at $\mathbf{x}^{i-1}$ and stopped if the Hessian $\boldsymbol{\nabla}^2 g(\mathbf{x}^i)$ is not positive definite or if the point $\mathbf{x}^i$ goes out of the simplex. In order to guarantee the convergence to the local minimum, the Newton increments are also controlled with a step length algorithm [29] in order to ensure that the iterates remain in the convex region $\Delta'_{s,\alpha}$ and the Hessian remains positive definite. More precisely, let c_{thres} be a given threshold (that corresponds to an approximation of the distance between convex regions), if $\det(\boldsymbol{\nabla}^2 g(\mathbf{x}^i))$ is close to zero, and $\|(\mathbf{p}_{\mathbf{x}}, p_\zeta)^T\|_2 > c_{\text{thres}}$, then the Newton iterates are computed as

$$\begin{pmatrix} \mathbf{x}^{i+1} \\ \zeta^{i+1} \end{pmatrix} = \begin{pmatrix} \mathbf{x}^{i} \\ \zeta^{i} \end{pmatrix} + \alpha_i \begin{pmatrix} \mathbf{p}_{\mathbf{x}} \\ p_\zeta \end{pmatrix}, \quad \alpha_i = \frac{c_{\text{thres}}}{\|(\mathbf{p}_{\mathbf{x}}, p_\zeta)^T\|_2}; \tag{3.2.7}$$

otherwise the new iterate $(\mathbf{x}^{i+1}, \zeta^{i+1})^T$ is computed with $\alpha_i = 1$.

Since the points $\mathbf{x}^i$ lie in the simplex Δ'_s, the parameter c_{thres} is initialized to 0.1. However, the distance between the convex areas could be smaller than 0.1 and, therefore the value of c_{thres} can be empirically updated at each time step by computing the minimal distance between all the $\mathbf{x}_\alpha$, $\alpha = 1, \ldots, p$.

Figure 3.2 illustrates the influence of the modification of the increments given by (3.2.7) for the scalar case $r = 1$. A 2-components chemical system composed of 1-hexacosanol and pinic acid is considered. The simplex Δ_1 is the segment $[0, 1]$ (0 meaning 100% of pinic acid in the system). As in Section 1.7 the index 1 is attributed to the constraint situated on the left extremity of Δ_1 and 2 stands for the one on the right extremity. In this example the inactive constraint is situated on the right and the active constraint is on the left.

The distance function d is represented with a bold curve, and the derivative ∇d is symbolized with a dashed curve. The tangent lines for the determination of the next iterate in the Newton method are the black straight lines. The black squares correspond to the successive Newton iterates, $\mathbf{x}^0$ being the starting point. The black circles are therefore the successive values $g(\mathbf{x}^k)$, $k = 0, \ldots, i, i+1, \ldots$. In this example d contains only one minimum that is $d(\mathbf{x}_2^{\mathcal{I},n+1})$ and the minimizer of d on $\Delta'_{1,1}$ is the right edge of $\Delta'_{1,1}$ where the Hessian of g becomes singular.

Figure 3.2 (left) shows the minimizing sequence obtained with the Newton method without the adaptive step length (3.2.7). The iterate $\mathbf{x}^{i+1}$ leaves the convex area $\Delta'_{1,1}$ and jumps to the convex area of the inactive constraint because the Newton system is ill-conditionned around $\mathbf{x}^i$. Consequently the sequence converges to the global minimizer and $\mathbf{x}_1^{\mathcal{A},n+1} = \mathbf{x}_2^{\mathcal{I},n+1}$. Since the Hessian of g at each iterate is positive definite, the jump is not detected and the sequence converges to the global minimizer instead of the minimizer belonging to $\Delta'_{1,1}$.

Figure 3.2 (right) illustrates the convergence of the sequence with step length modification. The iterate $\mathbf{x}^{i+1}$ is modified by (3.2.7) and when the new value falls in the area where the Hessian is not positive definite, the Newton method is stopped and $\mathbf{x}_1^{\mathcal{A},n+1}$ is set to $\mathbf{x}^i$ which is located near the local minimizer.

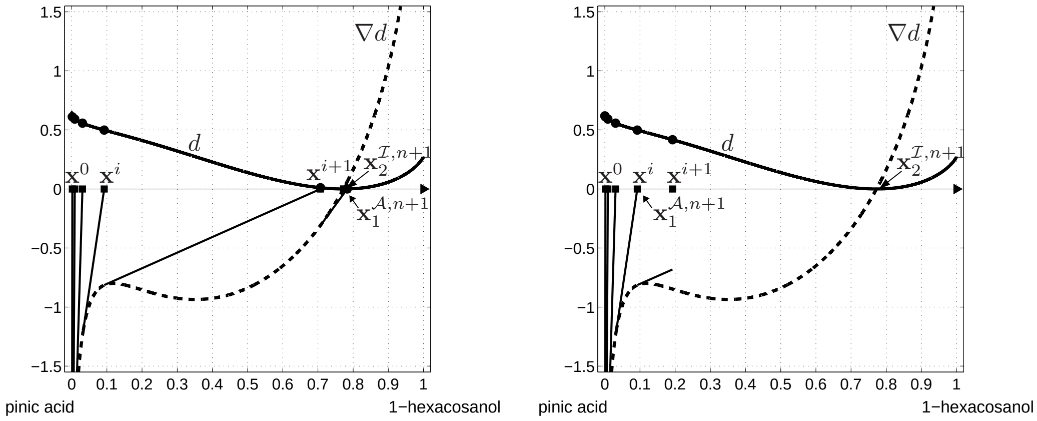


Figure 3.2: Left: Steps of the Newton algorithm for the computation of the point at minimal distance to the tangent plane without the criterion on the increment ($\alpha_i = 1$). Right: same but with the criterion on the increment.

At each iteration of the Newton method the distance is computed and the algorithm is stopped if the distance is negative. Otherwise the algorithm stops if the stopping criterion on the Euclidean norm of the residuals is smaller than a fixed tolerance, or if a maximal number of iterations K is reached.

The converged iterate of the Newton method serves as the initial guess of the Newton method at the next time step, i.e. a classical continuation method for the computation of the point at minimal distance of the tangent plane is used (see *e.g.* [6, 29]). The algorithm for the computation of the minimal distance is summarized as follows:

Algorithm 3.2.1. *At each time step* t^{n+1} *and for each inequality constraint such that* $\alpha \in \mathcal{A}(t^n)$, *initialize* $\mathbf{x}^0 = \mathbf{x}_\alpha^n$ *and* $\zeta^0 = \zeta^n$. *Then, for* $i = 1, \ldots, K$

(i) Build and solve the system (3.2.6) to obtain $\mathbf{p}_\mathbf{x}^i$ *and* p_ζ^i.

(ii) If $\det \boldsymbol{\nabla}^2 g(\mathbf{x}^{i-1}) \le \delta$, δ *given, and* $\|(\mathbf{p}_\mathbf{x}^i, p_\zeta^i)^T\|_2 > c_{\text{thres}}$ *then set*

$$\begin{pmatrix} \mathbf{x}^i \\ \zeta^i \end{pmatrix} = \begin{pmatrix} \mathbf{x}^{i-1} \\ \zeta^{i-1} \end{pmatrix} + \frac{c_{\text{thres}}}{\|(\mathbf{p}_\mathbf{x}^i, p_\zeta^i)^T\|_2} \begin{pmatrix} \mathbf{p}_\mathbf{x}^i \\ p_\zeta^i \end{pmatrix};$$

else compute $\mathbf{x}^i = \mathbf{x}^{i-1} + \mathbf{p}_\mathbf{x}^i$ *and* $\zeta^i = \zeta^{i-1} + p_\zeta^i$.

(iii) *If* $\nabla^2 g(\mathbf{x}^i)$ *is not positive definite, or* $\mathbf{x}^i$ *does not belong to the simplex* $\Delta_s^{'}$*, or if the Newton method does not converge, STOP and set* $\mathbf{x}_\alpha^{n+1} = \mathbf{x}^{i-1}$*.*

(iv) *If the distance to the supporting tangent plane is negative, if the stopping criterion is satisfied, or if* $i = K$*, STOP and set* $\mathbf{x}_\alpha^{n+1} = \mathbf{x}^i$*.*

Remark 3.2.2. *Since some constraints can have the same convex area, it is important to check at the end of the Newton method if* $\mathbf{x}^i$ *can be expressed as a linear combination of* $\mathbf{x}_\alpha^{n+1}$*,* $\alpha \in \mathcal{I}(t^{n+1})$*. In that case,* $\mathbf{x}^i$ *has to be reset to* $\mathbf{x}^0$*.*

Once an event (activation or deactivation) is detected, the exact time and points of discontinuity are computed. Their computation is detailed in the following section.

3.3 Computation of the discontinuity times and points

In the literature about ordinary differential equations with discontinuities, the location of the discontinuities is based on a combination of the *discontinuity locking approach* introduced by Cellier in [23] and interpolation procedures. The discontinuity locking approach consists in *locking* the system of equations during an integration step. In other words the system of equations is not modified during the integration even if an event occurs and implies a change in the system. The solver completes the integration as if no event occurs. This approach eliminates the difficulties of integration over discontinuities. Moreover according to Park and Barton [81] this approach is efficient and correct if the system of equations is mathematically well behaved in a small interval after the event, even if the solution is not physically meaningful.

Almost all of the algorithms present in the literature follow the same strategy: first to integrate through the discontinuity locking approach and second to look for events using an interpolant. The first use of interpolants is due to Shampine et al. in [97]. Then many techniques to define the interpolating polynomials were proposed. Esposito and Kumar in [33] and Mao and Petzold in [66] give a review of the main techniques. According to Esposito and Kumar [33] the event detection proposed by Park and Barton [81] seems to be the most reliable technique in the literature to date. However the authors point out that these methods fail to locate events which are close to regions where the right hand side of the differential system is undefined. The reason is that each of these methods attempts to evaluate the right hand side before determining if an event has occurred. In order to overcome this failure Esposito and Kumar elaborate a new technique that approaches the event from only one side. This technique consists in constructing an extrapolation polynomial in order to select the integration step size by checking for potential future events and avoiding the need to evaluate the differential equations in potentially singular regions.

The techniques to detect an activation or a deactivation of an inequality constraint follow the discontinuity locking approach. Concerning the computation of the discontinuity time and points, the extrapolation method is in fact more suitable than the use of interpolating polynomials as it is showned in the sequel. Let us distinguish the cases of an activation and a deactivation of a constraint, and assume that the event occurs in the time interval $[t^n, t^{n+1}]$.

3.3.1 Activation of an inequality constraint

Let us assume that the $\bar{\alpha}^{th}$ constraint activates and that the activation occurs during the interval $[t^n, t^{n+1}]$. This situation is depicted in Figure 3.3 (left) for a generic time evolution of the activating variable $y_{\bar{\alpha}}$ against the time. The variable $y_{\bar{\alpha}}$ is represented by a blue curve. Since the constraint activates, $y_{\bar{\alpha}}$ at t^n is positive and equal to zero at t^{n+1}. The purpose of this section is to approximate the discontinuity time t^* where $y_{\bar{\alpha}}$ is equal to zero for the first time.

One can observe in Figure 3.3 (left) that $y_{\bar{\alpha}}$ is truncated to zero at t^* and its first time derivative is discontinuous. This truncation implies the inefficiency of the interpolating polynomials to locate t^*. Indeed the interpolants are prone to be zero only in a neighborhood of t^{n+1}. Consequently the evaluation of the interpolants in the neighborhood of t^* is positive and t^* cannot be determined. An example of interpolation called $y_{\bar{\alpha},h}$ is given in Figure 3.3 (right). This interpolant is such that

$$\begin{aligned} y_{\bar{\alpha},h}(t^n) &= y_{\bar{\alpha}}(t^n), \quad & y_{\bar{\alpha},h}(t^{n+1}) &= y_{\bar{\alpha}}(t^{n+1}), \\ y'_{\bar{\alpha},h}(t^n) &= y'_{\bar{\alpha}}(t^n), \quad & y'_{\bar{\alpha},h}(t^{n+1}) &= y'_{\bar{\alpha}}(t^{n+1}). \end{aligned}$$

Clearly one has $y_{\bar{\alpha},h}(t^*) > 0$. However extrapolation techniques are fully practicable because they are not subjected to the truncation.

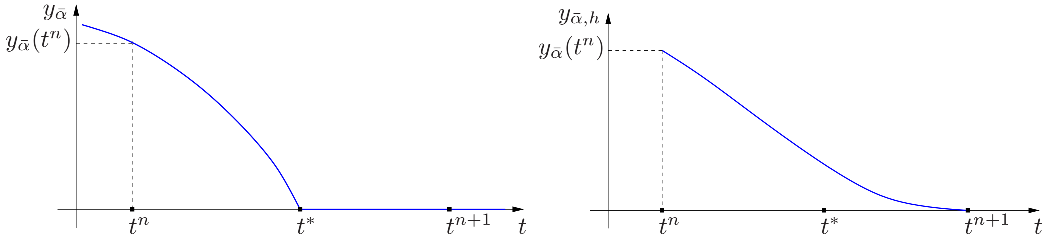


Figure 3.3: Left: The function $y_{\bar{\alpha}}$ on the time interval $[t^n, t^{n+1}]$. The $\bar{\alpha}^{th}$ constraint is activated at time t^*. Right: An interpolant of $y_{\bar{\alpha}}$ on the time interval $[t^n, t^{n+1}]$.

In order to define the extrapolation polynomial, let us follow the technique proposed by Esposito and Kumar in [33]. The underlying idea is to construct a polynomial, whose accuracy is of the same order as the underlying integration algorithm, and that extrapolates the event function on $[t^n, t^{n+1}]$. For the case of an activation, the event function is defined by $y_{\bar{\alpha}}$ and the event is characterized by the smallest fractional time step τ which satisfies $y_{\bar{\alpha}}(t^n + \tau) = 0$. A Taylor series expansion gives

$$0 = y_{\bar{\alpha}}(t^n + \tau) = y_{\bar{\alpha}}(t^n) + \tau \frac{\mathrm{d}}{\mathrm{d}t} y_{\bar{\alpha}}(t^n) + \frac{\tau^2}{2} \frac{\mathrm{d}^2}{\mathrm{d}t^2} y_{\bar{\alpha}}(t^n) + \mathcal{O}(\tau^3). \tag{3.3.1}$$

Let us remind that the variable $y_{\bar{\alpha}}$ is only known point-wise as a result of an optimization problem. Consequently the determination of its successive derivatives in time is not straightforward. For this reason let us only consider the first derivative in the definition of the extrapolation polynomial. This polynomial $P^a_{\bar{\alpha}}$ is defined by

$$P^a_{\bar{\alpha}}(\tau) \;=\; y_{\bar{\alpha}}(t^n) \;+\; \tau \frac{\mathrm{d}}{\mathrm{d}t} y_{\bar{\alpha}} \left(t^n + \frac{\tau}{2}\right). \tag{3.3.2}$$

The discontinuity fractional time τ is then given by the root of $P^a_{\bar{\alpha}}$.

The value $y_{\bar{\alpha}}(t^n)$ is already approximated by $y^n_{\bar{\alpha}}$. The first derivative $\frac{\mathrm{d}}{\mathrm{d}t} y_{\bar{\alpha}} \left(t^n + \frac{\tau}{2}\right)$ remains to be estimated. First since $y_{\bar{\alpha}}$ comes from the optimization problem let us transform the time derivative by using the chain rule

$$\frac{\mathrm{d}}{\mathrm{d}t} y_{\bar{\alpha}} \left(t^n + \frac{\tau}{2}\right) \;=\; \sum_{i=1}^{s} \frac{\partial y_{\bar{\alpha}}}{\partial b_i} \left(\mathbf{b}\left(t^n + \frac{\tau}{2}\right)\right) \frac{\mathrm{d}}{\mathrm{d}t} b_i \left(t^n + \frac{\tau}{2}\right). \tag{3.3.3}$$

Since no information from the time step t^{n+1} should be used and following the technique in [33] the derivatives $\frac{\mathrm{d}}{\mathrm{d}t} b_i \left(t^n + \frac{\tau}{2}\right)$, $i = 1, \ldots, s$, are approximated with the straight line passing through $\left(t^{n-1}, \mathbf{j}(\mathbf{b}^{n-1}, \mathbf{x}^{\mathcal{I},n-1}_{\alpha}, R^{n-1})\right)$ and $\left(t^n, \mathbf{j}(\mathbf{b}^n, \mathbf{x}^{\mathcal{I},n}_{\alpha}, R^n)\right)$, namely

$$\frac{\mathrm{d}}{\mathrm{d}t} \mathbf{b} \left(t^n + \frac{\tau}{2}\right) \;\approx\; \left(1 + \frac{\tau}{2h}\right) \mathbf{j}(\mathbf{b}^n, \mathbf{x}^{\mathcal{I},n}_{\alpha}, R^n) - \frac{\tau}{2h} \mathbf{j}(\mathbf{b}^{n-1}, \mathbf{x}^{\mathcal{I},n-1}_{\alpha}, R^{n-1}). \tag{3.3.4}$$

In this equation both fluxes $\mathbf{j}(\mathbf{b}^n, \mathbf{x}^{\mathcal{I},n}_{\alpha}, R^n)$ and $\mathbf{j}(\mathbf{b}^{n-1}, \mathbf{x}^{\mathcal{I},n-1}_{\alpha}, R^{n-1})$ come from the results of the discretized system (3.1.4) at the time step t^n and t^{n-1} and no further computation is required. Hence the first kind of derivatives in (3.3.3) is approximated. Concerning the second kind of derivatives, namely $\frac{\partial y_{\bar{\alpha}}}{\partial b_i} \left(\mathbf{b}\left(t^n + \frac{\tau}{2}\right)\right)$, $i = 1, \ldots, s$, the approximation of $\frac{\partial y_{\bar{\alpha}}}{\partial b_i} \left(\mathbf{b}\left(t^n\right)\right)$ is computed instead because optimization features furnish good estimates in that case.

Fiacco and McCormick proved in [36] that it is possible to be explicit about the derivative of the primal and dual variables of an optimization problem if appropriate assumptions are made about conditions holding at the optimum. The technique follows from the sensitivity analysis of the optimization problem. Let us present the sensitivity analysis in the following subsection and then apply this analysis to the approximation of the derivatives $\frac{\partial y_{\bar{\alpha}}}{\partial b_i} \left(\mathbf{b}\left(t^n\right)\right)$, $i = 1, \ldots, s$.

Sensitivity analysis in nonlinear optimization

The sensitivity analysis of a model in numerical analysis consists in estimating the variation in the solution when the inputs of the model are slightly perturbed. In optimization theory the purpose is similar: to estimate how much the optimum changes when changes are made in the constraints and the objective function. The theory developed below is inspired from Fiacco and McCormick [36].

For the statement of the sensitivity analysis let us consider the following nonlinear optimization problem: find $\mathbf{y}^* \in \mathbb{R}^n$ that satisfies

$$\begin{aligned} \min_{\mathbf{y}} \quad & \mathcal{G}(\mathbf{y}) \\ \text{s.t.} \quad & c_i(\mathbf{y}) \geq 0, \qquad \forall i = 1, \ldots, n_{\mathscr{I}}, \\ & c_i(\mathbf{y}) = 0, \qquad \forall i = 1, \ldots, n_{\mathscr{E}}. \end{aligned} \tag{3.3.5}$$

The associated perturbed optimization problem is defined as follows

$$\begin{aligned} \min_{\mathbf{y}} \quad & \mathcal{G}(\mathbf{y}) + a_0 \kappa_0(\mathbf{y}) \\ \text{s.t.} \quad & c_i(\mathbf{y}) + a_i \varphi_i(\mathbf{y}) \geq 0, \qquad \forall i = 1, \ldots, n_{\mathscr{I}}, \\ & c_i(\mathbf{y}) + a_{i+n_{\mathscr{I}}} \psi_i(\mathbf{y}) = 0, \qquad \forall i = 1, \ldots, n_{\mathscr{E}}, \end{aligned} \tag{3.3.6}$$

where κ_0, $\{\varphi_i\}_{i=1}^{n_{\mathscr{I}}}$ and $\{\psi_i\}_{i=1}^{n_{\mathscr{E}}}$ are scalar-valued functions of $\mathbf{y}$ and the $\{a_i\}_{i=0}^{n_{\mathscr{I}}+n_{\mathscr{E}}}$ are scalar. Let $\mathbf{a}$ denote the vector of the scalars a_i, $i = 0, \ldots, n_{\mathscr{I}} + n_{\mathscr{E}}$, i.e. $\mathbf{a}^T = (a_0, a_1, \ldots, a_{n_{\mathscr{I}}+n_{\mathscr{E}}})$.

As discussed in Subsection 1.4.3 let us define the following sets

- the set of inequality constraints: $\mathscr{I} = \{1, \ldots, n_{\mathscr{I}}\}$,
- the set of equality constraints: $\mathscr{E} = \{1, \ldots, n_{\mathscr{E}}\}$,
- the set of feasible points: $\Omega = \{\mathbf{y} \in \mathbb{R}^n \,|\, c_i(\mathbf{y}) = 0, \forall i \in \mathscr{E};\, c_i(\mathbf{y}) \geq 0, \forall i \in \mathscr{I}\}$,
- the active set at any feasible point $\mathbf{y} \in \Omega$: $\mathscr{A}(\mathbf{y}) = \mathscr{E} \cup \{i \in \mathscr{I} \,|\, c_i(\mathbf{y}) = 0\}$.

Given a feasible point $\mathbf{y}$, a direction vector, say $\mathbf{k}$, is feasible if there exists $\ell > 0$ such that $\mathbf{y} + \ell\,\mathbf{k}$ is feasible. One can define the set of linearized feasible directions $\mathscr{F}$ at a given feasible point $\mathbf{y}$.

Definition 3.3.1. *Given a feasible point* $\mathbf{y}$ *and the active set* $\mathscr{A}(\mathbf{y})$*, the set of linearized feasible directions* $\mathscr{F}(\mathbf{y})$ *is*

$$\mathscr{F}(\mathbf{y}) = \left\{ \mathbf{k} \in \mathbb{R}^n \,\middle|\, \begin{array}{ll} \mathbf{k}^T \boldsymbol{\nabla} c_i(\mathbf{y}) = \mathbf{0}, & \forall i \in \mathscr{E}, \\ \mathbf{k}^T \boldsymbol{\nabla} c_i(\mathbf{y}) \geq \mathbf{0}, & \forall i \in \mathscr{A}(\mathbf{y}) \cap \mathscr{I} \end{array} \right\}.$$

When the Karush-Kuhn-Tucker conditions are satisfied (see Theorem 1.4.1), a move along any feasible direction $\mathbf{k} \in \mathscr{F}(\mathbf{y}^*)$ either increases the first-order approximation to the objective function ($\mathbf{k}^T \boldsymbol{\nabla} \mathcal{G}(\mathbf{y}^*) > 0$), or else keeps this value the same ($\mathbf{k}^T \boldsymbol{\nabla} \mathcal{G}(\mathbf{y}^*) = 0$). For this latter case one cannot determine from the first derivative information alone whether a move along $\mathbf{k}$ increases or decreases the objective function. Second-order conditions examine the second derivative terms in the Taylor series expansions of $\mathcal{G}$ and c_i, $i \in \mathscr{E} \cup \mathscr{I}$, to see whether this extra information resolves the issue of increase or decrease in $\mathcal{G}$. The set of feasible directions for which it is not clear from first derivative information alone whether $\mathcal{G}$ increases or decreases, is called the *critical cone* and defined as follows [75]

Definition 3.3.2. *Given* $\mathscr{F}(\mathbf{y}^*)$ *and a Lagrange multiplier vector* $\boldsymbol{\lambda}^*$ *satisfying the Karush-Kuhn-Tucker conditions. The critical cone is the set* $\mathcal{C}(\mathbf{y}^*, \boldsymbol{\lambda}^*)$ *defined by*

$$\mathcal{C}(\mathbf{y}^*, \boldsymbol{\lambda}^*) = \{\mathbf{k} \in \mathscr{F}(\mathbf{y}^*) | \boldsymbol{\nabla} c_i(\mathbf{y}^*)^T\mathbf{k} = 0, \forall i \in \mathscr{A}(\mathbf{y}^*) \cap \mathscr{I} \text{ with } \lambda_i^* > 0\}.$$

The definition of $\mathscr{F}$ allows to express the definition of the critical cone in an equivalent way

$$\mathbf{k} \in \mathcal{C}(\mathbf{y}^*, \boldsymbol{\lambda}^*) \Leftrightarrow \begin{cases} \boldsymbol{\nabla} c_i(\mathbf{y}^*)^T\mathbf{k} = 0, & \forall i \in \mathscr{E}, \\ \boldsymbol{\nabla} c_i(\mathbf{y}^*)^T\mathbf{k} = 0, & \forall i \in \mathscr{A}(\mathbf{y}^*) \cap \mathscr{I} \text{ with } \lambda_i^* > 0, \\ \boldsymbol{\nabla} c_i(\mathbf{y}^*)^T\mathbf{k} \geq 0, & \forall i \in \mathscr{A}(\mathbf{y}^*) \cap \mathscr{I} \text{ with } \lambda_i^* = 0. \end{cases}$$

The theorem that provides the second-order sufficient conditions on $\mathcal{G}$ and c_i, $i \in \mathscr{E} \cup \mathscr{I}$ to ensure $\mathbf{y}^*$ is a local minimum, is the following [36, 75].

Theorem 3.3.1. *Sufficient conditions that a point* $\mathbf{y}^*$ *be an isolated (unique locally) local minimum of (3.3.5), where* f *and* c_i, $i \in \mathscr{E} \cup \mathscr{I}$ *are twice-differentiable functions, are that there exists a Lagrangian vector* $\boldsymbol{\lambda}^*$ *such that the following Karush-Kuhn-Tucker conditions are satisfied for* $(\mathbf{y}^*, \boldsymbol{\lambda}^*)$

$$\begin{aligned} \boldsymbol{\nabla}_{\mathbf{y}}\mathcal{L}(\mathbf{y}^*, \boldsymbol{\lambda}^*) &= \mathbf{0}, \\ c_i(\mathbf{y}^*) &= 0, \quad \forall\, i \in \mathscr{E}, \\ c_i(\mathbf{y}^*) &\geq 0, \quad \forall\, i \in \mathscr{I}, \\ \lambda_i^* &\geq 0, \quad \forall\, i \in \mathscr{I}, \\ \lambda_i^* c_i(\mathbf{y}^*) &= 0, \quad \forall\, i \in \mathscr{I}; \end{aligned} \tag{3.3.7}$$

and that $\forall \mathbf{k} \in \mathcal{C}(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ *with* $\mathbf{k} \neq \mathbf{0}$ *one has*

$$\mathbf{k}^T \boldsymbol{\nabla}^2_{\mathbf{yy}} \mathcal{L}(\mathbf{y}^*, \boldsymbol{\lambda}^*)\mathbf{k} > 0.$$

Then $\mathbf{y}^*$ *is a strict local solution of (3.3.5).*

Before stating the theorem about the optimum of the perturbed problem (3.3.6) from [36], let us separate the Lagrange multiplier $\boldsymbol{\lambda}$ between the equality and inequality constraints. Equivalently, let us write $\boldsymbol{\lambda}^T = (\mathbf{u}^T, \mathbf{w}^T)$ where $\mathbf{u} \in \mathbb{R}^{n_{\mathscr{I}}}$ is the Lagrange multiplier associated to the inequality constraints and $\mathbf{w} \in \mathbb{R}^{n_{\mathscr{E}}}$ is for the equality constraints. Moreover let us associate the index i to the inequality constraints and the index j to the equality constraints if nothing is specified. Finally let $\mathcal{L}^*$ denote $\mathcal{L}(\mathbf{y}^*, \mathbf{u}^*, \mathbf{w}^*)$ and so on for all functions evaluated at $\mathbf{y}^*, \mathbf{u}^*$ or $\mathbf{w}^*$. The theorem reads

Theorem 3.3.2. *If (a) the functions* $\mathcal{G}$ *and* c_i, $i \in \mathscr{E} \cup \mathscr{I}$, *are twice-differentiable, (b) the Karush-Kuhn-Tucker conditions (3.3.7) hold at* $\mathbf{y}^*$, *(c) the gradients* $\boldsymbol{\nabla} c_i(\mathbf{y}^*)$, $i \in \mathscr{A}(\mathbf{y}^*)$ *are linearly independent, and (d) strict complementarity holds (that is* $u_i^* > 0$ *when* $c_i(\mathbf{y}^*) = 0$ *for* $i \in \mathscr{I}$*), then*

- *(i) the multipliers* $\mathbf{u}_i^*$, $i \in \mathscr{I}$ *and* $\mathbf{w}_j^*$, $j \in \mathscr{E}$ *are unique;*
- *(ii) there exists a differentiable function* $(\mathbf{y}(\mathbf{a}), \mathbf{u}(\mathbf{a}), \mathbf{w}(\mathbf{a}))$ *in the neighborhood of* $\mathbf{0}$, *where* $\mathbf{y}(\mathbf{a})$ *is a local minimum of problem (3.3.6), and* $(\mathbf{u}(\mathbf{a}), \mathbf{w}(\mathbf{a}))$ *are the multipliers associated with it, where* $\lim_{\mathbf{a}\to\mathbf{0}}(\mathbf{y}(\mathbf{a}), \mathbf{u}(\mathbf{a}), \mathbf{w}(\mathbf{a})) = (\mathbf{y}^*, \mathbf{u}^*, \mathbf{w}^*)$,

(iii) a differential approximation to

$$\begin{pmatrix} \mathbf{y}(\mathbf{a}) - \mathbf{y}^* \\ \mathbf{u}(\mathbf{a}) - \mathbf{u}^* \\ \mathbf{w}(\mathbf{a}) - \mathbf{w}^* \end{pmatrix}$$

is given by

$$\begin{pmatrix} \boldsymbol{\nabla}^2_{\mathbf{yy}}\mathcal{L}^* & -\mathbf{C}^*_{\mathscr{I}} & \mathbf{C}^*_{\mathscr{E}} \\ \mathbf{U}^*(\mathbf{C}^*_{\mathscr{I}})^T & \operatorname{diag}(c_i^*) & \mathbf{0} \\ (\mathbf{C}^*_{\mathscr{E}})^T & \mathbf{0} & \mathbf{0} \end{pmatrix}^{-1} \begin{pmatrix} -\boldsymbol{\nabla}\kappa_o^* & \mathbf{U}^*\boldsymbol{\Phi}^* & -\mathbf{W}^*\boldsymbol{\Psi}^* \\ \mathbf{0} & \operatorname{diag}(-u_i^*\varphi_i^*) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \operatorname{diag}(-\psi_j^*) \end{pmatrix} \mathbf{a},$$

where $\mathbf{U}^* = \operatorname{diag}(u_i^*)$, $\mathbf{C}^*_{\mathscr{I}} = (-\boldsymbol{\nabla}c_1^*, \dots, -\boldsymbol{\nabla}c^*_{n_{\mathscr{I}}})$, $\mathbf{C}^*_{\mathscr{E}} = (\boldsymbol{\nabla}c_1^*, \dots, \boldsymbol{\nabla}c^*_{n_{\mathscr{E}}})$, $\boldsymbol{\Phi}^* = (\boldsymbol{\nabla}\varphi_1^*, \dots, \boldsymbol{\nabla}\varphi^*_{n_{\mathscr{I}}})$, $\mathbf{W}^* = \operatorname{diag}(w_j^*)$, *and* $\Psi^* = (\boldsymbol{\nabla}\psi_1^*, \dots, \psi^*_{n_{\mathscr{E}}})$,

(iv) the change in the optimum value of the objective function $\mathcal{G}(\mathbf{y}(\mathbf{a})) - \mathcal{G}(\mathbf{y}^*)$ *is approximated by*

$$-\sum_{i=1}^{n_{\mathscr{I}}} a_i u_i^* \varphi_i^* + \sum_{j=1}^{n_{\mathscr{E}}} a_{j+n_{\mathscr{E}}} w_j^* \psi_j^*.$$

The proof of the theorem can be found in [36]. In their proof, Fiacco and McCormick showed how to be explicit about the derivative of the differentiable functions $\mathbf{y}, \mathbf{u}$ and $\mathbf{w}$ whose existence is assured by the implicit function theorem. Let us summarize the main ideas of the proof.

Because of Assumption (b), the following set of equations is satisfied at $(\mathbf{y}, \mathbf{u}, \mathbf{w}) = (\mathbf{y}^*, \mathbf{u}^*, \mathbf{w}^*)$ and $\mathbf{a} = \mathbf{0}$

$$\begin{aligned} \boldsymbol{\nabla}\mathcal{G}(\mathbf{y}) + a_0\boldsymbol{\nabla}\kappa_0(\mathbf{y}) \quad & -\sum_{i=1}^{n_{\mathscr{I}}} u_i \left[\boldsymbol{\nabla}c_i(\mathbf{y}) + a_i\boldsymbol{\nabla}\varphi_i(\mathbf{y})\right] \quad + \sum_{j=1}^{n_{\mathscr{E}}} w_j \left[\boldsymbol{\nabla}c_j(\mathbf{y}) + a_{j+n_{\mathscr{I}}}\boldsymbol{\nabla}\psi_j(\mathbf{y})\right] = \mathbf{0}, \\ & u_i \left[c_i(\mathbf{y}) + a_i\varphi_i(\mathbf{y})\right] = 0, \qquad i \in \mathscr{I}, \\ & c_j(\mathbf{y}) + a_{j+n_{\mathscr{I}}}\psi_j(\mathbf{y}) = 0, \qquad j \in \mathscr{E}. \end{aligned} \tag{3.3.8}$$

This is a system of $n + n_{\mathscr{I}} + n_{\mathscr{E}}$ equations. The Jacobian matrix of this system with respect to $(\mathbf{y}, \mathbf{u}, \mathbf{w})$ at $(\mathbf{y}^*, \mathbf{u}^*, \mathbf{w}^*)$ and $\mathbf{a} = \mathbf{0}$ is

$$\mathbf{M}^* = \begin{pmatrix} \boldsymbol{\nabla}^2_{\mathbf{yy}}\mathcal{L}^* & -\mathbf{C}^*_{\mathscr{I}} & \mathbf{C}^*_{\mathscr{E}} \\ \mathbf{U}^*(\mathbf{C}^*_{\mathscr{I}})^T & \operatorname{diag}(c_i^*) & \mathbf{0} \\ (\mathbf{C}^*_{\mathscr{E}})^T & \mathbf{0} & \mathbf{0} \end{pmatrix},$$

where the elements of $\mathbf{M}^*$ are those defined in part *(iii)*. Under Assumptions (a)-(d) this matrix has an inverse. From the implicit function theorem a unique differentiable vector function $(\mathbf{y}(\mathbf{a}), \mathbf{u}(\mathbf{a}), \mathbf{w}(\mathbf{a}))$ in a neighborhood about $\mathbf{a} = \mathbf{0}$ is obtained, satisfying the system (3.3.8). Then treating the equations in the system (3.3.8) as functions of **a**, differentiating, evaluating at $(\mathbf{y}^*, \mathbf{u}^*, \mathbf{w}^*)$ and $\mathbf{a} = \mathbf{0}$, and rearranging yields

$$\begin{pmatrix} \boldsymbol{\nabla}^2 \mathcal{L}^* & -\mathbf{C}^*_{\mathscr{I}} & \mathbf{C}^*_{\mathscr{E}} \\ \mathbf{U}^*(\mathbf{C}^*_{\mathscr{I}})^T & \operatorname{diag}(c_i^*) & \mathbf{0} \\ (\mathbf{C}^*_{\mathscr{E}})^T & \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \frac{\mathrm{d}\mathbf{y}}{\mathrm{d}\mathbf{a}}(\mathbf{0}) \\ \frac{\mathrm{d}\mathbf{u}}{\mathrm{d}\mathbf{a}}(\mathbf{0}) \\ \frac{\mathrm{d}\mathbf{w}}{\mathrm{d}\mathbf{a}}(\mathbf{0}) \end{pmatrix} = \begin{pmatrix} -\boldsymbol{\nabla}\kappa_o^* & \mathbf{U}^*\boldsymbol{\Phi}^* & -\mathbf{W}^*\boldsymbol{\Psi}^* \\ \mathbf{0} & \operatorname{diag}(-u_i^*\varphi_i^*) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \operatorname{diag}(-\psi_j^*) \end{pmatrix}, \tag{3.3.9}$$

where the vector of derivatives represents in fact the following matrix

$$\begin{pmatrix} \frac{\mathrm{d}\mathbf{y}}{\mathrm{d}\mathbf{a}}(\mathbf{0}) \\ \frac{\mathrm{d}\mathbf{u}}{\mathrm{d}\mathbf{a}}(\mathbf{0}) \\ \frac{\mathrm{d}\mathbf{w}}{\mathrm{d}\mathbf{a}}(\mathbf{0}) \end{pmatrix} = \begin{pmatrix} \frac{\partial y_1}{\partial a_1} & \cdots & \frac{\partial y_1}{\partial a_{1+n_{\mathscr{I}}+n_{\mathscr{E}}}} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_n}{\partial a_1} & \cdots & \frac{\partial y_n}{\partial a_{1+n_{\mathscr{I}}+n_{\mathscr{E}}}} \\ \frac{\partial u_1}{\partial a_1} & \cdots & \frac{\partial u_1}{\partial a_{1+n_{\mathscr{I}}+n_{\mathscr{E}}}} \\ \vdots & \ddots & \vdots \\ \frac{\partial u_{n_{\mathscr{I}}}}{\partial a_1} & \cdots & \frac{\partial u_{n_{\mathscr{I}}}}{\partial a_{1+n_{\mathscr{I}}+n_{\mathscr{E}}}} \\ \frac{\partial w_1}{\partial a_1} & \cdots & \frac{\partial w_1}{\partial a_{1+n_{\mathscr{I}}+n_{\mathscr{E}}}} \\ \vdots & \ddots & \vdots \\ \frac{\partial w_{n_{\mathscr{E}}}}{\partial a_1} & \cdots & \frac{\partial w_{n_{\mathscr{E}}}}{\partial a_{1+n_{\mathscr{I}}+n_{\mathscr{E}}}} \end{pmatrix} (\mathbf{0}) \in \mathbb{R}^{(n+n_{\mathscr{I}}+n_{\mathscr{E}})\times(1+n_{\mathscr{I}}+n_{\mathscr{E}})}.$$

Hence the resolution of this system allows to compute the exact value of the first derivative of **y**, **u** and **w** with respect to **a** at **0**.

Application to the computation of the discontinuity time

In the relation (3.3.3) about the derivative of $y_{\bar{\alpha}}$ at t^n, an approximation for the partial derivatives $\frac{\partial y_{\bar{\alpha}}}{\partial b_i}(\mathbf{b}(t^n + \frac{\tau}{2}))$, $i = 1, \dots, s$, is needed. Since $y_{\bar{\alpha}}$ is solution of a constrained optimization problem, the sensitivity analysis developed above can be used.

One needs to approach the derivative of $y_{\bar{\alpha}}$ with respect to b_i, $i = 1, \dots, s$ at time $t^n + \frac{\tau}{2}$. However since the minimization problem is already solved at t^n, let us employ the sensitivity analysis to estimate $\frac{\partial y_{\bar{\alpha}}}{\partial b_i}(\mathbf{b}(t^n))$ instead and use this estimate as the approximation of $\frac{\partial y_{\bar{\alpha}}}{\partial b_i}(\mathbf{b}(t^n + \frac{\tau}{2}))$. Hence let us consider the optimization problem defined for $\mathbf{b}^n$

$$\begin{array}{lll} \displaystyle\min_{\{y_\alpha, \mathbf{x}_\alpha\}_{\alpha\in\mathcal{I}(t^n)}} & \displaystyle\sum_{\alpha\in\mathcal{I}(t^n)} y_\alpha\, g(\mathbf{x}_\alpha) & \\ \text{s.t.} & y_\alpha \geq 0, & \forall \alpha \in \mathcal{I}(t^n), \\ & \displaystyle\sum_{\alpha\in\mathcal{I}(t^n)} y_\alpha \mathbf{x}_\alpha - \mathbf{b}^n = \mathbf{0}, & \\ & 1 - \mathbf{e}^T\mathbf{x}_\alpha = 0, & \forall \alpha \in \mathcal{I}(t^n). \end{array}$$

The point $(\mathbf{x}_\alpha^n, y_\alpha^n, \boldsymbol{\lambda}^n, \theta_\alpha^n, \zeta_\alpha^n)$ is a KKT point and already computed. Furthermore since the objective function and the constraints are twice-differentiable functions of y_α and $\mathbf{x}_\alpha$

for $\alpha \in \mathcal{I}(t^n)$ the assumptions of the Theorem 3.3.2 are satisfied at this KKT point. So the sensitivity analysis can be applied and the perturbed optimization system for a general index $i \in \{1, \dots, s\}$ defined as

$$\begin{array}{lll} \displaystyle\min_{\{y_\alpha, \mathbf{x}_\alpha\}_{\alpha \in \mathcal{I}(t^n)}} & \displaystyle\sum_{\alpha \in \mathcal{I}(t^n)} y_\alpha \, g(\mathbf{x}_\alpha) & \\ \text{s.t.} & y_\alpha \ \geq \ 0, & \forall \alpha \in \mathcal{I}(t^n), \\ & \displaystyle\sum_{\alpha \in \mathcal{I}(t^n)} y_\alpha \mathbf{x}_\alpha - \mathbf{b}^n - \mathrm{db}_i \, \mathbf{e}_i \ = \ \mathbf{0}, & \\ & 1 - \mathbf{e}^T \mathbf{x}_\alpha \ = \ 0, & \forall \alpha \in \mathcal{I}(t^n). \end{array}$$

Because only the derivative with respect to db_i is required, the perturbation is only expressed on the second group of equality constraints. In comparison to (3.3.6) the vector $\mathbf{a}$ is defined as: $a_0 = 0$, $a_j = 0$, for $j = 1, \dots, p^{\mathcal{I}}$, for $j = 1, \dots, s$ one has

$$a_{j+p^{\mathcal{I}}} = \left\{ \begin{array}{rl} \mathrm{db}_i & \text{if } j = i, \\ 0 & \text{if } j \neq i, \end{array} \right.$$

and $a_{j+s+p^{\mathcal{I}}} = 0$, for $j = 1, \dots, p^{\mathcal{I}}$. The function ψ_i associated to the perturbation is the function identically equal to -1.

The sensitivity analysis requires first to consider a set of equations formed by the gradient of the Lagrangian associated to the perturbed problem with respect to the primal variables only, the complementarity equation for the inequality constraints and finally the equality constraints. This set reads in that case

$$\begin{array}{rcll} y_\alpha \left(\boldsymbol{\nabla} g(\mathbf{x}_\alpha) + \boldsymbol{\lambda} \right) + \zeta_\alpha \mathbf{e} & = & \mathbf{0}, & \forall \alpha \in \mathcal{I}(t^n), \\ g(\mathbf{x}_\alpha) + \boldsymbol{\lambda}^T \mathbf{x}_\alpha - \theta_\alpha & = & 0, & \forall \alpha \in \mathcal{I}(t^n), \\ \theta_a \, y_\alpha & = & 0, & \forall \alpha \in \mathcal{I}(t^n), \\ \displaystyle\sum_{\alpha \in \mathcal{I}(t^n)} y_\alpha \mathbf{x}_\alpha - \mathbf{b}^n - \mathrm{db}_i \, \mathbf{e}_i & = & \mathbf{0}, & \\ 1 - \mathbf{e}^T \mathbf{x}_\alpha & = & 0, & \forall \alpha \in \mathcal{I}(t^n). \end{array}$$

The next step is to consider this system as a function of $\mathbf{a}^T = (0, \dots, 0, \mathrm{db}_i, 0, \dots, 0)$, differentiating and evaluating at $\mathbf{a} = \mathbf{0}$ and the KKT point $(\mathbf{x}_\alpha^n, y_\alpha^n, \boldsymbol{\lambda}^n, \theta_\alpha^n, \zeta_\alpha^n)$. Since the sole non-zero component of $\mathbf{a}$ is db_i, the differentiation is done only with respect to db_i and one obtains the system

$$\begin{array}{rcl} y_\alpha^n \boldsymbol{\nabla}^2 g(\mathbf{x}_\alpha^n) \dfrac{\mathrm{d}\mathbf{x}_\alpha}{\mathrm{db}_i} + \dfrac{\mathrm{d}y_\alpha}{\mathrm{db}_i} \left(\boldsymbol{\nabla} g(\mathbf{x}_\alpha^n) + \boldsymbol{\lambda}^n \right) + y_\alpha^n \dfrac{\mathrm{d}\boldsymbol{\lambda}}{\mathrm{db}_i} + \dfrac{\mathrm{d}\zeta_\alpha}{\mathrm{db}_i} \mathbf{e} & = & \mathbf{0}, \ \forall \alpha \in \mathcal{I}(t^n), \\ \left(\boldsymbol{\nabla} g(\mathbf{x}_\alpha^n) + \boldsymbol{\lambda}^n \right)^T \dfrac{\mathrm{d}\mathbf{x}_\alpha}{\mathrm{db}_i} + \mathbf{x}_\alpha^{n,T} \dfrac{\mathrm{d}\boldsymbol{\lambda}}{\mathrm{db}_i} - \dfrac{\mathrm{d}\theta_\alpha}{\mathrm{db}_i} & = & 0, \ \forall \alpha \in \mathcal{I}(t^n), \\ \theta_\alpha^n \dfrac{\mathrm{d}\mathbf{x}_\alpha}{\mathrm{db}_i} + y_\alpha^n \dfrac{\mathrm{d}\theta_\alpha}{\mathrm{db}_i} & = & 0, \ \forall \alpha \in \mathcal{I}(t^n), \end{array}$$

$$\begin{aligned}\sum_{\alpha\in\mathcal{I}(t^n)} y_\alpha^n \frac{\mathrm{d}\mathbf{x}_\alpha}{\mathrm{db}_i} + \sum_{\alpha\in\mathcal{I}(t^n)} \frac{\mathrm{d}y_\alpha}{\mathrm{db}_i}\mathbf{x}_\alpha^n - \mathbf{e}_i &= \mathbf{0},\\ -\mathbf{e}^T\frac{\mathrm{d}\mathbf{x}_\alpha}{\mathrm{db}_i} &= 0,\ \forall\alpha\in\mathcal{I}(t^n).\end{aligned}$$

At the KKT point the primal variable y_α^n is positive and the dual variable θ_α^n is equal to 0, $\forall\alpha\in\mathcal{I}(t^n)$. Consequently the derivative $\frac{\mathrm{d}\theta_\alpha}{\mathrm{db}_i}=0$ and the above system simplifies in

$$\begin{aligned}y_\alpha^n\boldsymbol{\nabla}^2 g(\mathbf{x}_\alpha^n)\frac{\mathrm{d}\mathbf{x}_\alpha}{\mathrm{db}_i} + \frac{\mathrm{d}y_\alpha}{\mathrm{db}_i}(\boldsymbol{\nabla} g(\mathbf{x}_\alpha^n)+\boldsymbol{\lambda}^n) + y_\alpha^n\frac{\mathrm{d}\boldsymbol{\lambda}}{\mathrm{db}_i} + \frac{\mathrm{d}\zeta_\alpha}{\mathrm{db}_i}\mathbf{e} &= \mathbf{0},\ \forall\alpha\in\mathcal{I}(t^n),\\ (\boldsymbol{\nabla} g(\mathbf{x}_\alpha^n)+\boldsymbol{\lambda}^n)^T\frac{\mathrm{d}\mathbf{x}_\alpha}{\mathrm{db}_i} + \mathbf{x}_\alpha^{n,T}\frac{\mathrm{d}\boldsymbol{\lambda}}{\mathrm{db}_i} &= 0,\ \forall\alpha\in\mathcal{I}(t^n),\\ \sum_{\alpha\in\mathcal{I}(t^n)} y_\alpha^n \frac{\mathrm{d}\mathbf{x}_\alpha}{\mathrm{db}_i} + \sum_{\alpha\in\mathcal{I}(t^n)} \frac{\mathrm{d}y_\alpha}{\mathrm{db}_i}\mathbf{x}_\alpha^n - \mathbf{e}_i &= \mathbf{0},\\ -\mathbf{e}^T\frac{\mathrm{d}\mathbf{x}_\alpha}{\mathrm{db}_i} &= 0,\ \forall\alpha\in\mathcal{I}(t^n).\end{aligned}$$

The variable θ_α is removed from the system that can be rearranged in the following linear form

$$\begin{pmatrix} y_\alpha^n\boldsymbol{\nabla}^2 g(\mathbf{x}_\alpha^n) & \boldsymbol{\nabla} g(\mathbf{x}_\alpha^n)+\boldsymbol{\lambda}^n & y_\alpha^n\mathbf{I}_s & \mathbf{e}\\ (\boldsymbol{\nabla} g(\mathbf{x}_\alpha^n)+\boldsymbol{\lambda}^n)^T & 0 & \mathbf{x}_\alpha^{n,T} & 0\\ y_\alpha^n\mathbf{I}_s & \mathbf{x}_\alpha^n & \mathbf{0} & \mathbf{0}\\ \mathbf{e}^T & 0 & \mathbf{0} & 0\end{pmatrix}\begin{pmatrix}\frac{\mathrm{d}\mathbf{x}_\alpha}{\mathrm{db}_i}(0)\\ \frac{\mathrm{d}y_\alpha}{\mathrm{db}_i}(0)\\ \frac{\mathrm{d}\boldsymbol{\lambda}}{\mathrm{db}_i}(0)\\ \frac{\mathrm{d}\zeta_\alpha}{\mathrm{db}_i}(0)\end{pmatrix} = \begin{pmatrix}\mathbf{0}\\ 0\\ \mathbf{e}_i\\ 0\end{pmatrix}. \tag{3.3.10}$$

Hence by solving this linear system the approximation of the partial derivative of $y_{\bar\alpha}$ relative to b_i is given by

$$\frac{\partial y_{\bar\alpha}}{\partial b_i}\left(\mathbf{b}\left(t^n+\frac{\tau}{2}\right)\right)\approx\frac{\mathrm{d}y_{\bar\alpha}}{\mathrm{db}_i}(0).$$

Observe that the linear system (3.3.10) is similar to the system (2.2.11) stemmed from the interior-point method with $\mu=0$. Consequently the same method of resolution can be used. Furthermore the elements in the matrix are already computed since they are the result of the resolution of the optimization problem defined for $\mathbf{b}^n$, and the additional computational costs to build and solve this linear system are not expensive.

Then the extrapolation polynomial $P_{\bar\alpha}^a$ can be approximated by

$$\begin{aligned}P_{\bar\alpha}^a &\approx y_{\bar\alpha}^n+\tau\sum_{i=1}^s\frac{\mathrm{d}y_{\bar\alpha}}{\mathrm{db}_i}(0)\left[\left(1+\frac{\tau}{2h}\right)j_i^n-\frac{\tau}{2h}j_i^{n-1}\right]\\ &= y_{\bar\alpha}^n+\tau\sum_{i=1}^s\frac{\mathrm{d}y_{\bar\alpha}}{\mathrm{db}_i}(0)\,j_i^n+\frac{\tau^2}{2h}\sum_{i=1}^s\frac{\mathrm{d}y_{\bar\alpha}}{\mathrm{db}_i}(0)\,(j_i^n-j_i^{n-1}),\end{aligned}$$

where j_i^n and j_i^{n-1} denote respectively the i^{th} component of the fluxes $\mathbf{j}(\mathbf{b}^n,\mathbf{x}_\alpha^{\mathcal{I},n},R^n)$ and $\mathbf{j}(\mathbf{b}^{n-1},\mathbf{x}_\alpha^{\mathcal{I},n-1},R^{n-1})$.

Consequently looking for τ consists in solving the following second order equation in τ

$$y^n_{\bar{\alpha}} + \tau \sum_{i=1}^{s} \frac{\mathrm{d}y_{\bar{\alpha}}}{\mathrm{db}_i}(0)\, j^n_i \;+\; \frac{\tau^2}{2h} \sum_{i=1}^{s} \frac{\mathrm{d}y_{\bar{\alpha}}}{\mathrm{db}_i}(0)\, (j^n_i - j^{n-1}_i) \;=\; 0. \tag{3.3.11}$$

For a time step h sufficiently small, the numerical experiments show that the equation admits a unique positive root τ^* in $[0, h]$. The discontinuity time is then given by $t^*_h = t^n + \tau^*$.

Remark 3.3.1. *If the constant time step h used in the discretization is not small enough in order to ensure $\tau^* \in [0, h]$, then a smaller time step $h^{'}$ is considered and a new fixed-point iteration is done to approximate $\mathbf{b}$ at $t^n + h^{'}$. The time step is chosen such that the activation does not occur at time $t^n + h^{'}$, implying that the new time $t^n + h^{'}$ is closer to the discontinuity time than t^n. The technique for the computation of the discontinuity time is executed on the time interval $[t^n + h^{'}, t^{n+1}]$.*

Once the fractional time step τ^* is computed, the two-steps Adams-Bashforth with variable time step is used to approximate $\mathbf{b}$ at time $t^{n+1} := t^n + \tau^*$, namely

$$\mathbf{b}^{n+1} \;=\; \mathbf{b}^n \;+\; \tau^* \left[\left(1 + \frac{\tau^*}{2h} \right) \mathbf{j}^n \;-\; \frac{\tau^*}{2h} \mathbf{j}^{n-1} \right]. \tag{3.3.12}$$

All the terms in the above equation are already known before the computation of the discontinuity time. Hence the computation of $\mathbf{b}^{n+1}$ requires only addition and multiplication operations.

Finally the values of y^{n+1}_α, $\mathbf{x}^{n+1}_\alpha$, $\boldsymbol{\lambda}^{n+1}$, θ^{n+1}_α and ζ^{n+1}_α for all $\alpha \in \{1, \ldots, p\}$ are obtained by solving the optimization problem for $\mathbf{b}^{n+1}$ of (3.3.12) with a warm-start strategy defined by the solution at t^n with the exception of the removal of the constraint $\bar{\alpha}$ from $\mathcal{I}(t^n)$.

In conclusion, the computation of the discontinuity time and points for the activation uses a lot of data that are already known from the previous time step t^n, and the effort is essentially based on the resolution of a linear system which is in fact similar to the systems arising in the interior-point method. Hence the computation cost is not expensive.

3.3.2 Deactivation of an inequality constraint

Although, the computation of the discontinuity time and points is done with a similar method as for the activation case, it requires more attention. Let us assume that the deactivation of the phase $\bar{\alpha} \in \mathcal{A}(t^n)$ is detected in the time interval $[t^n, t^{n+1}]$, that is the point $\left(\mathbf{x}^{n+1}_{\bar{\alpha}}, g\left(\mathbf{x}^{n+1}_{\bar{\alpha}}\right)\right)$ is situated at a negative distance to the tangent plane defined by $\boldsymbol{\lambda}^{n+1}$.

As for the activation case, the event of the discontinuity can be expressed as an equation in the fractional time step τ. An event is detected if a point is situated at a negative distance to the supporting tangent plane and no event occurs if all the points on the graph of the energy are situated above the supporting tangent plane. In Section 1.7 the deactivation of a constraint is geometrically interpreted as the time when the supporting tangent plane has a new contact point with the energy graph without violating the Gibbs tangent plane criterion. This new contact point is then at a zero distance from the tangent plane.

Thus we are looking for the fractional time step $\tau \in [0, h]$ such that at time $t^n + \tau$ the supporting tangent plane has a new contact point with g in the convex region relative to the index $\bar{\alpha}$. Equivalently let us find τ satisfying

$$F(t^n + \tau) \;=\; g(\mathbf{x}_{\bar{\alpha}}(t^n + \tau)) + \boldsymbol{\lambda}(t^n + \tau)^T \mathbf{x}_{\bar{\alpha}}(t^n + \tau) \;=\; 0, \tag{3.3.13}$$

where $\mathbf{x}_{\bar{\alpha}}$ is solution of the minimization problem

$$\begin{aligned} \mathbf{x}_{\bar{\alpha}}(t^n + \tau) \;&=\; \arg\min_{\mathbf{x}} g(\mathbf{x}) + \boldsymbol{\lambda}(t^n + \tau)^T \mathbf{x} \\ &\quad \text{s.t.} \quad \mathbf{e}^T \mathbf{x} - 1 \;=\; 0. \end{aligned} \tag{3.3.14}$$

The expression for F resumes the definition (3.2.2) of the distance d between the supporting tangent plane and the graph of g. However unlike in (3.2.2) the variable $\boldsymbol{\lambda}$ is no more an input, but an unknown. Furthermore in (3.3.13) the variable $\mathbf{x}_{\bar{\alpha}}$ is an unknown that depends on $\boldsymbol{\lambda}$ since it is the point such that $(\mathbf{x}_{\bar{\alpha}}, g(\mathbf{x}_{\bar{\alpha}}))$ is situated at minimal distance to the tangent plane defined by $\boldsymbol{\lambda}$. In fact in (3.3.13) one should write $\mathbf{x}_{\bar{\alpha}}(\boldsymbol{\lambda}(t^n + \tau))$, but for avoiding heavy notations, only the dependence on τ is mentioned for $\mathbf{x}_{\bar{\alpha}}$. Hence two optimization problems are related to F. The first one is the original optimization problem with $\mathbf{b}(t^n + \tau)$ in order to get $\boldsymbol{\lambda}(t^n + \tau)$. The second one is the minimization problem (3.3.14) that defines the variable $\mathbf{x}_{\bar{\alpha}}(t^n + \tau)$.

The computation of the discontinuity time τ follows the strategy for the activation case. Thus let us consider the Taylor series expansion of F

$$F(t^n + \tau) \;=\; F(t^n) + \tau \frac{\mathrm{d}}{\mathrm{d}t} F(t^n) + \frac{\tau^2}{2} \frac{d^2}{dt^2} F(t^n) + \mathcal{O}(\tau^3).$$

Then the extrapolation polynomial is expressed as

$$P^d_{\bar{\alpha}}(\tau) \;=\; F(t^n) + \tau \frac{\mathrm{d}}{\mathrm{d}t} F(t^n + \frac{\tau}{2}).$$

The expression of the derivative is computed with the chain rule. By the definition of F one can write successively

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} F\left(t^n + \frac{\tau}{2}\right) \;&=\; \frac{\mathrm{d}}{\mathrm{d}t} g\left(\mathbf{x}_{\bar{\alpha}}\left(t^n + \frac{\tau}{2}\right)\right) + \frac{\mathrm{d}}{\mathrm{d}t}\left(\boldsymbol{\lambda}\left(t^n + \frac{\tau}{2}\right)^T \mathbf{x}_{\bar{\alpha}}\left(t^n + \frac{\tau}{2}\right)\right), \\ &=\; \boldsymbol{\nabla} g\left(\mathbf{x}_{\bar{\alpha}}\left(t^n + \frac{\tau}{2}\right)\right)^T \frac{\mathrm{d}}{\mathrm{d}t} \mathbf{x}_{\bar{\alpha}}\left(t^n + \frac{\tau}{2}\right) + \frac{\mathrm{d}}{\mathrm{d}t} \boldsymbol{\lambda}\left(t^n + \frac{\tau}{2}\right)^T \mathbf{x}_{\bar{\alpha}}\left(t^n + \frac{\tau}{2}\right) \\ &\quad + \boldsymbol{\lambda}\left(t^n + \frac{\tau}{2}\right)^T \frac{\mathrm{d}}{\mathrm{d}t} \mathbf{x}_{\bar{\alpha}}\left(t^n + \frac{\tau}{2}\right), \\ &=\; \left[\boldsymbol{\nabla} g\left(\mathbf{x}_{\bar{\alpha}}\left(t^n + \frac{\tau}{2}\right)\right) + \boldsymbol{\lambda}\left(t^n + \frac{\tau}{2}\right)\right]^T \frac{\mathrm{d}}{\mathrm{d}t} \mathbf{x}_{\bar{\alpha}}\left(t^n + \frac{\tau}{2}\right) \\ &\quad + \frac{\mathrm{d}}{\mathrm{d}t} \boldsymbol{\lambda}\left(t^n + \frac{\tau}{2}\right)^T \mathbf{x}_{\bar{\alpha}}\left(t^n + \frac{\tau}{2}\right). \end{aligned}$$

Now with the chain rule the time derivative of $\mathbf{x}_{\bar{\alpha}}$ is transformed in

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x}_{\bar{\alpha}}\left(t^n + \frac{\tau}{2}\right) = \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\lambda}}\mathbf{x}_{\bar{\alpha}}\left(t^n + \frac{\tau}{2}\right)\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\lambda}\left(t^n + \frac{\tau}{2}\right),$$

and the derivative of F at $t^n + \frac{\tau}{2}$ becomes

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}F\left(t^n + \frac{\tau}{2}\right) = {} & \left[\boldsymbol{\nabla} g\left(\mathbf{x}_{\bar{\alpha}}\left(t^n + \frac{\tau}{2}\right)\right) + \boldsymbol{\lambda}\left(t^n + \frac{\tau}{2}\right)\right]^T \frac{d}{d\boldsymbol{\lambda}}\mathbf{x}_{\bar{\alpha}}\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\lambda}\left(t^n + \frac{\tau}{2}\right) \\ & + \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\lambda}\left(t^n + \frac{\tau}{2}\right)^T \mathbf{x}_{\bar{\alpha}}\left(t^n + \frac{\tau}{2}\right).\end{aligned}$$

The difficulty consists therefore in the computation of the matrix $\frac{d}{d\boldsymbol{\lambda}}\mathbf{x}_{\bar{\alpha}}\left(t^n + \frac{\tau}{2}\right)$ and the vector $\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\lambda}\left(t^n + \frac{\tau}{2}\right)$. The technique is similar to the activation case with a sensitivity analysis on a perturbed optimization problem.

Let us begin with the simplest case: $\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\lambda}\left(t^n + \frac{\tau}{2}\right)$. Using the chain rule again, this expression becomes

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\lambda}\left(t^n + \frac{\tau}{2}\right) = \sum_{i=1}^{s}\frac{\partial\boldsymbol{\lambda}}{\partial b_i}\left(\mathbf{b}\left(t^n + \frac{\tau}{2}\right)\right)\frac{\mathrm{d}}{\mathrm{d}t}b_i\left(t^n + \frac{\tau}{2}\right).$$

The approximation of $\frac{\mathrm{d}}{\mathrm{d}t}b_i\left(t^n + \frac{\tau}{2}\right)$, $i = 1, \ldots, s$, is given by the straight line passing through $\left(t^{n-1}, \mathbf{j}(\mathbf{b}^{n-1}, \mathbf{x}_{\alpha}^{\mathcal{I},n-1}, R^{n-1})\right)$ and $\left(t^n, \mathbf{j}(\mathbf{b}^n, \mathbf{x}_{\alpha}^{\mathcal{I},n}, R^n)\right)$ as for the activation case. The other derivative $\frac{\partial\boldsymbol{\lambda}}{\partial b_i}\left(\mathbf{b}\left(t^n + \frac{\tau}{2}\right)\right)$, $i = 1, \ldots, s$, is in fact directly obtained by the resolution of the linear system (3.3.10) defined for the activation case. Hence the derivative $\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\lambda}\left(t^n + \frac{\tau}{2}\right)$ is estimated as in the previous section.

Concerning the matrix $\frac{d}{d\boldsymbol{\lambda}}\mathbf{x}_{\bar{\alpha}}\left(t^n + \frac{\tau}{2}\right)$, let us use the sensitivity analysis again. Because of the dependence of $\mathbf{x}_{\bar{\alpha}}$ on $\boldsymbol{\lambda}$ and the definition of $\mathbf{x}_{\bar{\alpha}}$, let us consider the minimization problem of the distance:

$$\begin{aligned} \min_{\mathbf{x}_{\bar{\alpha}}} \quad & g(\mathbf{x}_{\bar{\alpha}}) + \boldsymbol{\lambda}^T\mathbf{x}_{\bar{\alpha}} \\ \text{s.t.} \quad & \mathbf{e}^T\mathbf{x}_{\bar{\alpha}} - 1 = 0, \end{aligned}$$

considering $\boldsymbol{\lambda}$ as an input and not a variable.

The matrix $\frac{d}{d\boldsymbol{\lambda}}\mathbf{x}_{\bar{\alpha}}$ is defined by

$$\frac{d}{d\boldsymbol{\lambda}}\mathbf{x}_{\bar{\alpha}} = \left(\frac{\mathrm{d}\mathbf{x}_{\bar{\alpha}}}{\mathrm{d}\lambda_1}, \ldots, \frac{\mathrm{d}\mathbf{x}_{\bar{\alpha}}}{\mathrm{d}\lambda_s}\right).$$

Then let us approximate each $\frac{\mathrm{d}\mathbf{x}_{\bar{\alpha}}}{\mathrm{d}\lambda_i}$, $i = 1, \ldots, s$ with the sensitivity analysis method. For a general index i the perturbed minimization problem is defined by

$$\begin{aligned} \min_{\mathbf{x}_{\bar{\alpha}}} \quad & g(\mathbf{x}_{\bar{\alpha}}) + \boldsymbol{\lambda}^T\mathbf{x}_{\bar{\alpha}} - \mathrm{d}\lambda_i\mathbf{e}_i^T\mathbf{x}_{\bar{\alpha}} \\ \text{s.t.} \quad & \mathbf{e}^T\mathbf{x}_{\bar{\alpha}} - 1 = 0. \end{aligned}$$

In comparison to (3.3.6), the vector $\mathbf{a}^T = (\mathrm{d}\lambda_i, \mathbf{0})$ and the function ψ_i associated to the perturbation is the function defined by $-\mathbf{e}_i^T \mathbf{x}_{\bar{\alpha}}$.

The objective function and the equality constraint are twice-differentiable functions of $\mathbf{x}_{\bar{\alpha}}$. Furthermore $\mathbf{x}_{\bar{\alpha}}^n$ is a KKT point of the original minimization problem by construction. Thus the Theorem 3.3.2 can be applied and one can consider the set of equations issued from the derivative of the Lagrangian with respect to $\mathbf{x}_{\bar{\alpha}}$ and the equality constraint

$$\begin{aligned} \boldsymbol{\nabla} g(\mathbf{x}_{\bar{\alpha}}) + \boldsymbol{\lambda} - \mathrm{d}\lambda_i \mathbf{e}_i + \zeta_{\bar{\alpha}} \mathbf{e} &= \mathbf{0}, \\ \mathbf{e}^T \mathbf{x}_{\bar{\alpha}} - 1 &= 0, \end{aligned}$$

where $\zeta_{\bar{\alpha}} \in \mathbb{R}$ is the dual variable associated to the equality constraint.

The variable $\boldsymbol{\lambda}$ is considered as a given constant. Treating this set of equations as a function of $\mathrm{d}\lambda_i$, we differentiate it with respect to the independent variable $\mathrm{d}\lambda_i$ and evaluate it at $\mathbf{a} = \mathbf{0}$. It yields the following system:

$$\begin{aligned} \boldsymbol{\nabla}^2 g(\mathbf{x}_{\bar{\alpha}}^n) \frac{\mathrm{d}\mathbf{x}_{\bar{\alpha}}}{\mathrm{d}\lambda_i}(\mathbf{0}) \;-\; \mathbf{e}_i \;+\; \frac{\mathrm{d}\zeta_{\bar{\alpha}}}{\mathrm{d}\lambda_i}(\mathbf{0})\, \mathbf{e} &= \mathbf{0}, \\ \mathbf{e}^T \frac{\mathrm{d}\mathbf{x}_{\bar{\alpha}}}{\mathrm{d}\lambda_i}(\mathbf{0}) &= 0. \end{aligned}$$

Hence for all $i = 1, \ldots, s$ the derivative $\frac{\mathrm{d}\mathbf{x}_{\bar{\alpha}}}{\mathrm{d}\lambda_i}$ is obtained by solving the following linear system

$$\begin{pmatrix} \boldsymbol{\nabla}^2 g(\mathbf{x}_{\bar{\alpha}}^n) & \mathbf{e} \\ \mathbf{e}^T & 0 \end{pmatrix} \begin{pmatrix} \frac{\mathrm{d}\mathbf{x}_{\bar{\alpha}}}{\mathrm{d}\lambda_i}(\mathbf{0}) \\ \frac{\mathrm{d}\zeta_{\bar{\alpha}}}{\mathrm{d}\lambda_i}(\mathbf{0}) \end{pmatrix} = \begin{pmatrix} \mathbf{e}_i \\ 0 \end{pmatrix}. \tag{3.3.15}$$

Finally by combining all these previous results the derivative $\frac{\mathrm{d}}{\mathrm{d}t} F\left(t^n + \frac{\tau}{2}\right)$ is approximated by

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} F\left(t^n + \frac{\tau}{2}\right) &\approx \left[(\boldsymbol{\nabla} g(\mathbf{x}_{\bar{\alpha}}^n) + \boldsymbol{\lambda}^n)^T \frac{\mathrm{d}\mathbf{x}_{\bar{\alpha}}}{\mathrm{d}\boldsymbol{\lambda}}(\mathbf{0}) + \mathbf{x}_{\bar{\alpha}}^{n,T} \right] \frac{\mathrm{d}\boldsymbol{\lambda}}{\mathrm{d}t}\left(t^n + \frac{\tau}{2}\right) \\ &\approx \left[(\boldsymbol{\nabla} g(\mathbf{x}_{\bar{\alpha}}^n) + \boldsymbol{\lambda}^n)^T \frac{\mathrm{d}\mathbf{x}_{\bar{\alpha}}}{\mathrm{d}\boldsymbol{\lambda}}(\mathbf{0}) + \mathbf{x}_{\bar{\alpha}}^{n,T} \right] \frac{\mathrm{d}\boldsymbol{\lambda}}{\mathrm{d}\mathbf{b}}(\mathbf{0}) \left(\mathbf{j}^n + \frac{\tau}{2h}\left(\mathbf{j}^n - \mathbf{j}^{n-1}\right)\right), \end{aligned}$$

with $\mathbf{j}^n = \mathbf{j}(\mathbf{b}^n, \mathbf{x}_\alpha^{\mathcal{I},n}, R^n)$ and $\mathbf{j}^{n-1} = \mathbf{j}(\mathbf{b}^{n-1}, \mathbf{x}_\alpha^{\mathcal{I},n-1}, R^{n-1})$.
Let us define both scalars

$$\begin{aligned} A^n &= \left[(\boldsymbol{\nabla} g(\mathbf{x}_{\bar{\alpha}}^n) + \boldsymbol{\lambda}^n)^T \frac{\mathrm{d}\mathbf{x}_{\bar{\alpha}}}{\mathrm{d}\boldsymbol{\lambda}}(\mathbf{0}) + \mathbf{x}_{\bar{\alpha}}^{n,T} \right] \frac{\mathrm{d}\boldsymbol{\lambda}}{\mathrm{d}\mathbf{b}}(\mathbf{0}) \mathbf{j}^n, \\ B^n &= \left[(\boldsymbol{\nabla} g(\mathbf{x}_{\bar{\alpha}}^n) + \boldsymbol{\lambda}^n)^T \frac{\mathrm{d}\mathbf{x}_{\bar{\alpha}}}{\mathrm{d}\boldsymbol{\lambda}}(\mathbf{0}) + \mathbf{x}_{\bar{\alpha}}^{n,T} \right] \frac{\mathrm{d}\boldsymbol{\lambda}}{\mathrm{d}\mathbf{b}}(\mathbf{0}) (\mathbf{j}^n - \mathbf{j}^{n-1}). \end{aligned}$$

The approximation of the derivative $\frac{\mathrm{d}}{\mathrm{d}t} F\left(t^n + \frac{\tau}{2}\right)$ becomes

$$\frac{\mathrm{d}}{\mathrm{d}t} F\left(t^n + \frac{\tau}{2}\right) \approx A^n + \frac{\tau}{2h} B^n,$$

and the approximation of the associated extrapolation polynomial reads

$$P_{\bar{\alpha}}^{d,n}(\tau) \;=\; F^n \,+\, \tau A^n \,+\, \frac{\tau^2}{2h} B^n, \tag{3.3.16}$$

with $F^n = g(\mathbf{x}_{\bar{\alpha}}^n) + \boldsymbol{\lambda}^{n,T}\mathbf{x}_{\bar{\alpha}}^n \approx F(t^n)$.
As for the activation case the extrapolation polynomial is a second order polynomial in τ and the determination of τ consists in solving the second order equation in τ

$$P_{\bar{\alpha}}^{d,n}(\tau) \;=\; 0.$$

For a time step h sufficiently small, there exists a root τ^* of $P_{\bar{\alpha}}^{d,n}(\tau) = 0$ that is in $[0, h]$. The approximated discontinuity time is then given by $t_h^* \;=\; t^n + \tau^*$.

Once the fractional time step τ^* is computed, the two-step Adams-Bashforth method with variable time step is used to approximate $\mathbf{b}$ at $t^{n+1} := t^n + \tau^*$, namely

$$\mathbf{b}^{n+1} \;=\; \mathbf{b}^n \,+\, \tau^* \left[\left(1 + \frac{\tau^*}{2h}\right) \mathbf{j}^n - \frac{\tau^*}{2h}\mathbf{j}^{n-1} \right].$$

All the terms in the above relation are known. Then the computation of $\mathbf{b}^{n+1}$ only requires addition and multiplication operations.

Before restarting the simulation, the values of the variables $\mathbf{x}_\alpha, y_\alpha, \zeta_\alpha, \theta_\alpha$, for $\alpha = 1, \ldots, p$, and $\boldsymbol{\lambda}$ need to be approximated at t^{n+1}. As for the activation case their approximation is done through the optimization problem defined for $\mathbf{b}^{n+1}$. However the resolution of the optimization problem must be executed very carefully in order not to remove the new inactive constraint $\bar{\alpha}$ from the set $\mathcal{I}$. At the discontinuity time the new inactive constraints $y_{\bar{\alpha}}$ is in fact equal to 0 and its value is numerically forced to $y_{\bar{\alpha}}^{n+1} = \epsilon_y$. If the warm-start strategy as defined in Section 3.1.2 is applied, the variable $y_{\bar{\alpha}}$ is likely removed from $\mathcal{I}$ because of its small value close to the threshold ϵ_y. If the cold-start strategy is applied, all the informations from the time step t^n are lost even if the variables $\mathbf{x}_\alpha^{n,\mathcal{I}}$ are generally good approximations of $\mathbf{x}_\alpha^{n+1,\mathcal{I}}$ and that the resolution of the optimization problem may be difficult since it is defined at a discontinuity point.

Such a situation is represented in Figure 3.3.2. On this example the constraints y_1 and y_2 are inactive at time t^n and the corresponding phase simplex is the segment defined by $\mathrm{conv}(\mathbf{x}_1^n, \mathbf{x}_2^n) = [\mathbf{x}_1^n, \mathbf{x}_2^n]$. At the discontinuity time t^{n+1} the constraint y_3 is activated and the new phase simplex is the triangle whose vertices are the points $\mathbf{x}_1^{n+1}, \mathbf{x}_2^{n+1}$ and $\mathbf{x}_3^{n+1}$. The discontinuity point $\mathbf{b}^{n+1}$ belongs to the triangle. However since $\mathbf{b}^{n+1}$ is situated at the transition between both phase simplices, $\mathbf{b}^{n+1}$ belongs to the edge $[\mathbf{x}_1^{n+1}, \mathbf{x}_2^{n+1}]$ and $y_3^{n+1} = 0$. Moreover the points $\mathbf{x}_1^n$ and $\mathbf{x}_1^{n+1}$, and $\mathbf{x}_2^n$ and $\mathbf{x}_2^{n+1}$, are respectively close to eachother.

The warm-start strategy as defined in Section 3.1.2 will likely lead to the activation of the constraint $y_{\bar{\alpha}}$ and the cold-start strategy may also activate the constraint $y_{\bar{\alpha}}$. Then a new warm-start strategy has to be determined in order to ensure the belonging of the constraint $y_{\bar{\alpha}}$ in the set of inactive constraints $\mathcal{I}$. This new strategy takes the points

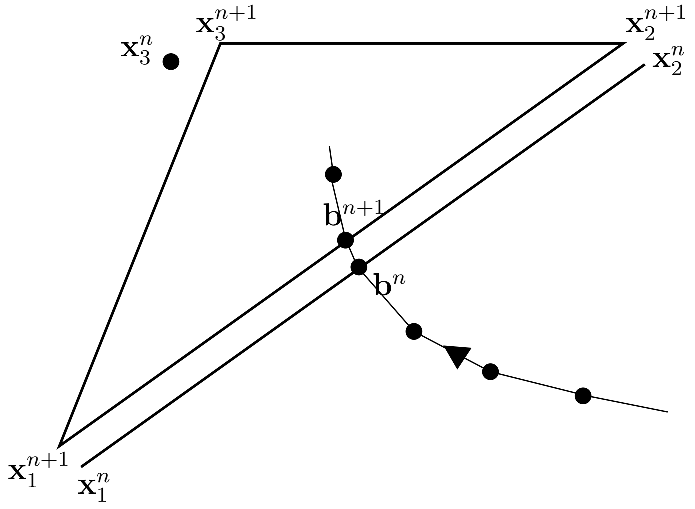


Figure 3.4: Deactivation of the constraint y_3 at t^{n+1} when the constraints y_1 and y_2 are already inactive at t^n.

$\mathbf{x}_\alpha^n$ as initialization of $\mathbf{x}_\alpha^{n+1}$ for $\alpha \in \mathcal{I}(t^n)$ in the optimization problem. Since $\mathbf{x}_{\bar{\alpha}}^n$ is the point at minimal distance from the supporting tangent plane defined by $\boldsymbol{\lambda}^n$ at t^n, this point constitutes a potentially good initialization of $\mathbf{x}_{\bar{\alpha}}^{n+1}$ (the point $\mathbf{x}_3^n$ in Figure 3.3.2 is situated in the neighborhood of $\mathbf{x}_3^{n+1}$). Furthermore in order not to favour any inactive constraints in $\mathcal{I}$ and to ensure the remain of $\bar{\alpha}$ in $\mathcal{I}$ the initialization of the variable y_α, $\alpha \in \mathcal{I}$ in the optimization problem is given by

$$y_\alpha = \frac{1}{p^{\mathcal{I}}}, \ \forall \alpha \in \mathcal{I}.$$

For the example in Figure 3.3.2 this initialization reads $y_\alpha = \frac{1}{3}$ with $\alpha \in \{1, 2, 3\}$. Then the remaining variables, namely $\zeta_\alpha, \theta_\alpha$ for $\alpha \in \mathcal{I}$ and $\boldsymbol{\lambda}$ are initialized as with the cold-start strategy but with $p^{\mathcal{I}}$ instead of p.

With this new warm-start strategy for the initialization of the resolution of the optimization problem, the approximations of the variables $\mathbf{x}_\alpha, y_\alpha, \theta_\alpha$ and ζ_α for $\alpha = 1, \dots, p$, and $\boldsymbol{\lambda}$ are obtained at the discontinuity time t^{n+1}. Then the resolution of the system (3.1.1) is restarted until the next event is detected.

Remark 3.3.2. *If the inactive constraint $\bar{\alpha}$ is removed from $\mathcal{I}$ even if the new warm-start strategy is applied, the variable $\mathbf{x}_{\bar{\alpha}}^{n+1}$ can still be determined since $\mathbf{x}_{\bar{\alpha}}^{n+1}$ defines the point situated in the convex area $\Delta'_{s,\bar{\alpha}}$ that minimizes the distance to the supporting tangent plane defined at t^{n+1}. Then the Algorithm 3.2.1 applied to $\bar{\alpha}$ leads to $\mathbf{x}_{\bar{\alpha}}^{n+1}$ and the set of inactive constraints is again increased by the index $\bar{\alpha}$ before the restart of the numerical simulation.*

3.4 A priori error estimation

Error estimates for the approximations of the discontinuity time and points are difficult to obtain in the case of the system (3.1.1) since one has to handle optimization and differential features. However if the optimization problem is supposed to be exact, the system (3.1.1) is reduced to a Cauchy's problem and a priori error estimates may be established following the theory developed for nonlinear problems [22, 91].

Let us assume that the optimization algorithm gives an exact solution, meaning that the primal variables y_α and $\mathbf{x}_\alpha$ for $\alpha = 1, \ldots, p$ are computed from $\mathbf{b}$ exactly and no error is made about their value. Moreover the variable R in (3.1.1) is already exact and fully depends on $\mathbf{b}$. Thus the error estimate for the approximations of the discontinuity time and points is solely based on the differential part present in the system (3.1.1), and this system can be reduced to the following Cauchy's problem for the establishment of error estimates: find $\mathbf{b}(t)$ satisfying

$$\left\{ \begin{array}{rcl} \dfrac{\mathrm{d}}{\mathrm{d}t}\mathbf{b}(t) & = & \mathbf{f}(\mathbf{b}(t)), \\ \\ \mathbf{b}(0) & = & \mathbf{b}_0. \end{array} \right. \tag{3.4.1}$$

As long as the number of inactive inequality constraints remains fixed, the solution of (3.1.1) is continuously differentiable. Once an event occurs, the variable $\mathbf{x}_\alpha$ and y_α, $\alpha \in \mathcal{I}$ lose their regularity, and consequently the flux $\mathbf{j}$ loses also its regularity. However one can extend the flux defined with the fixed number of inactive constraints, in the neighborhood of the discontinuity point. In the Cauchy's problem, the function $\mathbf{f}$ illustrates the flux function defined with a fixed number of inactive constraints. Hence the function $\mathbf{f}$ can be extended in the neighborhood of the discontinuity point, and assumed Lipschitz with the constant L and bounded. Moreover the Cauchy's problem is supposed to admit a solution. This solution is differentiable and can also be extended in the neighborhood of the discontinuity point. Let us denote by $\mathbf{b}^n$ the iterates obtained from the discretization of (3.4.1) with the Crank-Nicolson method. The last considered hypothesis is the positiveness of $\mathbf{b}(t)$, $\forall t > 0$ and of $\mathbf{b}^n$, $\forall n = 0, \ldots, m$.

The time interval is $[0, T]$ with T sufficiently large such that an activation or a deactivation occurs before T. Let $t^0, t^1, ..., t^m$ denote the discretization of $[0, T]$ with $t^0 = 0$, $t^m = T$, and $h^k = t^k - t^{k-1}$, the length of each subinterval, for $k = 1, \ldots, m$. Finally let us introduce the variable $h = \max_{1 \le k \le m} h^k$. Unlike in Section 3.1 the time step is no longer fixed. Indeed the time step is fixed except for the fractional time step needed to reach the discontinuity.

In the forthcoming theory the used norm is the infinity-norm that is simply denoted by $\| \cdot \|$. Moreover all the notations for the constants that are defined in the following lemmas and theorems are local notations. For reasons of simplification the case $s = 3$ is first studied. A generalization of the theoretical results to s dimensions is then discussed.

3.4.1 Theoretical results in 3 dimensions

Let us consider first the case $s = 3$. In this case the composition-vector $\mathbf{b}$ is made of 3 different chemical components and the trajectory of $\mathbf{b}$ is represented on a phase diagram of dimension $r = s - 1 = 2$ (see Section 1.7). A general representation of the ternary phase diagram is depicted in Figure 3.5.

In the system (3.1.1), $\mathbf{b}$ loses its regularity when an inequality constraint activates or deactivates. On a geometrical point of view this loss of regularity is characterized by a change in the dimension of the phase simplex associated to $\mathbf{b}$ as described in Section 1.7. In Figure 3.5 the trajectory of $\mathbf{b}$ on the phase diagram for a ternary system is presented. The initial composition-vector $\mathbf{b}_0$ is situated in the area 1 which means that $\mathbf{b}_0$ is a single-phase point. A second inequality constraint is deactivated at the first time step. The phase simplex becomes a segment represented by a dotted blue line and its dimension goes from $p^{\mathcal{I}} = 1$ to $p^{\mathcal{I}} = 2$. Then the last inequality constraint is deactivated and the corresponding phase simplex is a triangle.

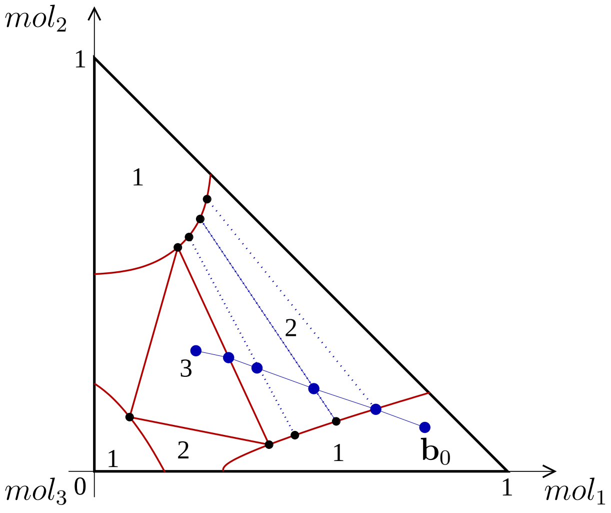


Figure 3.5: Phase boundaries on the phase diagram of a ternary system.

One can observe on the one hand in Figure 3.5 that the single-phase points which constitute the successive phase simplices when $p^{\mathcal{I}} \geq 2$ follow the frontiers that separate the different areas of the phase diagrams. The frontiers are the solid red lines in Figure 3.5. On the other hand both discontinuity points are located on the frontiers too. These two points are also called *phase transition points* because the dimension of the phase simplex changes at these points. In fact all the points on these solid red lines are phase transition points since they are all situated at the boundary between 2 areas where the dimension of the phase simplex differs. Thus for the system (3.1.1) $\mathbf{b}$ loses its regularity when its trajectory crosses a frontier of a phase diagram.

For the theoretical study of this section, let us assume that only one event occurs in the time interval $[0, T]$. In other words let us assume that $\mathbf{b}$ crosses only one frontier of the phase diagram and only once in the time interval $[0, T]$. Let us denote by t^* the exact time when $\mathbf{b}$ crosses the frontier and assume that there exists $n = n(h) \leq m - 1$ such that $t^* \in [t^n, t^{n+1}[$. Note that when h becomes smaller, the index $n = n(h)$ becomes larger. For the system (3.4.1) $\mathbf{b}$ does not lose regularity, but is extended in a $\mathcal{C}^2$ manner in the neighborhood of the discontinuity point.

In Figure 3.5 the frontiers are either segments or curves. However in the neighborhood of the event this frontier may be locally approximated by a straight line. Thus for this study let us consider the boundary as a segment $[\mathbf{C}, \mathbf{D}]$ defined by

$$[\mathbf{C}, \mathbf{D}] := \{\mathbf{z} \in \mathbb{R}^2 \,|\, \exists \ell \in [0, 1] \text{ s. t. } \mathbf{z} = \mathbf{OC} + \ell\, \mathbf{v}\},$$

with $\mathbf{v} = \mathbf{CD}$, the direction vector of the segment. Note that the frontier really is a segment when going from 2 to 3 inactive constraints.

Let $\mathbf{b}_h$ denotes the linear spline interpolation of $\mathbf{b}^n$ stemmed from the Crank-Nicolson method. The time at which $\mathbf{b}_h$ crosses the segment $[\mathbf{C}, \mathbf{D}]$ is denoted by t_h^*. The purpose of this section is to estimate on the one hand the error between the exact discontinuity time t^* and the approximated one t_h^*, namely $|t^* - t_h^*|$, and on the other hand the difference between the exact discontinuity point $\mathbf{b}(t^*)$ and the approximated one $\mathbf{b}_h(t_h^*)$. Let us remind that in fact the points represented on the phase diagram are normalized points belonging in $\mathbb{R}^r$ with $r = s - 1$. Therefore if the error between the points that cross the frontier is wanted, one should to compare their normalized value projected in $\mathbb{R}^2$. Let us denote by $\mathbf{d}$ the normalized projected points as in section 1.7. The error estimate is then given by $\|\mathbf{d}(t^*) - \mathbf{d}_h(t_h^*)\|$ with $\mathbf{d}(t) = \frac{1}{\mathbf{e}^T \mathbf{b}(t)} P\mathbf{b}(t)$ and $\mathbf{d}_h(t) = \frac{1}{\mathbf{e}^T \mathbf{b}_h(t)} P\mathbf{b}_h(t)$ for all $t \in [0, T]$.

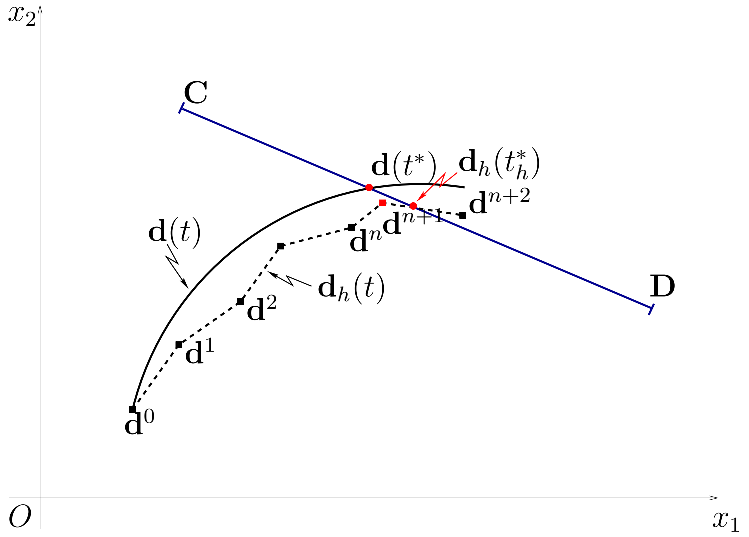


Figure 3.6: Illustration for exact and approximate trajectories.

The geometrical situation is represented in Figure 3.6. Let us define the following two functions:

$$\begin{array}{rcl} \mathbf{F} : \mathbb{R}_+ \times \mathbb{R} & \to & \mathbb{R}^2 \\ \mathbf{w}^T = (t, \ell) & \mapsto & \mathbf{F}(\mathbf{w}) = \mathbf{d}(t) - \mathbf{OC} - \ell\,\mathbf{v}, \end{array}$$

and

$$\begin{array}{rcl} \mathbf{F_h} : \mathbb{R}_+ \times \mathbb{R} & \to & \mathbb{R}^2 \\ \mathbf{w}^T = (t, \ell) & \mapsto & \mathbf{F_h}(\mathbf{w}) = \mathbf{d}_h(t) - \mathbf{OC} - \ell\,\mathbf{v}. \end{array}$$

The root of $\mathbf{F}$, respectively $\mathbf{F_h}$, is the crossing point between the segment and the trajectory $\mathbf{d}$, respectively $\mathbf{d}_h$. Hence the desired error estimates are given by the error between the zero of each function. Let us denote $\mathbf{u}^T = (t^*, \ell^*)$ the zero of $\mathbf{F}$ and $\mathbf{u_h}^T = (t_h^*, \ell_h^*)$ the zero of $\mathbf{F_h}$. The establishment of the error estimate follows nonlinear techniques presented in [22, 91]. Let us first give a property for the Jacobian matrix of $\mathbf{F}$ at $\mathbf{u}$.

Lemma 3.4.1. *If $\frac{d}{dt}\mathbf{d}(t^*)$ is not parallel to $\mathbf{v}$, then $DF(\mathbf{u})$ is regular.*

Proof. Let us denote by v_1, v_2 the components of the vector $\mathbf{v}$ and $\mathbf{n}$ the normal vector of the segment $[\mathbf{C}, \mathbf{D}]$ defined by:

$$\mathbf{n} = \begin{pmatrix} -v_2 \\ v_1 \end{pmatrix}.$$

The expression of $DF(\mathbf{u})$ is then given by:

$$DF(\mathbf{u}) = DF(t^*, \ell^*) = \begin{pmatrix} \frac{\mathrm{d}}{\mathrm{d}t} d_1(t^*) & -v_1 \\ \frac{\mathrm{d}}{\mathrm{d}t} d_2(t^*) & -v_2 \end{pmatrix}.$$

So the determinant of $DF(\mathbf{u})$ is equal to:

$$\det(DF(\mathbf{u})) \;=\; -v_2 \frac{\mathrm{d}}{\mathrm{d}t} d_1(t^*) + v_1 \frac{\mathrm{d}}{\mathrm{d}t} d_2(t^*) \;=\; \mathbf{n} \cdot \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{d}(t^*).$$

Since $\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{d}(t^*)$ is not parallel to $\mathbf{v}$ by hypothesis, the scalar product $\mathbf{n} \cdot \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{d}(t^*)$ is not equal to zero. This implies the regularity of $DF(\mathbf{u})$. ☐

In the following theorem the existence and uniqueness of a solution of $\mathbf{F_h}$ is proved and an a priori error estimate is established.

Theorem 3.4.2. *Assume that the functions $\mathbf{F}$ and $\mathbf{F}_h$ admit zeros in $(0, T)$, denoted by $\mathbf{u} = (t^*, \ell^*)$ and $\mathbf{u_h} = (t_h^*, \ell_h^*)$ respectively. Furthermore let us assume that $\mathbf{d}$ can be extended in a $\mathcal{C}^2$ manner in the neighborhood of the discontinuity point, and $\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{d}(t^*)$ is not parallel to $\mathbf{v}$. Then:*

(i) the following limit holds

$$\lim_{h \to 0} \|\mathbf{F_h}(\mathbf{u})\| \;=\; 0;$$

(ii) there exists $\bar{h} > 0$ *and a ball centered in* $\mathbf{u}$ *and with radius* $\delta > 0$ *denoted by* $\mathcal{B}(\mathbf{u}, \delta) \subset \mathbb{R}^2$ *such that* $\forall h < \bar{h}$ *there exists only one* $\mathbf{u_h} \in \mathcal{B}(\mathbf{u}, \delta)$ *satisfying* $\mathbf{F_h}(\mathbf{u_h}) = \mathbf{0}$. *Moreover there exists a constant* $\bar{C}$ *independent of* $h \leq \bar{h}$ *such that the following a priori error estimates holds*

$$\|\mathbf{u} - \mathbf{u_h}\| \;\leq\; \bar{C}\,\|\mathbf{F_h}(\mathbf{u})\|. \tag{3.4.2}$$

Proof. Since the Crank-Nicolson and Adams-Bashforth methods are consistent, one has immediately

$$\lim_{h\to 0} \mathbf{d}_h(t^*) \;=\; \mathbf{d}(t^*).$$

Then by writing

$$\|\mathbf{F_h}(\mathbf{u})\| \;=\; \|\mathbf{F_h}(\mathbf{u}) - \mathbf{F}(\mathbf{u})\| \;=\; \|\mathbf{d}_h(t^*) - \mathbf{d}(t^*)\|,$$

it follows

$$\lim_{h\to 0} \|\mathbf{F_h}(\mathbf{u})\| \;=\; 0$$

and the first part of the theorem is proved.

For the second part of the theorem let us define the operator $G : \mathbb{R}^2 \to \mathbb{R}^2$ as

$$G(\mathbf{w}) = \mathbf{w} - DF(\mathbf{u})^{-1}\mathbf{F_h}(\mathbf{w}).$$

By the Lemma 3.4.1, G is well defined. Moreover if $\mathbf{u_h}$ is a fixed point of G, then $\mathbf{u_h}$ is a zero of $\mathbf{F_h}$, i.e. $\mathbf{F_h}(\mathbf{u_h}) = 0$.
Let $\delta > 0$ and $\mathbf{w} = (t_w, \ell_w)^T$ be an element of the closed ball $\mathcal{B}(\mathbf{u}, \delta)$. Without loss of generality let us assume that $t_w < t^*$. We have:

$$\begin{aligned} G(\mathbf{w}) - G(\mathbf{u}) &= \mathbf{w} - DF(\mathbf{u})^{-1}\mathbf{F_h}(\mathbf{w}) - \mathbf{u} + DF(\mathbf{u})^{-1}\mathbf{F_h}(\mathbf{u}) \\ &= DF(\mathbf{u})^{-1}\left[DF(\mathbf{u})(\mathbf{w} - \mathbf{u}) - (\mathbf{F_h}(\mathbf{w}) - \mathbf{F_h}(\mathbf{u}))\right]. \end{aligned}$$

By definition of the Jacobian one can write:

$$DF(\mathbf{u})(\mathbf{w} - \mathbf{u}) = \begin{pmatrix} \frac{\mathrm{d}}{\mathrm{d}t}d_1(t^*) & -v_1 \\ \frac{\mathrm{d}}{\mathrm{d}t}d_2(t^*) & -v_2 \end{pmatrix} \begin{pmatrix} t_w - t^* \\ \ell_w - \ell^* \end{pmatrix} = \begin{pmatrix} \frac{\mathrm{d}}{\mathrm{d}t}d_1(t^*)(t_w - t^*) - (\ell_w - \ell^*)v_1 \\ \frac{\mathrm{d}}{\mathrm{d}t}d_2(t^*)(t_w - t^*) - (\ell_w - \ell^*)v_2 \end{pmatrix}.$$

It follows

$$\begin{aligned} \|DF(\mathbf{u})(\mathbf{w} - \mathbf{u}) - \mathbf{F_h}(\mathbf{w}) + \mathbf{F_h}(\mathbf{u})\| &= \max_{1\leq i\leq 2}\left(\left|\frac{\mathrm{d}}{\mathrm{d}t}d_i(t^*)(t_w - t^*) - (\ell_w - \ell^*)v_i - d_{h,i}(t_w)\right.\right. \\ &\qquad \left.\left. + C_i + \ell_w v_i + d_{h,i}(t^*) - C_i - \ell^* v_i\right|\right) \\ &= \max_{1\leq i\leq 2}\left(\left|\frac{\mathrm{d}}{\mathrm{d}t}d_i(t^*)(t_w - t^*) - d_{h,i}(t_w) + d_{h,i}(t^*)\right|\right), \end{aligned}$$

where C_i is the i^{th} component of **OC**. Now let us study the expression $|\frac{\mathrm{d}}{\mathrm{d}t}d_i(t^*)(t_w - t^*) - d_{h,i}(t_w) + d_{h,i}(t^*)|$. Even if the function $\frac{\mathrm{d}}{\mathrm{d}t}d_{h,i}$ is piecewise continuous, the following relation holds

$$d_{h,i}(t^*) - d_{h,i}(t_w) \;=\; \int_{t_w}^{t^*} \frac{\mathrm{d}}{\mathrm{d}t}d_{h,i}(s)\,ds.$$

The expression can then be reformulated in

$$\begin{aligned}\left|\frac{\mathrm{d}}{\mathrm{d}t}d_i(t^*)(t_w - t^*) - d_{h,i}(t_w) + d_{h,i}(t^*)\right| \;&=\; \left|\int_{t_w}^{t^*}\left(\frac{\mathrm{d}}{\mathrm{d}t}d_{h,i}(s) - \frac{\mathrm{d}}{\mathrm{d}t}d_i(t^*)\right)\,ds\right|,\\ &\le\; \left|\int_{t_w}^{t^*}\left(\frac{\mathrm{d}}{\mathrm{d}t}d_{h,i}(s) - \frac{\mathrm{d}}{\mathrm{d}t}d_i(s)\right)\,ds\right|\\ &\quad+\left|\int_{t_w}^{t^*}\left(\frac{\mathrm{d}}{\mathrm{d}t}d_i(s) - \frac{\mathrm{d}}{\mathrm{d}t}d_i(t^*)\right)\,ds\right|. \quad (3.4.3)\end{aligned}$$

On one hand since the function d_i is $\mathcal{C}^2$, one can consider the Taylor series expansion

$$\left|\frac{\mathrm{d}}{\mathrm{d}t}d_i(s) - \frac{\mathrm{d}}{\mathrm{d}t}d_i(t^*)\right| \;=\; \left|\frac{d^2}{dt^2}d_i(\xi)\right|\;|s - t^*|,\;\text{ with } 0 \;<\; |\xi - t^*| \;<\; |s - t^*|.$$

The second integral in (3.4.3) is then bounded from above by

$$\left|\int_{t_w}^{t^*}\left(\frac{\mathrm{d}}{\mathrm{d}t}d_i(s) - \frac{\mathrm{d}}{\mathrm{d}t}d_i(t^*)\right)\,ds\right| \;\le\; \max_{\xi\in[t^*-\delta,t^*+\delta]\setminus\{t^*\}}\left|\frac{d^2}{dt^2}d_i(\xi)\right|\;|t^* - t_w|^2. \tag{3.4.4}$$

On the other hand for a fixed h there exists a finite number $\bar{m}(\delta)$ of subintervals $[t_i, t_{i+1}[,\ i = n, \ldots, \bar{m}(\delta)$ such that $\cup_{i=n}^{\bar{m}(\delta)}[t_i, t_{i+1}[\supset [t_w, t^*]$. On each subinterval the consistency of the numerical scheme implies the existence of a constant $\tilde{C}_i$ independent of h such that

$$\left|\frac{\mathrm{d}}{\mathrm{d}t}d_{h,i}(s) - \frac{\mathrm{d}}{\mathrm{d}t}d_i(s)\right| \;\le\; \tilde{C}_i\,h,\;\text{ where } s \in [t_i, t_{i+1}[,\ \forall i = n, \ldots, \bar{m}(\delta). \tag{3.4.5}$$

The relations (3.4.4) and (3.4.5) in the inequality (3.4.3) gives

$$\left|\frac{\mathrm{d}}{\mathrm{d}t}d_i(t^*)(t_w - t^*) - d_{h,i}(t_w) + d_{h,i}(t^*)\right| \;\le\; \left[\tilde{C}\,h + \max_{\xi\in[t^*-\delta,t^*+\delta]\setminus\{t^*\}}\left|\frac{d^2}{dt^2}d_i(\xi)\right|\;|t^* - t_w|\right]\;|t^*{-}t_w|,$$

where $\tilde{C} = \sum_{i=n}^{\bar{m}}\tilde{C}_i$. An upper bound of the term $\tilde{C}\,h$ is given by

$$\begin{aligned}\tilde{C}\,h \;&=\; \sum_{i=n}^{\bar{m}}\tilde{C}_i\,h \;\le\; (\bar{m} - n + 1)\max_{n\le i\le\bar{m}}\tilde{C}_i\,h,\\ &=\; \max_{1\le i\le\bar{m}}\tilde{C}_i\,(\bar{m} - n + 1)\,h,\\ &\le\; \max_{1\le i\le\bar{m}}\tilde{C}_i\,(\delta + 2h).\end{aligned}$$

Therefore there exists $\bar{h}_1 > 0$ and let us choose $\delta > 0$ such that for $i = 1, 2$

$$\tilde{C}\,h + \max_{\xi \in [t^*-\delta, t^*+\delta]\setminus\{t^*\}} \left| \frac{d^2}{dt^2} d_i(\xi) \right| |t^* - t_w| \le \frac{1}{2\|DF(\mathbf{u})^{-1}\|},\ \forall h < \bar{h}_1.$$

We deduce:

$$\|DF(\mathbf{u})(\mathbf{w}-\mathbf{u}) - \mathbf{F_h}(\mathbf{w}) + \mathbf{F_h}(\mathbf{u})\| \ \le\ \frac{1}{2\|DF(\mathbf{u})^{-1}\|}\|\mathbf{u}-\mathbf{w}\|, \quad \forall\, h \le \bar{h}_1.$$

Consequently:

$$\begin{aligned} \|G(\mathbf{w}) - G(\mathbf{u})\| &\le \|DF(\mathbf{u})^{-1}\|\ \|DF(\mathbf{u})(\mathbf{w}-\mathbf{u}) - (\mathbf{F_h}(\mathbf{w}) - \mathbf{F_h}(\mathbf{u}))\|, \\ &\le \|DF(\mathbf{u})^{-1}\| \frac{1}{2\|DF(\mathbf{u})^{-1}\|}\|\mathbf{u}-\mathbf{w}\|, \\ &= \frac{1}{2}\|\mathbf{u}-\mathbf{w}\|. \end{aligned}$$

Now by the first part of the theorem, one can choose $\bar{h}_2$ in order to have $\|\mathbf{F_h}(\mathbf{u})\| \le \frac{\delta}{2\|DF(\mathbf{u})^{-1}\|}$. Consequently there exists $\bar{h} = \min(\bar{h}_1, \bar{h}_2)$ such that for $\mathbf{w} \in \mathcal{B}(\mathbf{u}, \delta)$ we have:

$$\begin{aligned} \|\mathbf{u} - G(\mathbf{w})\| &\le \|\mathbf{u} - G(\mathbf{u})\| \ +\ \|G(\mathbf{u}) - G(\mathbf{w})\|, \\ &\le \|DF(\mathbf{u})^{-1}\|\ \|\mathbf{F_h}(\mathbf{u})\| \ +\ \frac{1}{2}\|\mathbf{u}-\mathbf{w}\|, \\ &\le \frac{\delta}{2} + \frac{\delta}{2} = \delta. \end{aligned}$$

Hence $G(\mathbf{w}) \in \mathcal{B}(\mathbf{u}, \delta)$ when $\mathbf{w} \in \mathcal{B}(\mathbf{u}, \delta)$. So the operator G is a strict contraction of the ball $\mathcal{B}(\mathbf{u}, \delta)$ and admits a unique fixed point in $\mathcal{B}(\mathbf{u}, \delta)$ denoted by $\mathbf{u_h}$. The second part of the theorem is proved.

For the last part of the theorem, let us take $\mathbf{w} = \mathbf{u_h}$ in the above inequality. We obtain by noticing that $\mathbf{u_h} = G(\mathbf{u_h})$

$$\|\mathbf{u} - \mathbf{u_h}\| \ \le\ \|DF(\mathbf{u})^{-1}\|\ \|\mathbf{F_h}(\mathbf{u})\| \ +\ \frac{1}{2}\|\mathbf{u} - \mathbf{u_h}\|.$$

Hence

$$\|\mathbf{u} - \mathbf{u_h}\| \ \le\ 2\|DF(\mathbf{u})^{-1}\|\ \|\mathbf{F_h}(\mathbf{u})\|.$$

The relation (3.4.2) with $\bar{C} = 2\|DF(\mathbf{u})^{-1}\|$ is then obtained. □

Since the numerical methods used for the discretization of the system (3.4.1) are second order methods, one can be more precise on the first part of the theorem and prove that there exists $\tilde{h} > 0$ and a constant $\tilde{c}$ such that

$$\|\mathbf{F_h}(\mathbf{u})\| \ \le\ \tilde{c}\,h^2, \qquad \forall\, h \le \tilde{h}.$$

Then combination of this relation with (3.4.2) allow to conclude

$$\exists h_0 > 0,\ C_0 > 0 \text{ s. t. } |t^* - t_h^*| + |\ell^* - \ell_h^*| \ \le\ C_0 h^2, \quad \forall h < h_0. \tag{3.4.6}$$

3.4.2 Theoretical results in s dimensions

This theory can be generalized to several dimensions. The problem is to find the intersection between the curve $\mathbf{d}(\cdot)$ and the hyperplane defined by $\mathbf{OC} + \sum_{i=1}^{s-1} \lambda_i \mathbf{v}_i$, where $\mathbf{C}$ is a point in the plane and $\mathbf{v}_i$ are the direction vectors of the hyperplane. Thus the Lemma 3.4.1 is changed into

Lemma 3.4.3. *If* $\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{d}(t^*)$ *is not parallel to the hyperplane, then* $DF(\mathbf{u})$ *is regular.*

Proof. The Jacobian of $\mathbf{F}$ at $\mathbf{u}$ is given by

$$DF(\mathbf{u}) = \begin{pmatrix} \frac{\mathrm{d}}{\mathrm{d}t}d_1(t^*) & -v_{1,1} & \dots & -v_{s-1,1} \\ \vdots & \vdots & & \vdots \\ \frac{\mathrm{d}}{\mathrm{d}t}d_s(t^*) & -v_{1,s} & \dots & -v_{s-1,s} \end{pmatrix},$$

where $v_{i,j}$ denotes the j^{th} component of the vector $\mathbf{v}_i$.
If $\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{d}(t^*)$ is not parallel to the hyperplane, then $\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{d}(t^*)$ is not a linear combination of the vectors $\mathbf{v}_i$, $i = 1, \dots, s-1$. Consequently the determinant of the Jacobian matrix is not equal to zero. ☐

In the Theorem 3.4.2 the dimension appears in the computation of the matrix-vector product $DF(\mathbf{u})(\mathbf{w} - \mathbf{u})$. In this case one has

$$\begin{aligned} DF(\mathbf{u})(\mathbf{w}-\mathbf{u}) &= \begin{pmatrix} \frac{\mathrm{d}}{\mathrm{d}t}d_1(t^*) & -v_{1,1} & \dots & -v_{s-1,1} \\ \vdots & \vdots & & \vdots \\ \frac{\mathrm{d}}{\mathrm{d}t}d_s(t^*) & -v_{1,s} & \dots & -v_{s-1,s} \end{pmatrix} \begin{pmatrix} t^* - t_w \\ \ell_1^* - \ell_{w,1} \\ \vdots \\ \ell_{n-1}^* - \ell_{n-1}^v \end{pmatrix} \\ &= \begin{pmatrix} \frac{\mathrm{d}}{\mathrm{d}t}d_1(t^*)(t^*-t_w) & -v_{1,1}(\ell_1^* - \ell_{w,1}) & \dots & -v_{s-1,1}(\ell_{s-1}^* - \ell_{w,s-1}) \\ \vdots & \vdots & & \vdots \\ \frac{\mathrm{d}}{\mathrm{d}t}d_s(t^*)(t^*-t_w) & -v_{1,s}(\ell_1^* - \ell_{w,1}) & \dots & -v_{s-1,s}(\ell_{s-1}^* - \ell_{w,s-1}) \end{pmatrix}, \end{aligned}$$

where $\mathbf{w}^T = (t_w, \ell_{w,1}, \dots, \ell_{w,s-1})$ and $\mathbf{u}^T = (t^*, \ell_1^*, \dots, \ell_{s-1}^*)$.
Then the expression $DF(\mathbf{u})(\mathbf{w}-\mathbf{u}) - \mathbf{F_h}(\mathbf{w}) + \mathbf{F_h}(\mathbf{u})$ equals to

$$DF(\mathbf{u})(\mathbf{w}-\mathbf{u}) - \mathbf{F_h}(\mathbf{w}) + \mathbf{F_h}(\mathbf{u}) = \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{d}(t^*)(t^* - t_w) - \mathbf{d}_h(t_w) + \mathbf{d}_h(t^*),$$

which is exactly the same expression as in the Theorem 3.4.2. Thus the theorem is still valid.

3.4.3 Theoretical example

Under the assumption that the optimization algorithm is exact, the system (3.1.1) can be reduced when additional assumptions are considered. Indeed in the particular case of the p-simplex (i.e. when $s = p$), the phase simplex is a polyhedron of dimension s and its edges are constant and exactly known once an optimization problem is computed for a $\mathbf{b}(t)$ belonging in the interior of the polyhedron. Consequently for any given analytical function $\mathbf{f}$ and for an initial point $\mathbf{b}_0$ situated in the p-simplex, the intersection between the trajectory of $\mathbf{b}$ and the p-simplex can be computed exactly.

Let us work with 3 chemical components. Hence the p-simplex is a triangle. Furthermore let us consider a function $\mathbf{f}$ which is a simplified case of the flux in the original problem (3.1.1). In other words let us assume

- only one particle is in the system, i.e. $N = 1$,
- in the definition of the flux $\mathbf{j}$ the matrix $\mathbf{H}$ is equal to the identity matrix.

So the Cauchy's problem of (3.4.1) reads

$$\left\{ \begin{array}{rcl} \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{b}(t) & = & -\mathbf{b}(t) + \mathbf{K}, \\ \mathbf{b}(0) & = & \mathbf{b}_0, \end{array} \right. \tag{3.4.7}$$

with $\mathbf{K} = \mathbf{b}^{\mathrm{tot}} - \frac{1}{\mathcal{R}_c T}\exp(-\boldsymbol{\nabla} g(\mathbf{x}_\alpha) + \ln(\mathbf{p}_g^o)) = \mathbf{b}^{\mathrm{tot}} - \mathbf{c}_g^{surf}$. First by definition the vector $\mathbf{K}$ is positive. Second as long as $\mathbf{b}$ stays in the p-phase simplex, the vector $\boldsymbol{\nabla} g(\mathbf{x}_\alpha)$ is constant and in fact equal to $-\boldsymbol{\lambda}$, the dual variable. Hence as long as $\mathbf{b}$ stays in the p-phase simplex, the vector $\mathbf{K}$ is constant.

The solution of the Cauchy's problem is exactly known and given by:

$$\mathbf{b}(t) = (\mathbf{b}_0 - \mathbf{K})e^{-t} + \mathbf{K}.$$

Let us denote by $\mathbf{z}_1, \mathbf{z}_2$ and $\mathbf{z}_3$ the vertices of the p-simplex defined on the phase diagram. Moreover without loss of generality let us assume that the curve $\mathbf{d}(\cdot)$ crosses the edge $[\mathbf{z}_1, \mathbf{z}_3]$. One can represent analytically this edge by:

$$\begin{pmatrix} c_1 + \ell v_1 \\ c_2 + \ell v_2 \end{pmatrix}, \quad \text{with } \ell \in [0, 1],$$

where $\mathbf{z}_1^T = (c_1, c_2)$ and $\mathbf{v} = \overrightarrow{\mathbf{z}_3\mathbf{z}_1}$.

Before starting the computation of the intersection point, let us write the expression of $\mathbf{d}(\cdot) = \frac{1}{\mathbf{e}^T\mathbf{b}(\cdot)}\mathbf{b}(\cdot)$ more simply. Let us note the expression of $\mathbf{b}(t)$

$$\begin{pmatrix} b_1(t) \\ b_2(t) \\ b_3(t) \end{pmatrix} = \begin{pmatrix} A_1 e^{-t} + B_1 \\ A_2 e^{-t} + B_2 \\ A_3 e^{-t} + B_3 \end{pmatrix},$$

where $A_i = b_{0,i} - K_i$ and $B_i = K_i$ for $i = 1, \ldots, 3$.

Hence

$$\mathbf{e}^T\mathbf{b}(t) = (A_1 + A_2 + A_3)e^{-t} + B_1 + B_2 + B_3 = Ae^{-t} + B,$$

where $A := A_1 + A_2 + A_3$ and $B := B_1 + B_2 + B_3$. Consequently finding the intersection between the curve $\mathbf{d}$ and the segment $[\mathbf{z}_1, \mathbf{z}_3]$ is given by solving the following system

$$\left\{ \begin{array}{lcl} \frac{A_1 e^{-t} + B_1}{A e^{-t} + B} & = & c_1 + \ell v_1, \\ \\ \frac{A_2 e^{-t} + B_2}{A e^{-t} + B} & = & c_2 + \ell v_2, \end{array} \right.$$

where t and ℓ are the unknowns.

This system is equivalent to

$$\left\{ \begin{array}{lcl} A_1 e^{-t} + B_1 & = & (c_1 + \ell v_1)(A e^{-t} + B), \\ A_2 e^{-t} + B_2 & = & (c_2 + \ell v_2)(A e^{-t} + B). \end{array} \right.$$

Multiplying the first equation by v_2, the second one by $-v_1$ and additioning these new equations yield

$$v_2 A_1 e^{-t} + v_2 B_1 - v_1 A_2 e^{-t} - v_1 B_2 = (A e^{-t} + B)(v_2 c_1 - c_2 v_1).$$

We deduce

$$e^{-t} = \frac{B(v_2 c_1 - c_2 v_1) - v_2 B_1 + v_1 B_2}{v_2 A_1 - v_1 A_2 - A(v_2 c_1 - c_2 v_1)},$$

and obtain

$$t = \ln \left(\frac{v_2 A_1 - v_1 A_2 - A(v_2 c_1 - c_2 v_1)}{B(v_2 c_1 - c_2 v_1) - v_2 B_1 + v_1 B_2} \right).$$

So the intersection point occurs at $t^* = \ln \left(\frac{v_2 A_1 - v_1 A_2 - A(v_2 c_1 - c_2 v_1)}{B(v_2 c_1 - c_2 v_1) - v_2 B_1 + v_1 B_2} \right)$ and is given by $B(t^*)$.

To apply the theory developed in the previous section we need to verify whether the required hypotheses are satisfied. The first one to check is the positiveness of the solution.

Lemma 3.4.4. *The solution of the differential equation* $\mathbf{b}(t)$ *is positive.*

Proof. The solution $\mathbf{b}(t)$ is defined by

$$\mathbf{b}(t) \ = \ (\mathbf{b}_0 - \mathbf{K})\, e^{-t} + \mathbf{K},$$

which can be transformed into

$$\mathbf{b}(t) \ = \ e^{-t}\, \mathbf{b}_0 + (1 - e^{-t}\,)\, \mathbf{K}.$$

Since the initial condition is chosen to be positive, and the vector $\mathbf{K}$ and the variable t are by definition positive, we deduce immediately the positiveness of the solution. ☐

We need to verify the Lipschitz continuity of the function $\mathbf{f}$.

Lemma 3.4.5. *The function* $\mathbf{f}$ *of the Cauchy's problem (3.4.7) is Lipschitz and the Lipschitz constant equals to 1.*

Proof. By definition of $\mathbf{f}$ one has for all $\mathbf{b}_1, \mathbf{b}_2 \geq 0$

$$\begin{aligned}\|\mathbf{f}(\mathbf{b}_1) - \mathbf{f}(\mathbf{b}_2)\| &= \| - \mathbf{b}_1 + \mathbf{K} + \mathbf{b}_2 - \mathbf{K}\| \\ &= \|\mathbf{b}_1 - \mathbf{b}_2\|.\end{aligned}$$

It means that the function $\mathbf{f}$ is Lipschitz and the Lipschitz's constant equals to 1. ☐

The last hypothesis concerns the positiveness of the numerical scheme. One has to prove that $\mathbf{b}^k > 0$ for $k = 0, 1, \dots, m$. The first iterate is initially set to the positive value $\mathbf{b}^0 = \mathbf{b}_0$. The positiveness of the other iterates is proved in the following lemma. Let us recall that the numerical scheme is the Crank-Nicolson method except for the iterate that hits the frontier. In this last case the numerical scheme is the Adams-Bashforth method.

Lemma 3.4.6. *For all $h \in]0, 2[$, the numerical scheme Crank-Nicolson/Adams-Bashforth is positive.*

Proof. Let us assume that the event occurs during the interval $[t^n, t^{n+1}[$. Our algorithm implies that until the step t^n the Crank-Nicolson scheme is used and it follows for $k = 0, \dots, n-1$

$$\begin{aligned}\frac{1}{h}(\mathbf{b}^{k+1} - \mathbf{b}^k) &= \frac{1}{2}\mathbf{f}(\mathbf{b}^k) + \frac{1}{2}\mathbf{f}(\mathbf{b}^{k+1}), \\ &= -\frac{1}{2}\mathbf{b}^k - \frac{1}{2}\mathbf{b}^{k+1} + \mathbf{K},\end{aligned}$$

which is equivalent to

$$\mathbf{b}^{k+1} = \frac{2-h}{2+h}\mathbf{b}^k + \frac{2h}{2+h}\mathbf{K}, \qquad \text{for } k = 0, \dots, n-1.$$

Repeating iteratively the process implies

$$\begin{aligned}\mathbf{b}^{k+1} &= \frac{2-h}{2+h}\mathbf{b}^k + \frac{2h}{2+h}\mathbf{K} \\ &= \frac{2-h}{2+h}\left(\frac{2-h}{2+h}\mathbf{b}^{k-1} + \frac{2h}{2+h}\mathbf{K}\right) + \frac{2h}{2+h}\mathbf{K} \\ &= \left(\frac{2-h}{2+h}\right)^2 \mathbf{b}^{k-1} + \frac{2h}{2+h}\left(1 + \frac{2-h}{2+h}\right)\mathbf{K} \\ &\vdots \\ &= \left(\frac{2-h}{2+h}\right)^{k+1} \mathbf{b}_0 + \left[\frac{2h}{2+h}\sum_{p=0}^{k}\left(\frac{2-h}{2+h}\right)^p\right]\mathbf{K}.\end{aligned}$$

Since $h > 0$ the above sum is a geometric series equals to

$$\sum_{p=0}^{k}\left(\frac{2-h}{2+h}\right)^p = \frac{2+h}{2h}\left[1 - \left(\frac{2-h}{2+h}\right)^{k+1}\right].$$

Hence $\mathbf{b}^{k+1}$ becomes

$$\mathbf{b}^{k+1} = \left(\frac{2-h}{2+h}\right)^{k+1} \mathbf{b}_0 + \left[1 - \left(\frac{2-h}{2+h}\right)^{k+1}\right] \mathbf{K}. \tag{3.4.8}$$

Let us now consider the fraction $\frac{2-h}{2+h}$. One has

$$\frac{2-h}{2+h} = \frac{2+h-2h}{2+h} = 1 - \frac{2h}{2+h}.$$

The time step $h \in]0,2[$ yields

$$0 \le \frac{2h}{2+h} \le 1,$$

and consequently

$$0 \le \frac{2-h}{2+h} \le 1.$$

It implies for $k = 0, \ldots, n-1$

$$0 \le \left(\frac{2-h}{2+h}\right)^{k+1} \le 1 \quad \text{and} \quad 0 \le 1 - \left(\frac{2-h}{2+h}\right)^{k+1} \le 1.$$

The expression (3.4.8) for $\mathbf{b}^{k+1}$ for $k = 0, \ldots, n-1$ is then positive.

The next iterate $\mathbf{b}^{n+1}$ corresponds to the approximation of the discontinuity point and is obtained with the 2-steps Adams-Bashforth method, with $\tau = t^{n+1} - t^n$. This method writes

$$\begin{aligned}
\mathbf{b}^{n+1} &= \mathbf{b}^n + \tau \left[\left(1 + \frac{\tau}{2h}\right)(-\mathbf{b}^n + \mathbf{K}) - \frac{\tau}{2h}(-\mathbf{b}^{n-1} + \mathbf{K})\right] \\
&= \left(1 - \tau - \frac{\tau^2}{2h}\right) \mathbf{b}^n + \frac{\tau^2}{2h} \mathbf{b}^{n-1} + \tau \mathbf{K}.
\end{aligned}$$

Then with the relation (3.4.8) the expression of $\mathbf{b}^{n+1}$ becomes

$$\begin{aligned}
\mathbf{b}^{n+1} &= \left(1 - \tau - \frac{\tau^2}{2h}\right) \mathbf{b}^n + \frac{\tau^2}{2h} \mathbf{b}^{n-1} + \tau \mathbf{K}, \\
&= \left(1 - \tau - \frac{\tau^2}{2h}\right) \left(\frac{2-h}{2+h}\right)^n \mathbf{b}_0 + \left(1 - \tau - \frac{\tau^2}{2h}\right) \left[1 - \left(\frac{2-h}{2+h}\right)^n\right] \mathbf{K} \\
&\quad + \frac{\tau^2}{2h} \left(\frac{2-h}{2+h}\right)^{n-1} \mathbf{b}_0 + \frac{\tau^2}{2h} \left[1 - \left(\frac{2-h}{2+h}\right)^{n-1}\right] \mathbf{K} + \tau \mathbf{K}, \\
&= \left(\frac{2-h}{2+h}\right)^n \left[1 - \tau - \frac{\tau^2}{2h} + \frac{\tau^2}{2h} \frac{2+h}{2-h}\right] \mathbf{b}_0 \\
&\quad + \left[1 - \tau - \frac{\tau^2}{2h} + \frac{\tau^2}{2h} - \left(\frac{2-h}{2+h}\right)^n \left(1 - \tau - \frac{\tau^2}{2h} + \frac{\tau^2}{2h} \frac{2+h}{2-h}\right) + \tau\right] \mathbf{K}, \\
&= \left(\frac{2-h}{2+h}\right)^n \left(1 - \tau + \frac{\tau^2}{2-h}\right) \mathbf{b}_0 + \left[1 - \left(\frac{2-h}{2+h}\right)^n \left(1 - \tau + \frac{\tau^2}{2-h}\right)\right] \mathbf{K}.
\end{aligned}$$

Let us write $\mathbf{b}^{n+1}$ in the following manner, which is more convenient to prove the positiveness of $\mathbf{b}^{n+1}$

$$\begin{aligned}\mathbf{b}^{n+1} &= \frac{(2-h)^{n-1}}{(2+h)^n}\left(2-h-(2-h)\tau+\tau^2\right)\mathbf{b}_0 + \\ &\qquad \left[1-\frac{(2-h)^{n-1}}{(2+h)^n}\left(2-h-(2-h)\tau+\tau^2\right)\right]\mathbf{K}. \qquad (3.4.9)\end{aligned}$$

By definition the vectors $\mathbf{K}$ and $\mathbf{b}_0$ are positive. It remains to prove the positiveness of $2-h-(2-h)\tau+\tau^2$ and $1-\frac{(2-h)^{n-1}}{(2+h)^n}\left(2-h-(2-h)\tau+\tau^2\right)$. In that way let us define the polynomial function $q:]0,h] \to \mathbb{R}$ as $q(\tau) = \tau^2-(2-h)\,\tau+2-h$. Then q is a second order polynomial in τ and its discrimant is given by

$$\triangle = (2-h)^2-4(2-h) = h^2-4.$$

This discriminant is negative $\forall h \in]0,2[$. We conclude that $q(\tau)$ is positive $\forall \tau \in]0,h]$. Then the factor $2-h-(2-h)\tau+\tau^2$ is always positive. In order to determine the positiveness of $1-\frac{(2-h)^{n-1}}{(2+h)^n}[2-h-(2-h)\tau+\tau^2]$, let us prove

$$0 \le \frac{(2-h)^{n-1}}{(2+h)^n}\left[2-h-(2-h)\tau+\tau^2\right] \le 1,\ \forall \tau \in]0,1],\ \forall h \in]0,2[.$$

Since $0 \le \frac{(2-h)^{n-1}}{(2+h)^{n-1}} \le 1$ and the function q is positive, it remains to establish

$$\frac{1}{2+h}[2-h-(2-h)\tau+\tau^2] \le 1.$$

Let us define the polynomial function $\tilde{q}:]0,h] \to \mathbb{R}$ as $\tilde{q}(\tau) = \frac{1}{2+h}[2-h-(2-h)\tau+\tau^2]$. Then $\tilde{q}$ is a second order polynomial in τ and one has

$$\begin{aligned}\tilde{q}(0) &= \frac{2-h}{2+h} \le 1, \forall h \in]0,2[; \\ \tilde{q}(h) &= \frac{2-h-(2-h)h+h^2}{2+h}, \\ &= \frac{2-3h+2h^2}{2+h}, \\ &= 1-2h\frac{2-h}{2+h} \le 1, \forall h \in]0,2[.\end{aligned}$$

Consequently $\tilde{q}$ is bounded from above by 1 and $1-\frac{(2-h)^{n-1}}{(2+h)^n}\left(2-h-(2-h)\tau+\tau^2\right)$ is positive. Thanks to (3.4.9) we deduce the positiveness of $\mathbf{b}^{n+1}$.

For the next iterates $\mathbf{b}^k$ with $k \ge n+2$ the Crank-Nicolson scheme is used again. So by definition one has for $k = n+1, n+2, \ldots$

$$\begin{aligned}\mathbf{b}^{k+1} &= \frac{2-h}{2+h}\mathbf{b}^k + \frac{2h}{2+h}\mathbf{K} \\ &= \ldots = \left(\frac{2-h}{2+h}\right)^{k-n}\mathbf{b}^{n+1} + \frac{2h}{2+h}\sum_{p=0}^{k-n-1}\left(\frac{2-h}{2+h}\right)^p\mathbf{K}.\end{aligned}$$

Since $h \in]0,1[$ and $\mathbf{b}^{n+1}$ and $\mathbf{K}$ are positive, the above expression is positive too for all $k \geq n+1$.
In conclusion we have proven that $\mathbf{b}^k > \mathbf{0}$ for all $k \in \mathbb{N}$. $\square$

The numerical results for this example when $s = 3$ are presented in the following section.

3.5 Numerical results

In atmospheric chemistry literature phase diagrams of only few organic aerosol particles are available. Moreover for particles with more than 4 chemical components, visualization of the associated phase diagrams becomes very difficult. In [3, 30, 100, 105] phase diagrams are presented for aerosols made of 2 or 3 chemicals. In this section let us consider the examples of [3] in order to check the correctness of the solution, i.e. if the solution follows the phase repartition described on the phase diagrams.

First examples concern particles with 2 chemical species ($s = 2$). Then examples with 3 chemicals ($s = 3$) are considered. In addition to these examples, the choice of the warm-start strategy versus the cold-start one is justified. Finally the theoretical results of Section 3.4 on the example of Subsection 3.4.3 are applied for different initial conditions $\mathbf{b}_0$.

Then 2 examples with $s = 4$ are presented. No phase diagram is available. However the behaviour of the approximations $\mathbf{b}^n$, $n = 0, 1, \dots$ can be observed on the tetrahedron Δ_3 as well as the time evolution of y_α^n, $n = 0, 1, \dots$ with the detection and computation of the discontinuities.

Finally some examples for an aerosol made of $s = 18$ chemical species are studied. Once again no phase diagram is available and no result exist in the literature that may confirm or invalidate our results. Nevertheless the efficiency of the numerical method to solve the optimization-constrained differential system and to detect the discontinuities are highlighted.

For all the following numerical examples, the temperature of the gas-aerosol system is equal to 298.15 K and the pressure is 1 atm. Furthermore the interaction parameters that define the molar Gibbs free energy g are derived from vapor-liquid equilibrium data (Hansen et al. in [51]) except for one class of examples when $s = 3$. For this case the parameters are taken from the liquid-liquid equilibrium data referred in Magnussen et al. [64]. Unless specified otherwise, the number of aerosol particles in the system per unit of volume N is taken equal to 10^7 m^{-3}, the tolerance of the stopping criterion for the fixed-point method is 10^{-5}, and the tolerance for the interior-point method is 10^{-13} whereas the maximal number of iterations is fixed to 300.

All computations are executed on a workstation with an Intel processor of 2.4 GHz and 2 GB of RAM memory.

3.5.1 Numerical results in one dimension

In these examples, the index 1 is associated to the convex area situated on the left side of the phase diagram and the index 2 stands for the other side. Moreover the regions on the phase diagram for which one inequality constraint is inactive is denoted *area 1*, and the other region with both inactive constraints is called *area 2*.

The first example concerns an aerosol particle made of 1-hexacosanol and pinic acid. For this example the molar Gibbs free energy g is represented by a black curve in Figure 3.7 (left). The corresponding phase diagram is depicted at the bottom of the figure. The phase diagram is divided in 3 areas as follows

$$\Delta_1 = [0,1] = [0,\, 0.0665672] \cup]0.0665672,\, 0.4633492[\cup [0.4633492,\, 1],$$

the subinterval in the middle being the area 2 and the two other subintervals are the areas 1. For this example and the remaining of the section, the area 2 is colored in red whereas the black color is associated to area 1.

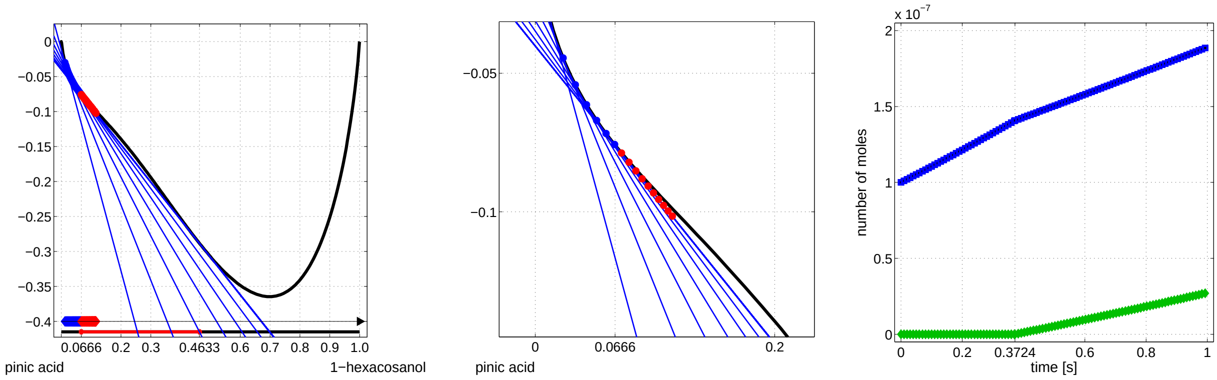


Figure 3.7: Organic aerosol made of 1-hexacosanol and pinic acid with initial composition-vector $\mathbf{b}_0^T = (0.01 \cdot 10^{-7},\, 0.99 \cdot 10^{-7})$ mol. The deactivation occurs at $t = 0.3724$ s. Left: evolution of $\mathbf{b}^n$, $n = 0, 1, 2, \ldots$, the corresponding supporting tangent plane evolves until making contact with the graph of g. Middle: zoomed-in view of the deactivation. Right: time evolution of y_1^n and y_2^n, $n = 0, 1, 2, \ldots$.

In Figure 3.7 (left) the approximations $\mathbf{b}^n$, $n = 0, 1, 2, \ldots$ are represented by colored diamonds situated on an axis above the phase diagram. The corresponding values for the energy $g(\mathbf{b}^n)$, $n = 0, 1, 2, \ldots$ are drawn with colored circles. As for the phase diagram, when for $\mathbf{b}^n$ both inequality constraints are inactive, $\mathbf{b}^n$ is in red. If the number of inactive constraint is equal to 1, the approximation $\mathbf{b}^n$ is in blue.

For this first example the initial composition vector is $\mathbf{b}_0^T = (0.01{\cdot}10^{-7},\, 0.99{\cdot}10^{-7})$ mol, and its normalized value is situated in the left area 1. The initial gas concentration vector is initialized by $\mathbf{c}_{g,0}^{\infty,T} = (3,\, 7)$ mol/m^3 and $h = 0.01$ s. The simulation of this example leads to a motion of $\mathbf{b}^n$ on the right until the area 2. The result illustrated in Figure 3.7 (left) shows that the approximations $\mathbf{b}^n$, $n = 0, 1, 2, \ldots$ follow the phase diagram correctly since they become red as they enter in the area 2. Moreover the computed discontinuity point is equal to $\mathbf{b}^{n+1} = (0.94 \cdot 10^{-8}, 0.13 \cdot 10^{-6})$ mol. The error committed between the projection $\mathbf{d}^{n+1} = \frac{1}{\mathbf{e}^T \mathbf{b}^{n+1}} P\mathbf{b}^{n+1}$ and the boundary point between the left area 1 and the area 2 of the phase diagram, namely 0.0665672, is then given by $|0.0665672 - 0.0665655| = 1.7 \cdot 10^{-6}$.

In Figure 3.7 (left) the time evolution of the supporting tangent plane is also depicted. One can observe how the sequence of the tangent planes tends to the new contact point as $\mathbf{b}^n$, $n = 0, 1, 2, \ldots$ tend to the boundary point 0.0665672. A zoomed-in view of the points $g(\mathbf{b}^n)$ around the deactivation of the second inequality constraint is proposed in Figure 3.7 (middle). The points $g(\mathbf{b}^n)$ before the deactivation are blue and the ones after the deactivation are red. Moreover these points follow correctly the convex envelope of g since the blue points are on g and the red points follow the supporting tangent plane which defines the convex envelope as both inequality constraints are inactive.

Finally Figure 3.7 (right) shows the time evolution of the constraints y_1^n (blue plot) and y_2^n (green plot), $n = 0, 1, 2, \ldots$. At time t^0, $y_1^0 = 10^{-7}$ and $y_2^0 = 0$, which corresponds to the initial state. Then at time $t = 0.3724$ s the deactivation occurs and y_2^n becomes positive afterwards. At this time one can observe that the evolution of the points y_2^n undergoes a discontinuity in its derivative, and that it is also the case for y_1^n.

Figure 3.8 illustrates the computation of the points $(\mathbf{x}_\alpha, g(\mathbf{x}_\alpha))$, $\alpha \in \mathcal{A}$ at minimal distance to the supporting tangent plane for the first example. The computation at three different time steps is considered. Figure 3.8 (left) stands for a time step far before the deactivation. The graph in the middle shows the computation at the time step t^n, just before the deactivation, whereas the graph on the right is at t^{n+1}, just after the deactivation. The blue curve represents the distance function d. The phase diagram is illustrated at the bottom of the graph as in Figure 3.7 (left). Finally the black diamonds on the axis are the successive Newton iterations and their correspondants on the curve d are denoted by black circles.

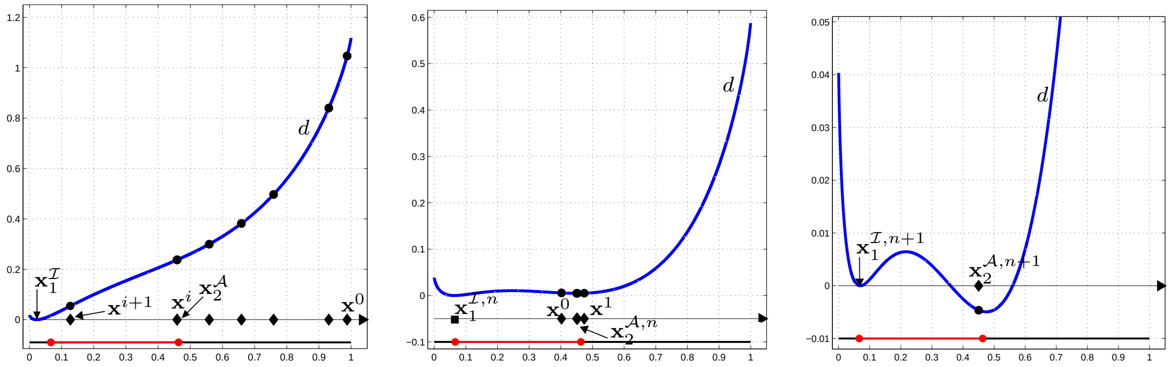


Figure 3.8: Computation of the minimal distance between the graph of g and the supporting tangent plane. Left: distance function when $\mathbf{b}$ is far away from the event. Middle: at time t^n, a local minimum appears. Right: at time t^{n+1}, the point $\mathbf{x}_2^{\mathcal{A},n+1}$ is located at a negative distance to the tangent plane.

When the time step is far before the deactivation (Figure 3.8 (left)) the distance function d is convex and admits one unique global minimizer on $[0,1]$. The Newton sequence naturally tends to the global minimizer, but does not go out from the area 1. As depicted in Figure 3.8 (left) the algorithm detects when the sequence goes out, stops the Newton method and sets $\mathbf{x}_2^{\mathcal{A}}$ to the last admissible Newton iterate.

In the middle figure the time step t^n is near the discontinuity time. The distance function d is stretched and a local minimum appears. The Newton sequence converges to this local minimizer. At time t^{n+1} the Newton sequence converges to a point with a negative distance to the tangent plane. It indicates that the deactivation occurs in $[t^n,\, t^{n+1}]$. Note that $\mathbf{x}_2^{\mathcal{A},n+1}$ is a good approximation of the deactivation point.

In Figure 3.9 the aerosol particle has the same composition and the initialization of the example is given by

$$\mathbf{b}_0^T \;=\; (3\cdot 10^{-8},\, 7\cdot 10^{-8})\ \text{mol},\ \mathbf{c}_{g,0}^{\infty,T} \;=\; (9,\, 1)\ \text{mol/m}^3,\ h = 0.1\ \text{s}.$$

Hence the initial point $\mathbf{b}_0$ is situated in the area 2 and the initial gas-concentration in the right area 1. The motion of $\mathbf{b}^n$, $n = 0,1,2,\ldots$ is from left to right and an activation occurs during the simulation. Figure 3.9 uses the same notations as in Figure 3.7. The evolution of $\mathbf{b}^n$, $n = 0,1,2,\ldots$, follows the colored repartition of the phase diagram and the discontinuity point $\mathbf{b}^{n+1}$ is well situated on the boundary point between the area 2 and the right area 1 of the phase diagram. The difference between these two points is given by $|0.46334919 - 0.46334949| = 3.0\cdot 10^{-7}$. The successive supporting tangent planes are also depicted. The planes respect the Gibbs criterion and progressively release from the first contact point as the activation occurs.

The time evolution of y_1^n and y_2^n, $n = 0,1,2,\ldots$, is shown in Figure 3.9 (right). As for the previous example both evolutions lose their regularity at the discontinuity time.

Figures 3.10 and 3.11 illustrate two other examples with $s = 2$. In Figure 3.10 an example of a gas-aerosol system with a very small area 1 on the left of its phase diagram is considered. The chemical components of the system are water and 1-hexacosanol, and the initial conditions are given by

$$\mathbf{b}_0^T \;=\; (9.9\cdot 10^{-8},\, 0.1\cdot 10^{-8})\ \text{mol},\ \mathbf{c}_{g,0}^{\infty,T} \;=\; (3,\, 7)\ \text{mol/m}^3,\ h = 0.01\ \text{s}.$$

The initial vector $\mathbf{b}_0$ is then situated in the right area 1 and the approximations $\mathbf{b}^n$, $n = 0,1,2,\ldots$, move from right to left of the phase diagram. A deactivation occurs at $t = 0.0221$ s. The time evolution of $\mathbf{b}^n$, $n = 0,1,2,\ldots$, on the phase diagram and of $g(\mathbf{b}^n)$, $n = 0,1,2,\ldots$, are illustrated in Figure 3.10 (left). Figure 3.10 (middle) proposes a zoomed-in view of the deactivation whereas Figure 3.10 (right) shows the evolution of y_1^n and y_2^n, $n = 0,1,2,\ldots$, with their respective discontinuity at $t = 0.0221$ s. The error between the boundary point and the discontinuity point is given by $|0.89884 - 0.89838| = 0.00046$.

Figure 3.11 shows an example of a gas-aerosol system for which the distinction between the energy function g and its convex envelope is very small on the convex area 2. The numerical example illustrates a complete time evolution with a deactivation and an activation. The initial conditions are

$$\mathbf{b}_0^T \;=\; (9.5\cdot 10^{-8},\, 0.5\cdot 10^{-8})\ \text{mol},\ \mathbf{c}_{g,0}^{\infty,T} \;=\; (1,\, 9)\ \text{mol/m}^3,\ h = 0.1\ \text{s}.$$

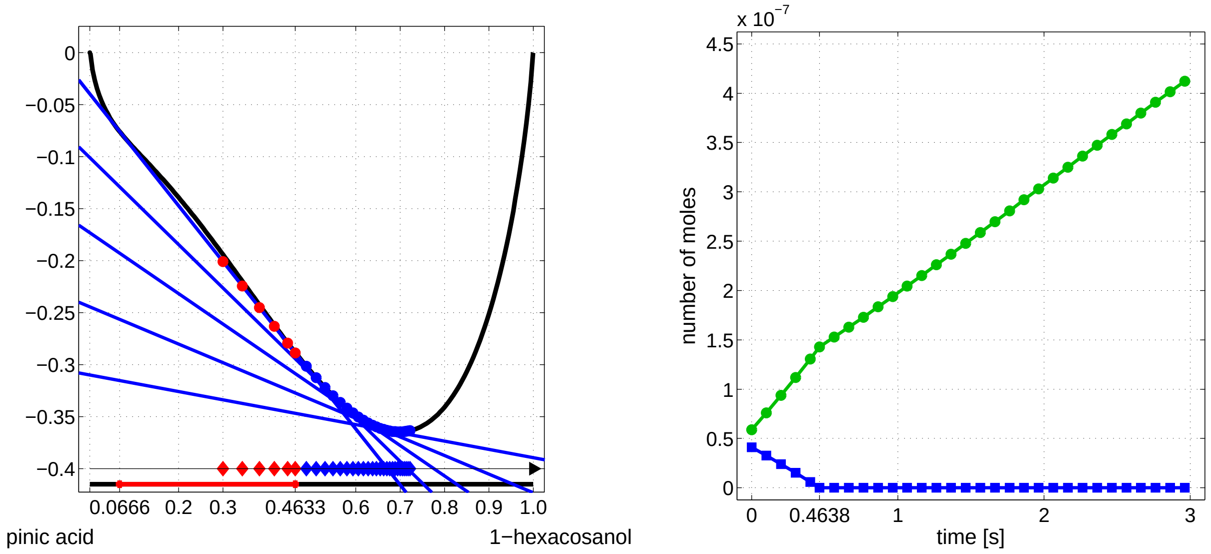


Figure 3.9: Organic aerosol made of 1-hexacosanol and pinic acid with initial composition-vector $\mathbf{b}_0^T = (3 \cdot 10^{-8}, 7 \cdot 10^{-8})$ mol. The activation occurs at $t = 0.4633$ s. Left: evolution of $\mathbf{b}^n$, $n = 0, 1, 2, \ldots$, the corresponding supporting tangent plane evolves after leaving the contact with the left convex region on the graph of g. Right: time evolution of y_1^n and y_2^n, $n = 0, 1, 2, \ldots$.

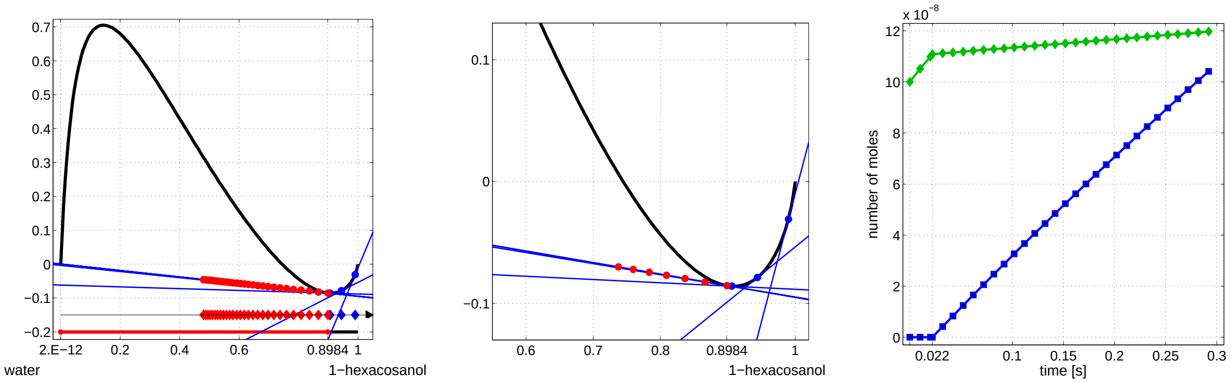


Figure 3.10: Organic aerosol made of water and 1-hexacosanol with initial composition-vector $\mathbf{b}_0^T = (9.9 \cdot 10^{-8}, 0.1 \cdot 10^{-8})$ mol. The deactivation occurs at $t = 0.0221$ s. Left: evolution of $\mathbf{b}^n$, $n = 0, 1, 2, \ldots$ and the corresponding supporting tangent planes. Middle: zoomed-in view of the deactivation. Right: evolution of y_1^n and y_2^n, $n = 0, 1, 2, \ldots$.

The initial composition vector $\mathbf{b}_0$ is located in the right area 1 and the approximations $\mathbf{b}^n$, $n = 0, 1, 2, \ldots$ moves from right to left. Hence the simulation encounters first a deactivation at $t = 0.542$ s and second, an activation at $t = 4.517$ s. The error between the boundary points and the discontinuity points are given by

- for the deactivation: $|0.3485 - 0.3453| = 0.0032$,
- for the activation: $|0.0936539099 - 0.0936539076| = 2.3 \cdot 10^{-9}$.

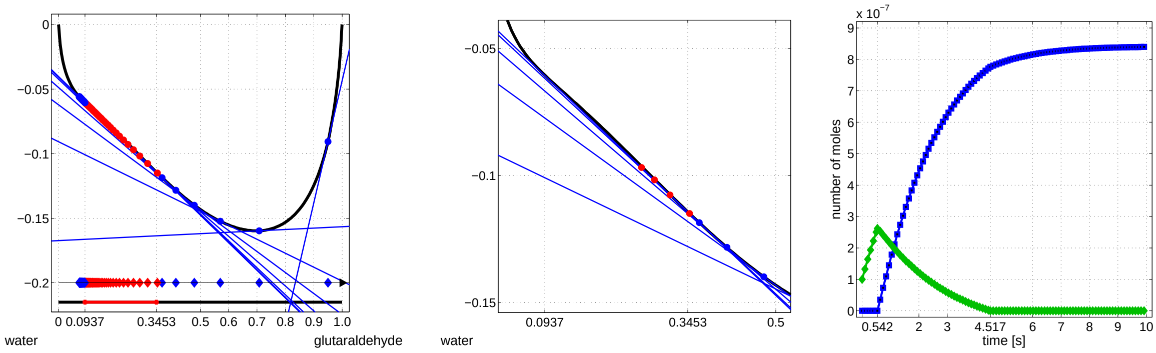


Figure 3.11: Organic aerosol made of water and glutaraldehyde with initial composition-vector $\mathbf{b}_0^T = (9.5 \cdot 10^{-8}, 0.5 \cdot 10^{-8})$ mol. The deactivation occurs at $t = 0.542$ s and the activation at $t = 4.517$ s. Left: evolution of $\mathbf{b}^n$, $n = 0, 1, 2, \ldots$, and the corresponding supporting tangent planes. Middle: zoomed-in view of the deactivation. Right: evolution of y_1^n and y_2^n, $n = 0, 1, 2, \ldots$.

3.5.2 Numerical results in two dimensions

The second class of numerical examples stands for $s = 3$. The gas-aerosol system is made of pinic acid, 1-hexacosanol and water. In Figure 3.12 two different phase diagrams are illustrated for this gas-aerosol system. The difference between both phase diagram resides on the interaction parameters that define the energy function g leading actually to two different models for the objective function g. For the phase diagram on the left, the interaction parameters are derived from vapor-liquid equilibrium data of [51]. For the other phase diagram, the parameters are derived from liquid-liquid equilibrium data of [64]. The phase diagram on the left has the same shape as the Figure 1.4 with extremely small bottom left area 1 and bottom area 2, whereas the second phase diagram has two areas 1 that are common. So the first phase diagram is classical and the second is not classical. Let the first example be called *example VL* and the second, *example LL* in reference to the respective data that generate g.

For the numerical results presented in this subsection the color notation on the phase diagram is as follows when the phase simplices and the composition-vectors $\mathbf{b}^n$ are plotted:

- green: the solution of the PEP defined at $\mathbf{b}^n$ has one inactive inequality constraint,
- blue: the solution of the PEP defined at $\mathbf{b}^n$ has two inactive inequality constraints,

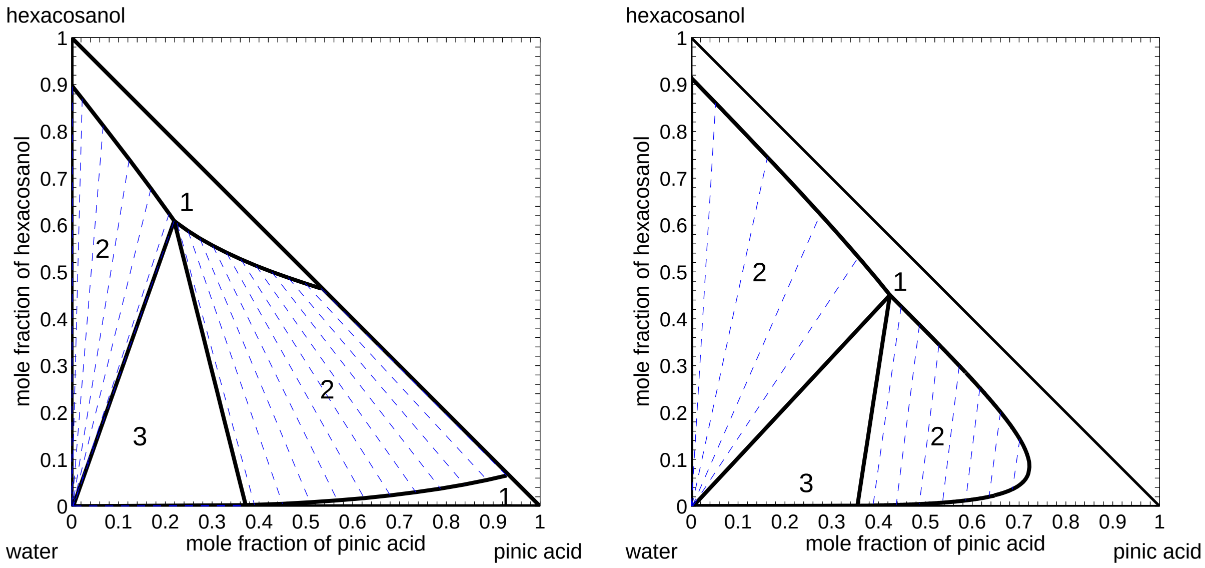


Figure 3.12: Phase diagrams for the system pinic acid/ 1-hexacosanol/ water with two different sets of interaction parameters: (left) from vapor-liquid equilibrium data [51], (right) from liquid-liquid equilibrium data [64].

- red: the solution of the PEP defined at $\mathbf{b}^n$ has three inactive inequality constraints.

If the single-phase points $\mathbf{x}_\alpha$, $\alpha = 1, 2, 3$, are plotted on the phase diagram, then the color notation means

- ■: the single-phase point $\mathbf{x}_1$,
- ●: the single-phase point $\mathbf{x}_2$,
- ♦: the single-phase point $\mathbf{x}_3$.

First examples are concerned with the comparison of the warm-start and the cold-start strategies on both phase diagrams.

Warm-start vs cold-start

In order to compare the warm-start to the cold-start strategies and to quantify the gain provided by the warm-start strategy, let us consider a fixed artificial trajectory on the phase diagram and look at the number of correct solutions as well as the computation times of each strategy.

The trajectory starts from the bottom left corner of the phase diagram and zigzags until the top left corner as depicted in Figures 3.13 or 3.14. On this trajectory the composition-vectors $\mathbf{b}^n$ are fixed and equidistant from eachother. Hence if $\bar{n}$ is the number of composition-vector situated on the first row of the trajectory, then the distance between

2 vectors $\mathbf{b}^n$ and $\mathbf{b}^{n+1}$ along the axis Ox is equal to $|b_1^n - b_1^{n+1}| = \frac{1}{\bar{n}+1}$ and the total number of $\mathbf{b}^n$ on the trajectory is $\frac{\bar{n}(\bar{n}+1)}{2}$. Four different values of $\bar{n}$ are considered: 10, 20, 40 and 80; given respectively a total number of composition vectors: 55, 210, 820 and 3240.

Since the composition-vectors $\mathbf{b}^n$, $n = 1, \dots, \frac{\bar{n}(\bar{n}+1)}{2}$, are fixed, the fixed-point algorithm is not used. Only the primal-dual interior-point method is employed to compute the phase equilibrium problem defined for the successive $\mathbf{b}^n$, $n = 1, \dots, \frac{\bar{n}(\bar{n}+1)}{2}$. Thus the efficiency of the warm-start and cold-start strategies are fully revealed.

In addition to the warm-start and cold-start procedures for the initialization of the interior-point method, a technique of backwards check is implemented in order to improve the robustness of the algorithm. Suppose that the number of inactive inequality constraints changes at $\mathbf{b}^n$. The backwards check consists in verifying the correctness of the solution of the last optimization problem (i.e. at $\mathbf{b}^{n-1}$) by solving this last optimization problem with the solution of the actual problem (i.e. at $\mathbf{b}^n$) as the warm-start. If the solution is different from the first resolution, then the algorithm returns back to the previous optimization problems with the warm-start strategy until the solution of the new computation coincides with the former solution. Hence this technique is useful to correct the solution when the activation or deactivation is detected too late. In order to estimate the need and the cost of this technique, the trajectory is solved without and with the backwards check. Summarizing, four different methods are exploited for the computation of the successive optimization problems:

1. warm-start: the initialization procedure for the interior-point method is the warm-start strategy described in Section 3.1.2.

2. warm-start & back: in addition to the warm-start strategy, the backwards check is applied as soon as the number of inactive inequality constraints changes. Suppose this number changes for $\mathbf{b}^n$. Once the backwards check is done, the simulation restarts at $\mathbf{b}^{n+1}$ with a warm-start based on the solution at $\mathbf{b}^n$.

3. cold-start: the initialization of the primal-dual interior-point method for the successive PEP is done according to the cold-start strategy described in Section 2.2.3.

4. cold-start & back: in addition to the cold-start strategy, the backwards check is applied as soon as the number of inactive inequality constraints changes. Suppose this number changes for $\mathbf{b}^n$. Once the backwards check is done, the simulation restarts at $\mathbf{b}^{n+1}$ with a cold-start.

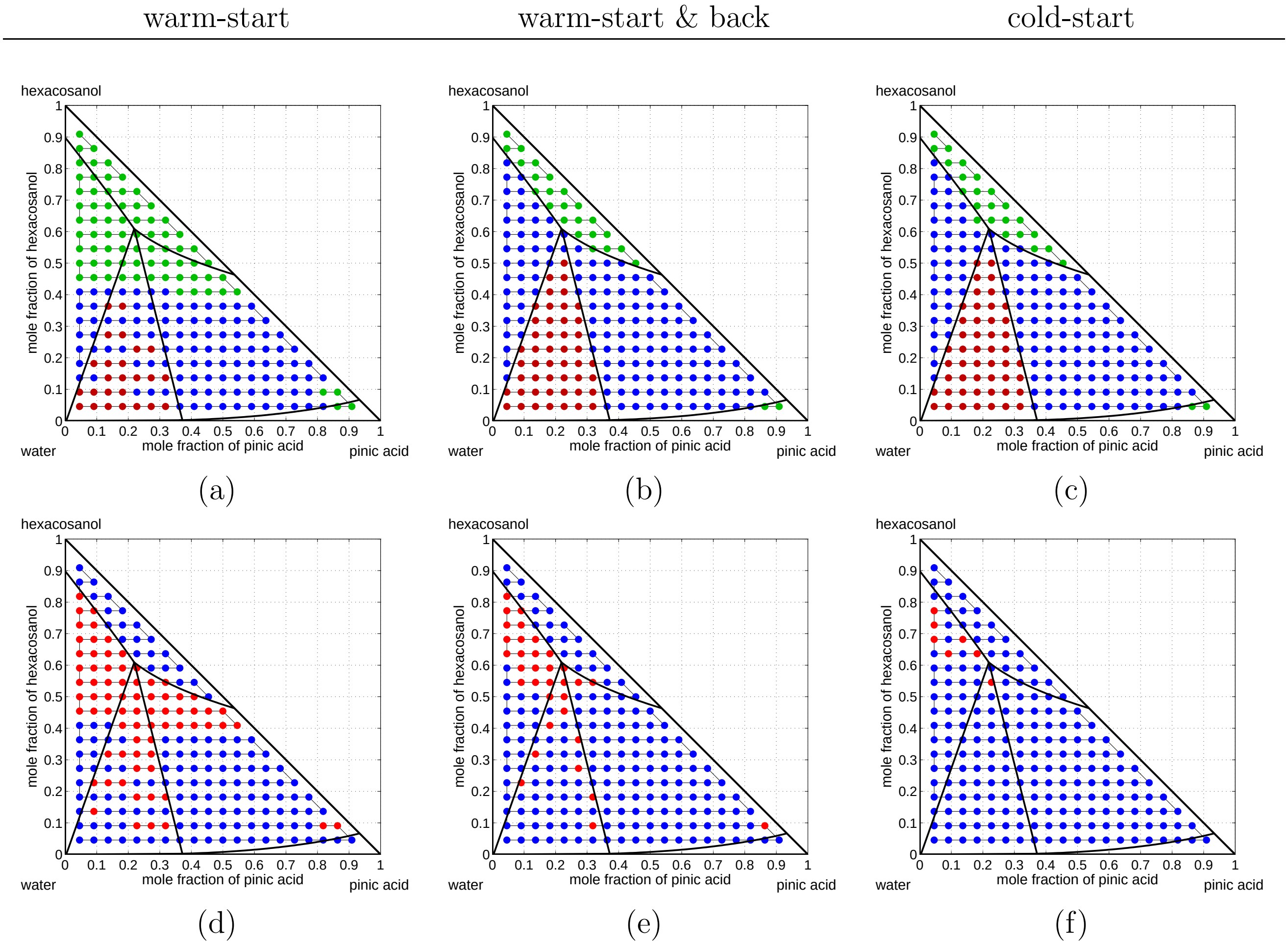
Thanks to the phase diagrams of Figure 3.12 one can check the correctness of the solution for each $\mathbf{b}^n$, $n = 1, \dots, \frac{\bar{n}(\bar{n}+1)}{2}$. The number and percentage of correct solutions for each method and each $\bar{n}$ are registered in Tables 3.1 and 3.2. Furthermore the mean number of iterations of the primal-dual interior-point method and the CPU times are considered to compare the warm-start to the cold-start strategies. These comparisons allow to quantify the computational gain issued from the warm-start. The computational cost of the backwards check may also be estimated by the comparison of the CPU times of each method.

		number of correct solutions	% of correct solutions	mean number of it. for i.p.m.	CPU time [s]
$\bar{n} = 10$	warm-start	36	65.45	25.71	0.09
	warm-start & back	49	89.09	27.05	0.15
	cold-start	54	98.18	68.20	0.22
	cold-start & back	55	100.00	68.29	0.32
$\bar{n} = 20$	warm-start	137	65.24	26.00	0.34
	warm-start & back	177	84.29	24.41	0.47
	cold-start	203	96.67	68.28	0.84
	cold-start & back	210	100.00	68.30	0.96
$\bar{n} = 40$	warm-start	530	64.63	25.65	1.26
	warm-start & back	730	89.02	23.83	1.54
	cold-start	809	98.66	68.22	3.26
	cold-start & back	817	99.63	68.24	3.47
$\bar{n} = 80$	warm-start	2023	62.44	25.34	4.92
	warm-start & back	2928	90.37	23.51	6.24
	cold-start	3146	97.10	68.25	12.86
	cold-start & back	3225	99.54	68.27	13.48

Table 3.1: Number of correct solutions, percentage of correct solutions, mean number of iterations for the primal-dual interior-point method (i.p.m.) and CPU time for the phase diagram with interaction parameters stemmed from the vapor-liquid equilibrium data of [51].

The number and percentage of correct solutions, the mean number of iterations for the primal-dual interior-point method and the CPU times for the system with interaction parameters stemmed from the vapor-liquid equilibria (see the phase diagram in Figure 3.12 (left)) are summarized in Table 3.1. One can observe first in this table that the most efficient method, with almost 100% of correct solutions, is the cold-start strategy with the backwards check. The least efficient is the warm-start strategy alone. Nevertheless its percentage of correct solutions is not too bad with 64.44% in average. The warm-start & back method is far better than the warm-start method with 88.19% of correct solutions in average.

To understand which points $\mathbf{b}^n$ on the zigzag trajectory induce incorrect solutions for each method, the solutions in the case $\bar{n} = 20$ are presented in Figure 3.13. For the warm-start method most of the incorrect solutions are situated in the upper part of the phase diagram. Indeed once the last activation occurs, all the forthcoming vectors $\mathbf{b}^n$ on the trajectory are in the state with one inactive constraint and are encouraged to remain



(a) (b) (c)

(d) (e) (f)

Figure 3.13: Solutions of the $\mathbf{b}^n$, $n = 1, \ldots, \frac{\bar{n}(\bar{n}+1)}{2}$ with $\bar{n} = 20$ for the method warm-start, warm-start & back and cold-start. Graphs (a)-(b)-(c): vectors $\mathbf{b}^n$ on the phase diagram stemmed from the vapor-liquid equilibrium data with the colored repartition: 🟢 when one inequality constraint is inactive, 🔵 when two inequality constraints are inactive and 🔴 when all inequality constraints are inactive. Graphs (d)-(e)-(f): vectors $\mathbf{b}^n$ on the phase diagram stemmed from the vapor-liquid equilibrium data with the colored repartition: 🔵 when the solution is correct, 🔴 otherwise.

in this state with the warm-start. The warm-start method seems also to have difficulties for the $\mathbf{b}^n$ situated in area 3. The solutions have a tendency to stay in the state with two inactive constraints and the deactivation is often missed. If the backwards check is added to the method, then most of the wrong solutions in the area 3 are corrected. The remaining errors are situated near the boundary of area 3 which is a region that generates numerical difficulties (the algorithm has a tendency to activate the inequality constraints). Furthermore in the upper part of the phase diagram, the repartition between the state with one or two inactive constraints seems to be correct. However regarding the Figure 3.13 (e), most of these points are considered as wrong even if they are situated in an area 2. The reason is that the single-phase points generating the phase simplex are incorrect. With regards to the cold-start method, one can observe that the incorrect solutions are located in the neighborhood of the phase boundaries which are difficult regions for the optimization.

The numerical results for the energy function g defined from the liquid-liquid equilibrium data are presented in Table 3.2 and Figures 3.14 and 3.15. This example is more difficult to compute than the latest because two convex areas are in common. In this case the convex areas $\Delta_{2,1}$ and $\Delta_{2,2}$ are common. Suppose that the constraint 1 is inactive and 2 is active. The associated vector $\mathbf{x}_2$ tends to converge to $\mathbf{x}_1$ instead of encouraging the deactivation of y_2 since $\mathbf{x}_1$ and $\mathbf{x}_2$ belong to the same convex area.

Regarding the results of Table 3.2 the observation is that the warm-start & back method is the best and the cold-start & back method makes errors. The warm-start method is still the less efficient and the cold-start gives similar results as cold-start & back.

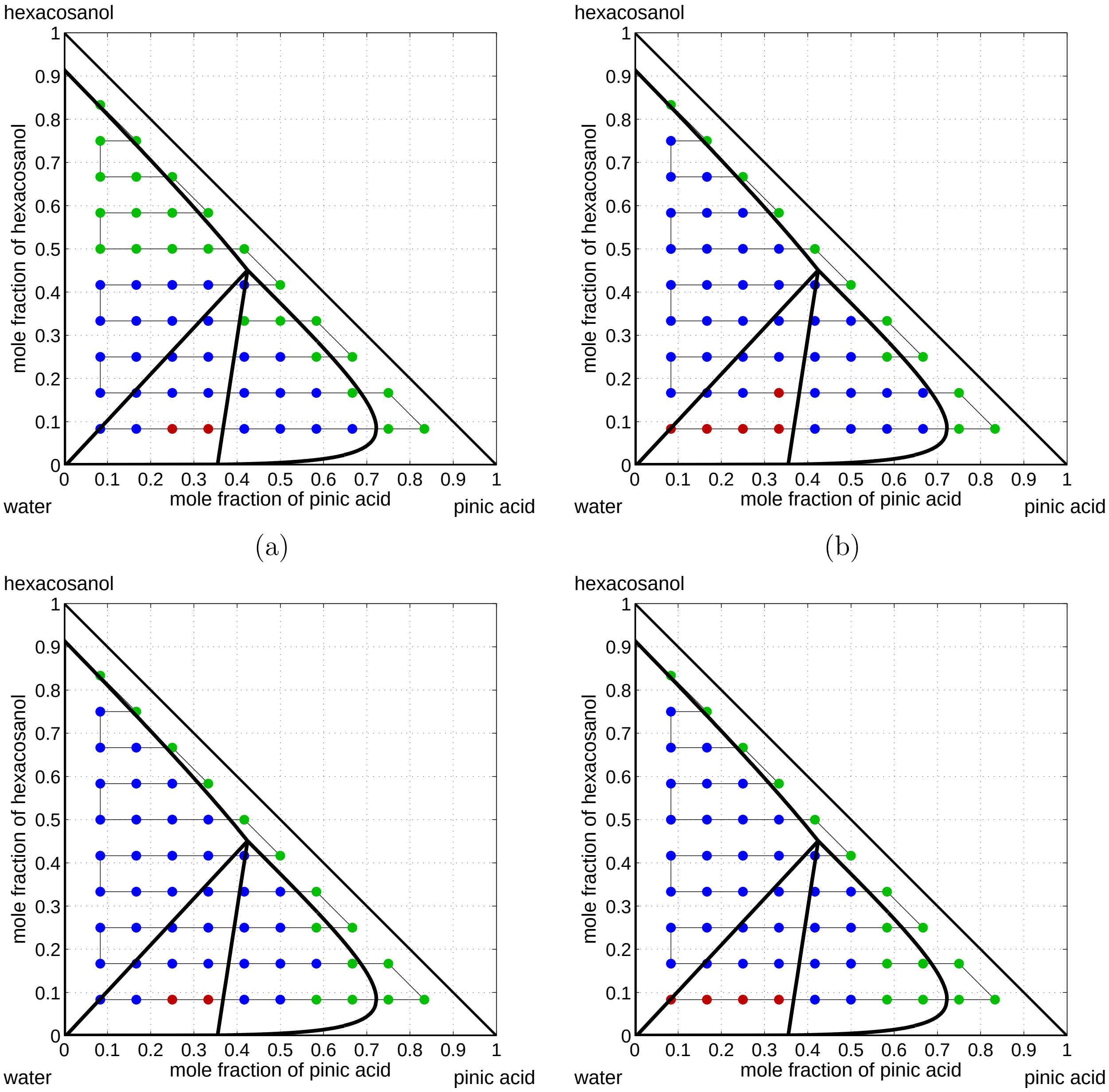
Let us consider the representations of Figures 3.14 and 3.15 to understand the differences with the example VL and where are located on the zigzag trajectory the $\mathbf{b}^n$ for which the solution of the PEP is wrong. The case $\bar{n} = 10$ is represented on both figures. Figure 3.14 illustrates the solution of the successive PEP for each method. Figure 3.15 shows the correctness of the successive solutions with the colored repartition: blue circle when the solution is correct and a red circle otherwise.

For the four methods the solution of the PEP defined at $\mathbf{b}^n$ located in the area 3 are almost all incorrect. Generally these wrong solutions remain with two inactive constraints and the deactivation of the last active constraint is missed. For the warm-start method, the other wrong solutions are situated near the phase boundaries, the solution remaining a single-phase point instead to deactive an active constraint.

For the cold-start and cold-start & back methods, the other wrong solutions are located near the phase boundary between the right area 2 and the common area 1. For the vectors $\mathbf{b}^n$ lying in this region the primal-dual interior-point method has a tendency to encourage the vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ to converge to $\mathbf{b}^n$ if their initialization is badly chosen, meaning too far from the phase simplex at $\mathbf{b}^n$. Consequently the solution is a single-phase point and the deactivation is missed. This situation is illustrated in Figure 3.16 (left). In the warm-start & back method the initialization of the interior-point method is well chosen since the initial value of $\mathbf{x}_1$ and $\mathbf{x}_2$ are close to the phase simplex defined at $\mathbf{b}^n$ (see Figure 3.16 (right)). Then $\mathbf{x}_1$ and $\mathbf{x}_2$ converge to the phase simplex of dimension 2.

		number of correct solutions	% of correct solutions	mean number of it. for i.p.m.	CPU time [s]
n=10	warm-start	32	58.18	31.42	0.11
	warm-start & back	47	85.45	29.30	0.15
	cold-start	42	76.36	66.95	0.22
	cold-start & back	43	78.18	65.69	0.29
n=20	warm-start	124	59.05	30.00	0.38
	warm-start & back	144	68.57	29.05	0.52
	cold-start	145	69.05	68.03	0.83
	cold-start & back	135	64.29	68.05	0.97
n=40	warm-start	497	60.61	29.14	1.44
	warm-start & back	753	91.83	30.33	2.16
	cold-start	657	80.12	68.11	3.38
	cold-start & back	697	85.00	68.11	3.93
n=80	warm-start	1970	60.80	29.78	5.70
	warm-start & back	2979	91.94	29.79	7.94
	cold-start	2591	79.97	68.06	12.68
	cold-start & back	2744	84.69	68.04	16.31

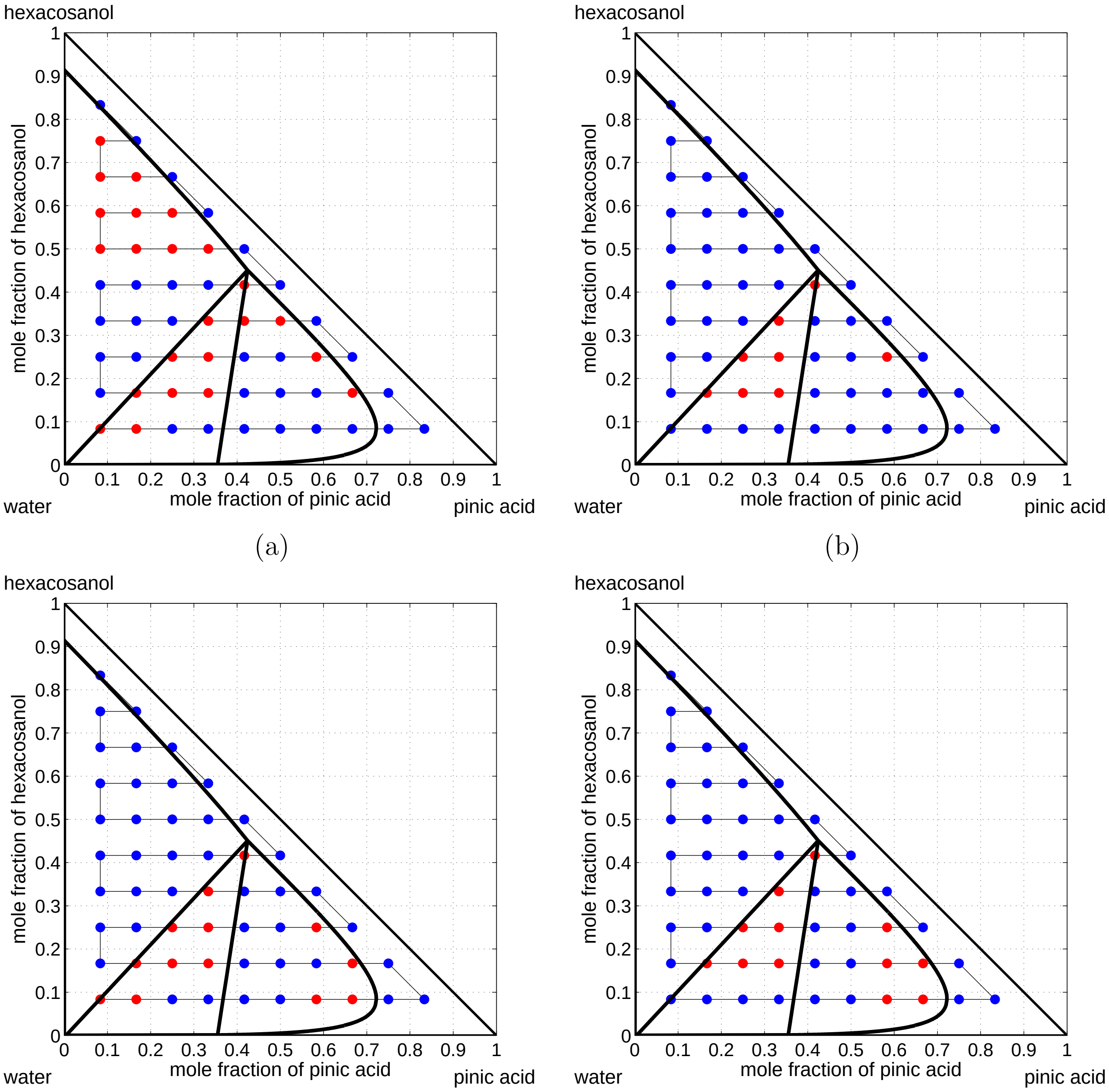
Table 3.2: Number of correct solutions, percentage of correct solutions, mean number of iterations for the primal-dual interior-point method (i.p.m.) and CPU time for the phase diagram with interaction parameters stemmed from the liquid-liquid equilibrium data of [64].



(a) (b)

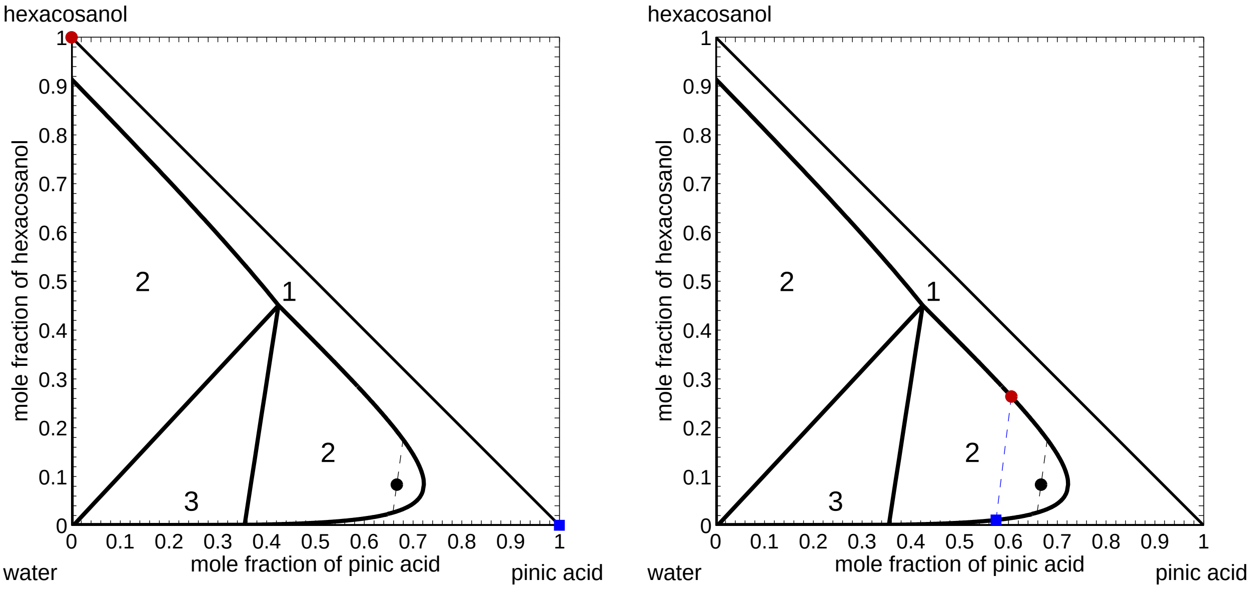
(c) (d)

Figure 3.14: Solutions of the $\mathbf{b}^n$, $n = 1, \ldots, \frac{\bar{n}(\bar{n}+1)}{2}$ with $\bar{n} = 10$ for the methods warm-start (a), warm-start & back (b), cold-start (c) and cold-start & back (d). The vectors $\mathbf{b}^n$ on the phase diagram stemmed from the vapor-liquid equilibrium data follow the colored repartition: ● when one inequality constraint is inactive, ● when two inequality constraints are inactive and ● when all inequality constraints are inactive.



(a) (b) (c) (d)

Figure 3.15: Solutions of the $\mathbf{b}^n$, $n = 1, \ldots, \frac{\bar{n}(\bar{n}+1)}{2}$ with $\bar{n} = 10$ for the methods warm-start (a), warm-start & back (b), cold-start (c) and cold-start & back (d). The vectors $\mathbf{b}^n$ on the phase diagram stemmed from the vapor-liquid equilibrium data follow the colored repartition: ● when the solution is correct, ● otherwise.



(a)

(b)

Figure 3.16: Cold-start (a) and warm-start (b) strategies in the initialization of the interior-point method for a composition-vector $\mathbf{b}^n$ (black circle) located near the phase boundary between area 2 and area 1. The initialization of $\mathbf{x}_1$ and $\mathbf{x}_2$ are respectively represented by a blue square and a red circle. The phase simplex at $\mathbf{b}^n$ is depicted with a dotted black line.

In conclusion the warm-start & back method has proved to be competitive in the example VL and the best method for the example LL. Let us observe now the computational cost in using this method.

First from Tables 3.1 and 3.2 the mean number of iterations for the primal-dual interior-point method is 67.87 for the cold-start & back method and 27.16 for the warm-start & back method in average. Hence the warm-start strategy allows to decrease the number of iterations for the interior-point method by 60%. This gain is also represented in the difference between the CPU times of the warm-start & back and cold-start & back methods. Although the CPU times for both examples and four methods are very small. The maximal time is equal to 16.31 s which is quite fast.

In Figure 3.17 (a) and (b), the CPU time versus $\bar{n}$ is plotted in the log-log scale for both examples. The plots have a second order behaviour as $\bar{n}$ increases and one can remark that the methods using the cold-start strategy are slower than the others. Furthermore by observing the small difference between the cold-start and cold-start & back methods, and warm-start and warm-start & back methods, one can conclude that the backwards check is not time-consuming. In Figure 3.17 (c) and (d), the normalized discrepancy in CPU times defined by $|\frac{t-t_B}{t_B}|$ where t is the CPU time for the methods without the backwards check and t_B is the CPU time with the backwards check, is plotted. This discrepancy is almost constant and one can deduce that the backwards check is responsible for 16.28% of the CPU times for the cold-start strategy, and 23.63% of the CPU time for the warm-start strategy.

Examples on the phase diagram VL

In a second part, we perform simulations on the example VL whose phase diagram is classical and depicted in Figure 3.12 (left).

Numerical results show the time evolution of the variables y_α, $\alpha = 1, 2, 3$, $\boldsymbol{\lambda}$, $\mathbf{b}$, $\mathbf{c}_g^\infty$ and $\mathbf{c}_g^{surf}$ along the time. For inequality constraints y_α with $\alpha = 1, 2, 3$, or equivalently for the number of moles in each liquid phase, the colored repartition follows the color code for $\mathbf{x}_\alpha$, $\alpha = 1, 2, 3$, namely

- ■: the inequality constraint y_1,
- ●: the inequality constraint y_2,
- ◆: the inequality constraint y_3.

With regards to the vectors $\mathbf{b}$, $\boldsymbol{\lambda}$, $\mathbf{c}_g^\infty$ and $\mathbf{c}_g^{surf}$, their components are associated to the chemical species present in the aerosol. For this section, the colored repartition is defined by

pinic acid (◆), hexacosanol (■) and water (●).

The first example has the following initial conditions

- composition-vector: $\mathbf{b}_0^T = (1.5 \cdot 10^{-8},\ 8.0 \cdot 10^{-8},\ 0.5 \cdot 10^{-8})$ mol,
- gas concentration-vector: $\mathbf{c}_{g,0}^{\infty,T} = (3.5,\ 3,\ 3.5)$ mol/m^3,
- time step: $h = 0.015$ s.

Hence the initial composition-vector $\mathbf{b}_0$ is located in the area 1 associated to the constraint 2 and the initial gas concentration-vector lies in the area 2 on the right. Figure 3.18 represents respectively the time evolution of the approximations $\mathbf{b}^n$, $n = 0, 1, 2, \ldots$, the associated phase simplices and $\mathbf{x}_\alpha^n$, $n = 0, 1, 2, \ldots$, $\alpha = 1, 2, 3$, on the phase diagram of the example VL. One can see that the simulation encounters first the deactivation of the constraint $y_3 = 0$, then the deactivation of the constraint $y_1 = 0$, and finally the activation of $y_3 > 0$.

Thanks to the colored repartition one can observe in Figure 3.18 (left) that the phase equilibrium associated to each $\mathbf{b}^n$, $n = 0, 1, 2, \ldots$, follows the phase diagram. Moreover the Figure 3.18 (middle) shows that the successive phase simplices computed for each $\mathbf{b}^n$ coincide with the phase boundaries on the phase diagram. The last figure in Figure 3.19 (right) presents the time evolution of the vector $\mathbf{x}_\alpha^n$, $\alpha = 1, 2, 3$. For a better view of the evolution of $\mathbf{x}_1^n$ and $\mathbf{x}_3^n$ a zoomed-in view is proposed in Figure 3.19.

The constraint $y_2^n > 0$ is inactive during all the simulation. Hence $\mathbf{x}_2^n$ belongs to the successive phase simplices of the simulation. When $\mathbf{b}^n$ is a single-phase point, $\mathbf{b}^n = y_2^n \mathbf{x}_2^n$,

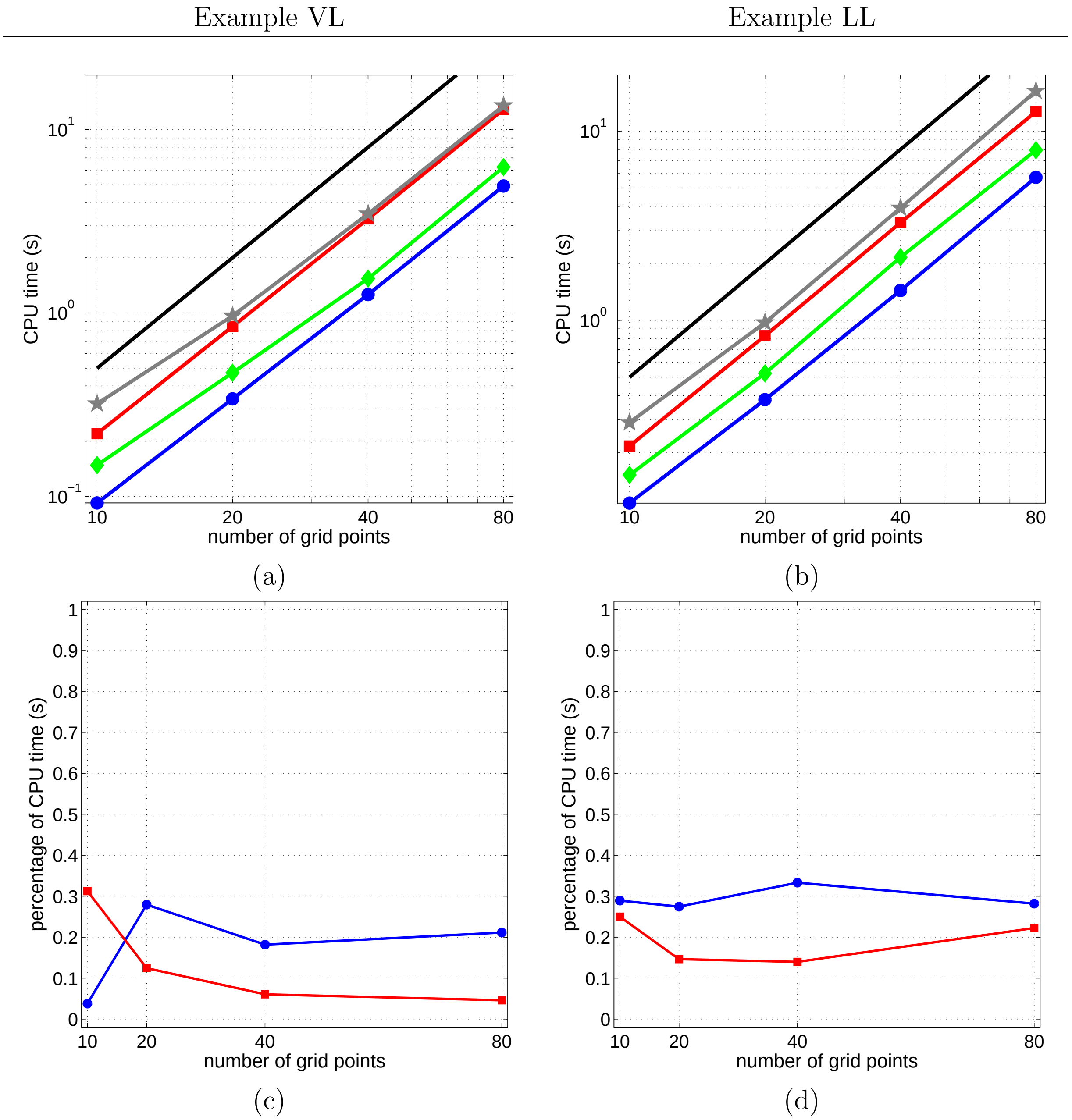


Figure 3.17: CPU times for the computation of the 4 methods on both examples. (a)-(b): CPU times in log-log scale with (—) for the function $\bar{n}^2$, (—●—) for the warm-start method, (—◆—) for the warm-start & back method, (—■—) for the cold-start method and (—★—) for the cold-start & back method. (c)-(d): normalized discrepancy between the CPU times of the warm-start methods with and without the backwards check (— ● —), and of the cold-start methods with and without the backwards check (—■—).

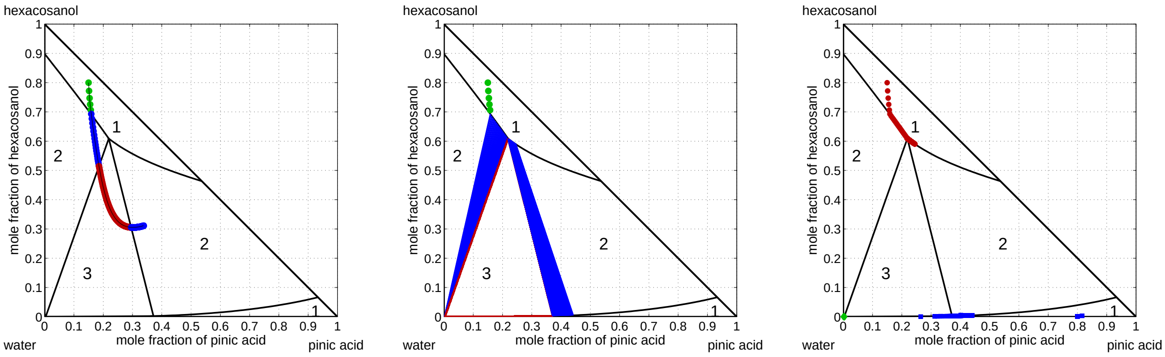


Figure 3.18: Example VL with the initial conditions $\mathbf{b}_0^T = (1.5 \cdot 10^{-8}, 8.0 \cdot 10^{-8}, 0.5 \cdot 10^{-8})$ mol and $\mathbf{c}_{g,0}^{\infty,T} = (3.5,\ 3,\ 3.5)$ mol/m^3. Left: time evolution of $\mathbf{b}^n$, $n = 0, 1, 2, \dots$ on the phase diagram. Middle: time evolution of the phase simplices on the phase diagram. Right: time evolution of the single-phase points $\mathbf{x}_\alpha^n$, $\alpha = 1, 2, 3$ and $n = 0, 1, 2, \dots$, on the phase diagram.

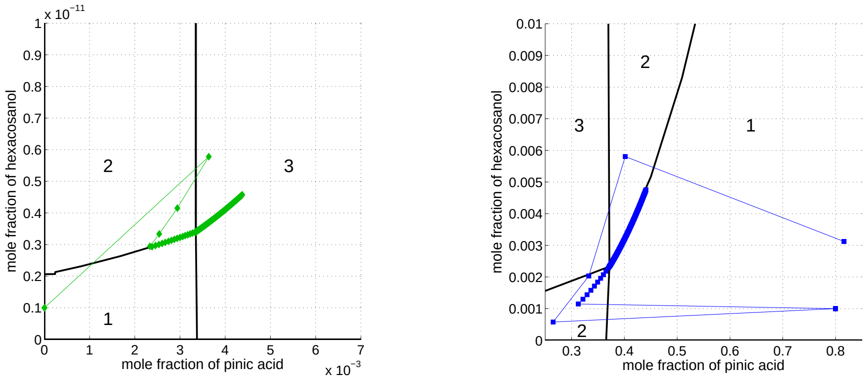
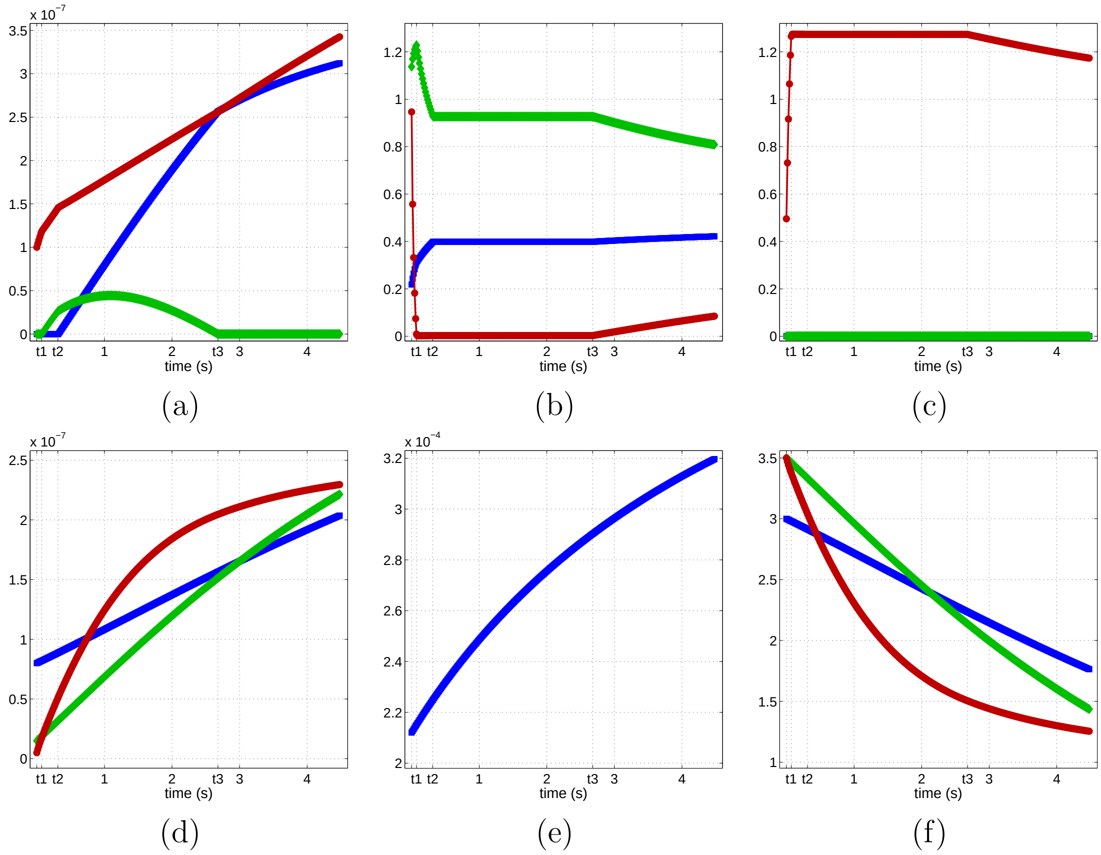


Figure 3.19: Example VL with the initial conditions $\mathbf{b}_0^T = (1.5 \cdot 10^{-8}, 8.0 \cdot 10^{-8}, 0.5 \cdot 10^{-8})$ mol and $\mathbf{c}_{g,0}^{\infty,T} = (3.5,\ 3,\ 3.5)$ mol/m^3. Zoomed-in view of the time evolution of $\mathbf{x}_3^n$ (left) and $\mathbf{x}_1^n$ (right), $n = 0, 1, 2, \dots$, on the phase diagram.

and $\mathbf{x}_2^n$ and $\mathbf{b}^n$ are equal on the phase diagram (since the normalized values are plotted on the phase diagram). Once the number of inactive constraints at the equilibrium is greater than 2, $\mathbf{x}_2^n$ follows the phase boundaries that lie on the boundary of $\Delta_{2,2}$. The vectors $\mathbf{x}_2^n$ on the phase diagram of Figure 3.18 (right) follows this behaviour.

With regards to $\mathbf{x}_1^n$ and $\mathbf{x}_3^n$, both associated constraints are active at $t = 0$ s. As long as these constraints are active, the vectors $\mathbf{x}_1^n$ and $\mathbf{x}_3^n$, $n = 0, 1, 2, \dots$ are defined as the minimizer of the distance between the supporting tangent hyperplane and $(\mathbf{x}_i^n, g(\mathbf{x}_i^n))$, $i = 1, 3$. When $\mathbf{b}^n$ comes closer to the deactivation point, the vectors $\mathbf{x}_1^n$ and $\mathbf{x}_3^n$ tend to the discontinuity located on the phase boundaries. This situation is observable in Figure 3.19.

Figure 3.20 shows the time evolution of y_α^n, $\alpha = 1, 2, 3$, $\boldsymbol{\lambda}^n$, $\mathbf{c}_g^{surf,n}$, $\mathbf{b}^n$, R^n and $\mathbf{c}_g^{\infty,n}$. The evolution of y_α^n, $\alpha = 1, 2, 3$, is illustrated on the graph (a). The time evolution indicates that initially $y_2^0 > 0$ and $y_1^0 = y_3^0 = 0$ and then at $t^1 = 0.0717$ s, y_3^n becomes positive, at $t^2 = 0.3181$ s, y_1^n becomes positive and finally at $t^3 = 2.68$ s, y_3^n is null again. This evolution follows the phase diagram exactly and one can observe the discontinuity in y_α^n, $\alpha = 1, 2, 3$ at t^1, t^2 and t^3. These discontinuities are also present in the time evolution of $\boldsymbol{\lambda}^n$ (graph (b)) and $\mathbf{c}_g^{surf,n}$ (graph (c)).



(a) (b) (c)

(d) (e) (f)

Figure 3.20: Example VL with the initial conditions $\mathbf{b}_0^T = (1.5 \cdot 10^{-8}, 8.0 \cdot 10^{-8}, 0.5 \cdot 10^{-8})$ mol and $\mathbf{c}_{g,0}^{\infty,T} = (3.5,\ 3,\ 3.5)$ mol/m^3. Time evolution of (a) the number of moles y_α^n of each liquid phase $\alpha = 1, 2, 3$; (b) the components of the Lagrange multiplier $\boldsymbol{\lambda}^n$; (c) the gas concentration of each species at the surface of the aerosol $c_{g,i}^{surf,n}$, $i = 1, 2, 3$; (d) the composition of each species b_i^n, $i = 1, 2, 3$; (e) the radius of the aerosol and (f) the gas concentration of each species far from the particle $c_{g,i}^{\infty,n}$, $i = 1, 2, 3$.

On graphs (b) and (c) one can observe furthermore that $\boldsymbol{\lambda}^n$ and $\mathbf{c}_g^{surf,n}$ are constant as all 3 inequality constraints are inactive. This fact illustrates the relation $\boldsymbol{\lambda} = -\boldsymbol{\nabla} g(\mathbf{x}_\alpha)$, $\forall \alpha \in \mathcal{I}$ and these variables are constant in area 3 since $\mathbf{x}_\alpha^{\mathcal{I}}$ are constant in this area.

The plots on graphs (d)-(f) of Figure 3.20 represent the time evolution of $\mathbf{b}^n$, R^n and $\mathbf{c}_g^{\infty,n}$. These evolutions do not lose their regularity at t^1, t^2 and t^3. Furthermore the evolution of $\mathbf{b}^n$ is complementary with the time evolution of $\mathbf{c}_g^{\infty,n}$. Hence the total mass in the system is conserved. Thanks to the evolution of the radius, one knows that the aerosol particle is growing on this example.

The simulation presented in Figure 3.21 shows a trajectory of $\mathbf{b}^n$, $n = 0, 1, 2, \ldots$ that evolves near the phase boundary between the areas 1 and 2. The initial conditions of this example are

- composition-vector: $\mathbf{b}_0^T = (2.8 \cdot 10^{-8},\ 7.0 \cdot 10^{-8},\ 0.2 \cdot 10^{-8})$ mol,
- gas concentration-vector: $\mathbf{c}_{g,0}^{\infty,T} = (0.02,\ 7.3,\ 1.68)$ mol/m^3,
- time step: $h = 0.015$ s.

The representations in Figure 3.21(a)-(b) show a time evolution of $\mathbf{b}^n$, $n = 0, 1, 2 \ldots$ and the corresponding phase simplices that follow the phase diagram. The time evolution of y_α^n, $\alpha = 1, 2, 3$, $n = 0, 1, 2 \ldots$, depicted in Figure 3.21 (c) emphasizes the difficulty of this example by the weak variation in y_3^n during the time interval of the deactivation of this latter (namely between the time $t^1 = 0.6173$ s and $t^2 = 2.2663$ s.)

Figure 3.21(d)-(e)-(f) are zoomed-in view of the trajectory $\mathbf{b}^n$, $n = 0, 1, 2, \ldots$ near the phase boundary between the areas 1 and 2. On these graphs one can observe that the phases are correctly computed, as shown by the correct color code in the Figure. Furthermore the computed discontinuity points are very close to the phase boundary and thus good approximations. Finally in Figure 3.21(e) the fractional time step needed to reach the discontinuity point is clearly depicted.

The last example for the phase diagram VL is illustrated in Figure 3.22 in order to show the robustness of the algorithm in extreme situations. This example has the particularity to be subject to several activations/deactivations in a short time interval by crossing the region on the phase diagram that is at the intersection of areas 1, 2 and 3. The initial conditions for this example are

- composition-vector: $\mathbf{b}_0^T = (1.5 \cdot 10^{-8},\ 8.0 \cdot 10^{-8},\ 0.5 \cdot 10^{-8})$ mol,
- gas concentration-vector: $\mathbf{c}_{g,0}^{\infty,T} = (2.85,\ 4.2,\ 1.95)$ mol/m^3,
- time step: $h = 0.015$ s.

The graphs (a), (b) and (c) in Figure 3.22 depict the time evolution of $\mathbf{b}^n$, $n = 0, 1, 2, \ldots$, the phase simplices and $\mathbf{x}_\alpha^n$, $n = 0, 1, 2, \ldots$, $\alpha = 1, 2, 3$ respectively. Hence a first deactivation occurs at $t^1 = 0.3672$ s, a second deactivation at $t^2 = 0.4633$ s immediately followed by an activation at $t^3 = 0.4889$ s. The time evolution of y_α^n, $\alpha = 1, 2, 3$ and $n = 0, 1, 2, \ldots$ is proposed in Figure 3.22 (d) with a zoomed-in view of y_3^n when this constraint is deactivated (Figure 3.22 (e)). The last picture in Figure 3.22 proposes a zoomed-in view of $\mathbf{b}^n$, $n = 0, 1, 2, \ldots$ in the vicinity of the intersection between the areas 1, 2 and 3. Let us note the perfect match of the colored repartition (and consequently of the successive phase simplices) in this region.

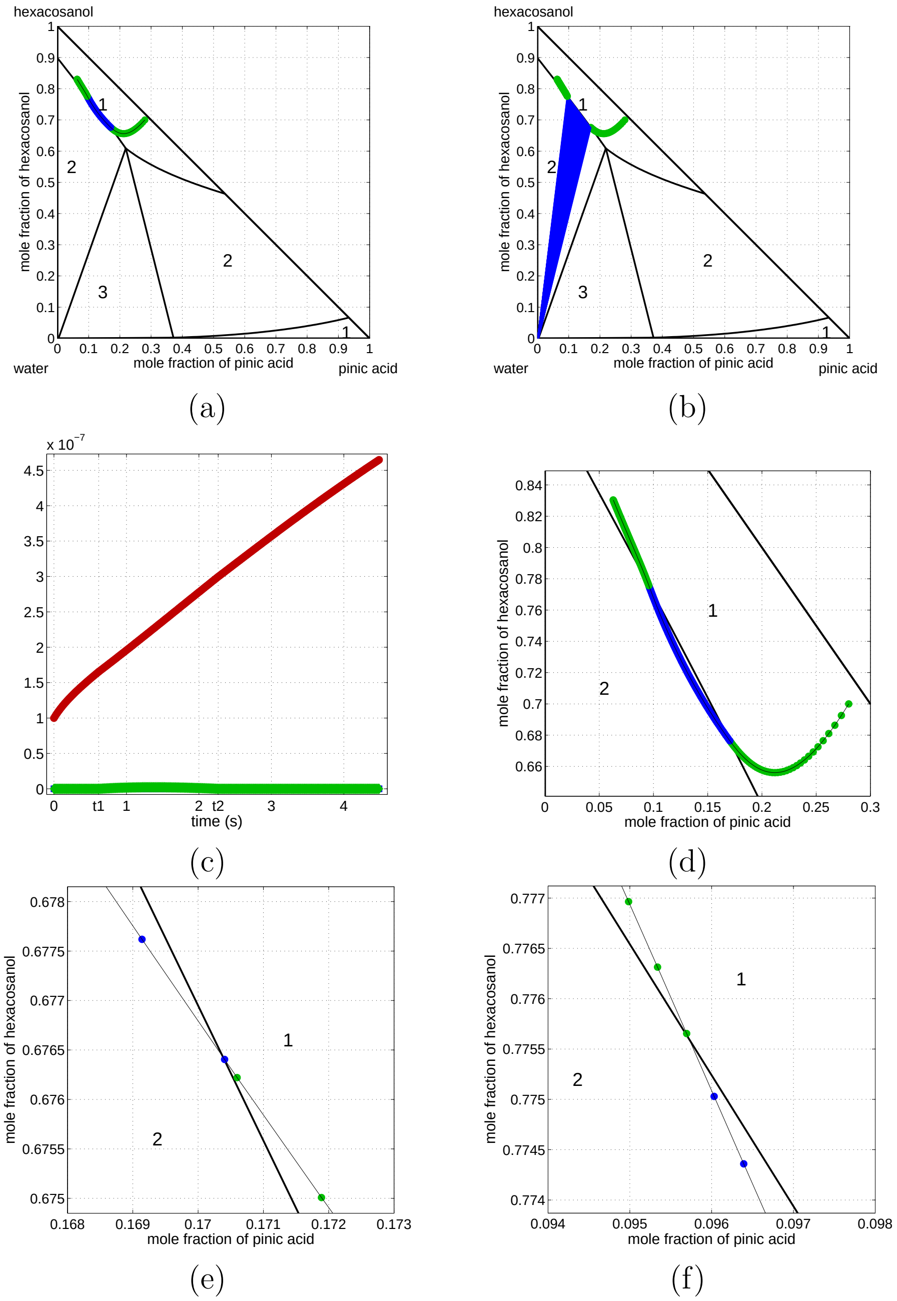
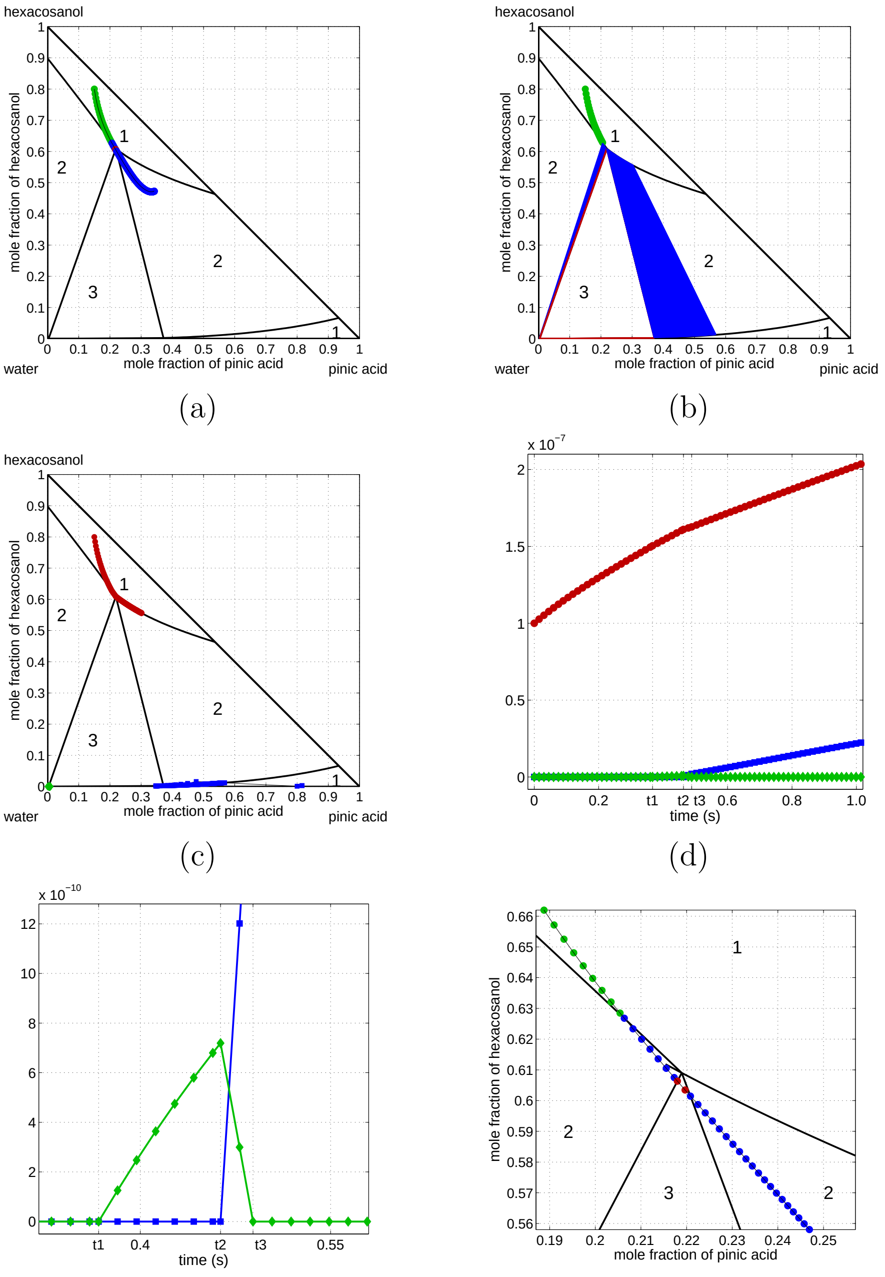


Figure 3.21: Example VL with the initial conditions $\mathbf{b}_0^T = (2.8 \cdot 10^{-8}, 7.0 \cdot 10^{-8}, 0.2 \cdot 10^{-8})$ mol and $\mathbf{c}_{g,0}^{\infty,T} = (0.02, 7.3, 1.68)$ mol/m^3. Time evolution of (a) $\mathbf{b}^n$, $n = 0, 1, 2, \ldots$ on the phase diagram, (b) the phase simplices on the phase diagram; (c) the number of moles y_α^n, $\alpha = 1, 2, 3$ and $n = 0, 1, 2, \ldots$. The graphs (d),(e) and (f) are a zoomed-in view of the trajectory of the $\mathbf{b}^n$, $n = 0, 1, 2, \ldots$ near the phase boundary between the areas 2 and 1, near the deactivation and near the activation respectively.



(a) (b)

(c) (d)

(e) (f)

Figure 3.22: Example VL with the initial conditions $\mathbf{b}_0^T = (1.5 \cdot 10^{-8},\, 8.0 \cdot 10^{-8},\, 0.5 \cdot 10^{-8})$ mol and $\mathbf{c}_{g,0}^{\infty,T} = (2.85,\, 4.2,\, 1.95)$ mol/m^3. Time evolution of (a) $\mathbf{b}^n$, $n = 0, 1, 2, \ldots$ on the phase diagram, (b) the phase simplices on the phase diagram; (c) the mole-fraction vectors $\mathbf{x}_\alpha^n$, $\alpha = 1, 2, 3$ and $n = 0, 1, 2, \ldots$, on the phase diagram; (d) the number of moles y_α^n, $\alpha = 1, 2, 3$ and $n = 0, 1, 2, \ldots$. The graphs (e) and (f) are respectively zoomed-in views of (e) the number of moles y_α^n, $\alpha = 1, 2, 3$ and $n = 0, 1, 2, \ldots$, and (f) the trajectory of $\mathbf{b}^n$, $n = 0, 1, 2, \ldots$.

This example and the previous one prove the efficiency and robustness of the optimization-based numerical method on a phase diagram of the classical type presented in Figure 1.4. In the next class of examples let us work on a non classical phase diagram.

Examples on the phase diagram LL

The first example considered on the non classical phase diagram LL has the initial conditions

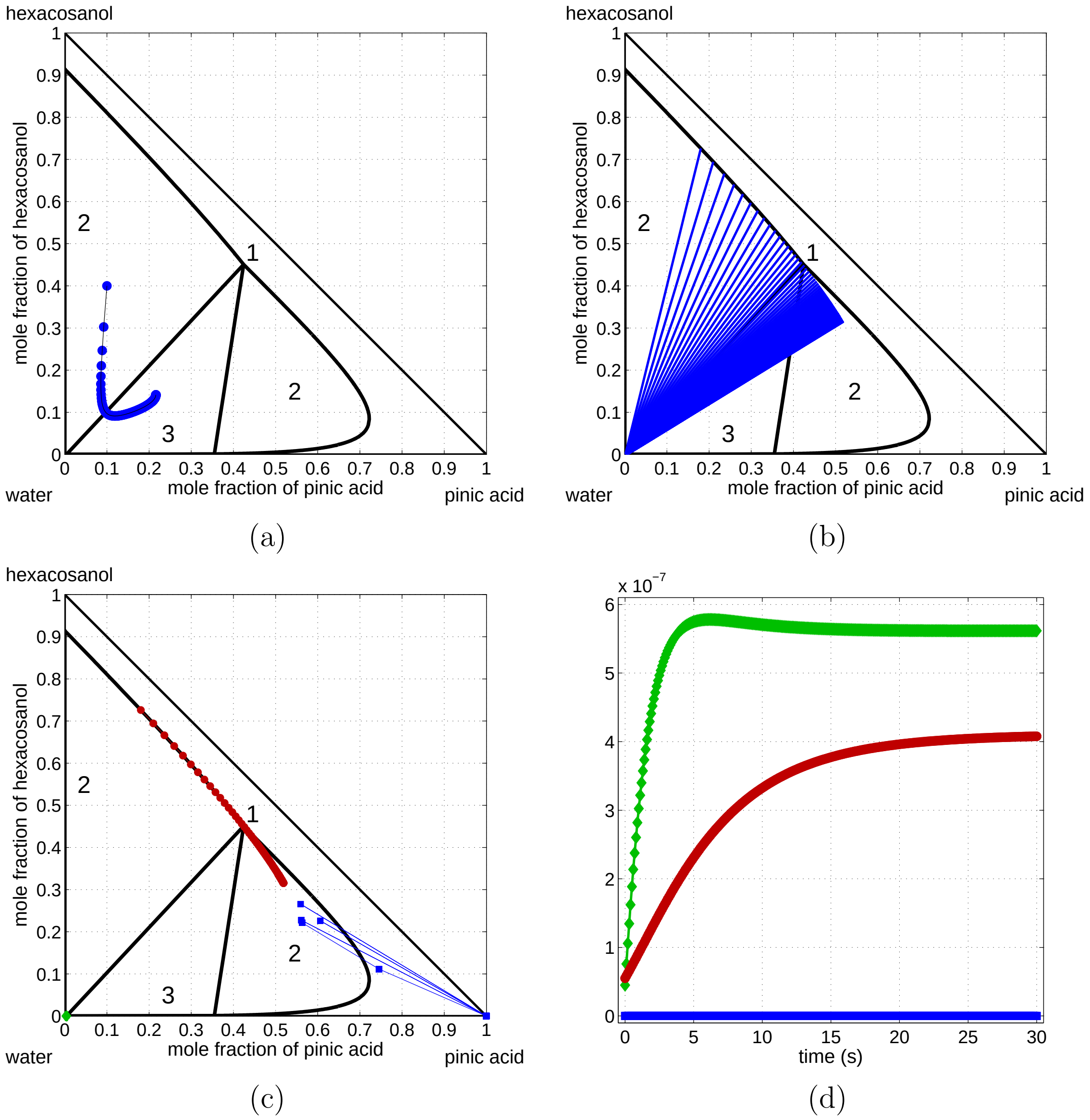
- composition-vector: $\mathbf{b}_0^T = (1.0 \cdot 10^{-8},\ 4.0 \cdot 10^{-8},\ 5. \cdot 10^{-8})$ mol,
- gas concentration-vector: $\mathbf{c}_{g,0}^{\infty,T} = (2.0,\ 1.0,\ 7.0)$ mol/m^3,
- time step: $h = 0.1$ s.

The results of the simulation are presented in Figure 3.23. The picture (a) in Figure 3.23 depicts the evolution of $\mathbf{b}^n$, $n = 0, 1, 2, \ldots$ on the phase diagram and one can remark that the simulation starts in the area 2 and then enters in the area 3 but the color code of the vectors $\mathbf{b}^n$, $n = 0, 1, 2, \ldots$ is still blue, meaning that the deactivation of the first constraint is not detected. This situation is confirmed in Figure 3.23 (b) where the time evolution of the phase simplices is depicted. Even if the time step is decreased, the deactivation is still missed and the solution bifurcates into a branch of local minima or saddle-points.

To understand why the deactivation is not detected, let us observe Figure 3.23 (c) which represents the time evolution of $\mathbf{x}_\alpha^n$, $\alpha = 1, 2, 3$ and $n = 0, 1, 2, \ldots$. The variable $\mathbf{x}_1$ should indicate the deactivation of the constraint $y_1 = 0$ by converging to the bottom left vertex of the area 3 (i.e. under the area 2 on the right) as $\mathbf{b}$ comes closer to the phase boundary between the area 2 on the left and the area 3. However Figure 3.23 (c) indicates that $\mathbf{x}_1^n$, $n = 0, 1, 2, \ldots$, tends to $\mathbf{x}_2^n$ as $\mathbf{b}^n$ tends to the deactivation point. The blue squares that represent the evolution of $\mathbf{x}_1^n$, $n = 0, 1, 2, \ldots$, are successively in the area 2 on the right, close to $\mathbf{x}_2^n$, and in the corner of the phase diagram. This alternance is justified by the Algorithm 3.2.1 and the remark 3.2.2. If the sequence built in the Algorithm 3.2.1 converges to $\mathbf{x}_2^n$, then $\mathbf{x}_1^n$ is set to the initial point $\mathbf{x}_1^0$ located in the bottom left corner of the phase diagram. At the next time step, the sequence starts from the corner and iterates until the Hessian is not positive definite. The point $\mathbf{x}_1^{n+1}$ is then set to the last correct iterate of the sequence that is located here in the area 2 on the right. For the next time step, the sequence starts from this point and converges to the inactive phase $\mathbf{x}_2^{n+2}$. The point $\mathbf{x}_1^{n+2}$ is then set to $\mathbf{x}_1^0$. Since the constraints 1 and 2 share the same convex areas, $\mathbf{x}_1^n$ always converges to $\mathbf{x}_2^n$ with this initialization. Let us change this initialization by a point located closer to the deactivation point

$$\mathbf{x}_1^{0,T} = (0.7,\ 3.0 \cdot 10^{-8},\ 0.29999997).$$

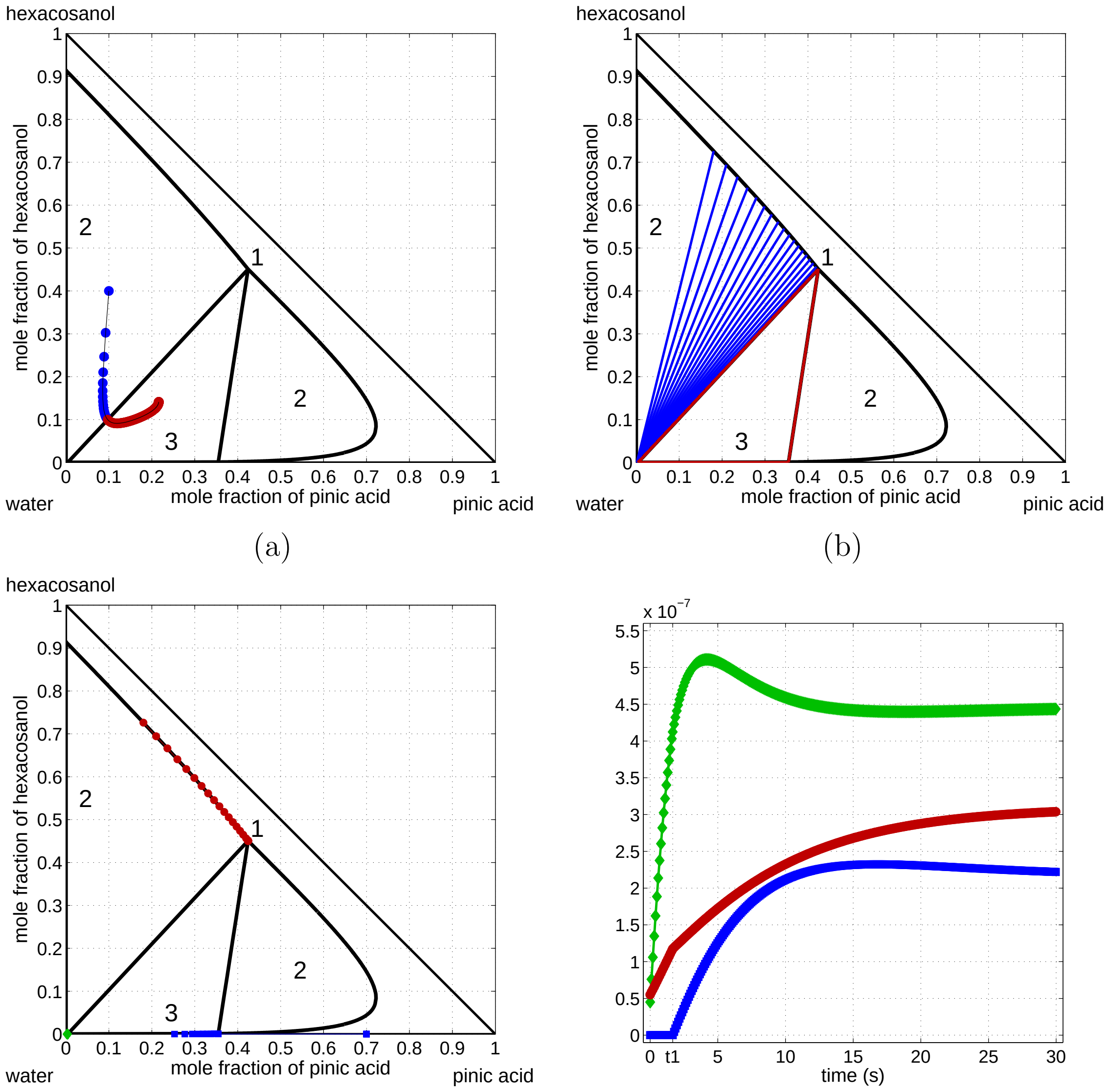
The results of the simulation with this new initialization is illustrated in Figure 3.24. In this case the deactivation is detected at $t^1 = 1.6684\,[s]$ and correctly computed: the color code is respected and the phase simplices follow the phase boundaries of the phase diagram.



(a) (b) (c) (d)

Figure 3.23: Example LL with the initial conditions $\mathbf{b}_0^T = (1.0{\cdot}10^{-8},\ 4.0{\cdot}10^{-8},\ 5.{\cdot}10^{-8})$ mol and $\mathbf{c}_{g,0}^{\infty,T} = (2.0,\ 1.0,\ 7.0)$ mol/m^3 when $\mathbf{x}_\alpha^0$, $\alpha = 1, 2, 3$ are initialized in the corner of the phase diagram. Time evolution of (a) $\mathbf{b}^n$, $n = 0, 1, 2, \ldots$; (b) the phase simplices; (c) the molar-fraction vectors $\mathbf{x}_\alpha^n$, $\alpha = 1, 2, 3$ and $n = 0, 1, 2, \ldots$; (d) the number of moles y_α^n, $\alpha = 1, 2, 3$ and $n = 0, 1, 2, \ldots$.

Finally let us compare the plots of Figure 3.23 (d) and Figure 3.24 (d) that represent the time evolution of y_α^n, $\alpha = 1, 2, 3$ and $n = 0, 1, 2, \ldots$. On the first plot, only y_2^n and y_3^n, $n = 0, 1, 2, \ldots$ evolve whereas on the second plot all variables evolve in time. Furthermore let us remark that the deactivation of $y_1^n = 0$ has an influence on the evolution of y_2^n and



(a) (b)

(c) (d)

Figure 3.24: Example LL with the initial conditions $\mathbf{b}_0^T = (1.0\cdot10^{-8},\ 4.0\cdot10^{-8},\ 5.\cdot10^{-8})$ mol and $\mathbf{c}_{g,0}^{\infty,T} = (2.0,\ 1.0,\ 7.0)$ mol/m^3 when $\mathbf{x}_3^0$ is initialized closer to the bottom left vertex of the area 3. Time evolution of (a) $\mathbf{b}^n$, $n = 0, 1, 2, \ldots$; (b) the phase simplices; (c) the molar-fraction vectors $\mathbf{x}_\alpha^n$, $\alpha = 1, 2, 3$ and $n = 0, 1, 2, \ldots$; (d) the number of moles y_α^n, $\alpha = 1, 2, 3$ and $n = 0, 1, 2, \ldots$.

y_3^n. In addition to the loss of regularity at t^1, these two variables have a different evolution from the plot of Figure 3.23 (d). Consequently the deactivation of the constraint influences the whole system.

The next examples on the non classical phase diagram LL are concerned with simulation that starts in the common area 1 and enters in the area 2 located on the right. The initial conditions for the first example are

- composition-vector: $\mathbf{b}_0^T = (9.0 \cdot 10^{-8},\ 0.5 \cdot 10^{-8},\ 0.5 \cdot 10^{-8})$ mol,
- gas concentration-vector: $\mathbf{c}_{g,0}^{\infty,T} = (4.0,\ 2.0,\ 4.0)$ mol/m^3,
- time step: $h = 0.001$ s.

In this example the deactivation is detected in the time interval [0.167, 0.168]. The computation of the deactivation time indicates that the time interval is not correct since the computed fractional time step is negative. The algorithm returns back over 14 time steps and the new time interval is then defined by [0.154, 0.155]. In this interval, the computation of the discontinuity time succeeds. The discontinuity time is equal to $t^{n+1} = 0.154106$ s and the discontinuity point is given by

$$\mathbf{b}^{T,n} = (9.72 \cdot 10^{-8},\ 0.72 \cdot 10^{-8},\ 3.26 \cdot 10^{-8}).$$

The next step in the numerical resolution is the computation of the other variables at the discontinuity time and then the restart of the simulation with the new number of inactive constraints. This procedure fails for this example. The reason is that the interior-point method returns a solution that is a single-phase point at t^{n+1}. Furthermore even if the second inequality constraint is artificially set to inactive, the interior-point method continues to return single-phase solution at t^{n+2}, preventing the simulation from restarting.

A representation of this example is given in Figure 3.25 where the time evolution of $\mathbf{b}^n$, $n = 0, 1, 2, \ldots$ is illustrated until the deactivation. The zoomed-in view in Figure 3.25 (right) shows that the computed deactivation time and point are good approximations. Observing the trajectory of $\mathbf{b}^n$, $n = 0, 1, 2, \ldots$, on the phase diagram, one can understand the difficulty for the algorithm to restart from the discontinuity point: the trajectory hits the phase boundary nearly perpendicularly to the phase simplices defined in area 2, and the phase simplex at the discontinuity point is in fact given by $\mathbf{x}_1 = \mathbf{x}_2$. Yet no solution exists to enforce the simulation to restart.

The last example on the phase diagram LL illustrates another simulation that starts in area 1 and enters in area 2. Unlike in the previous example, the trajectory of $\mathbf{b}^n$, $n = 0, 1, 2, \ldots$ does not come perpendicularly to the phase simplices defined in area 2. The situation is illustrated in Figure 3.26. The initial conditions are

- composition-vector: $\mathbf{b}_0^T = (7.5 \cdot 10^{-8},\ 2.0 \cdot 10^{-8},\ 0.5 \cdot 10^{-8})$ mol,
- gas concentration-vector: $\mathbf{c}_{g,0}^{\infty,T} = (3.0,\ 3.0,\ 4.0)$ mol/m^3,
- time step: $h = 0.001$ s.

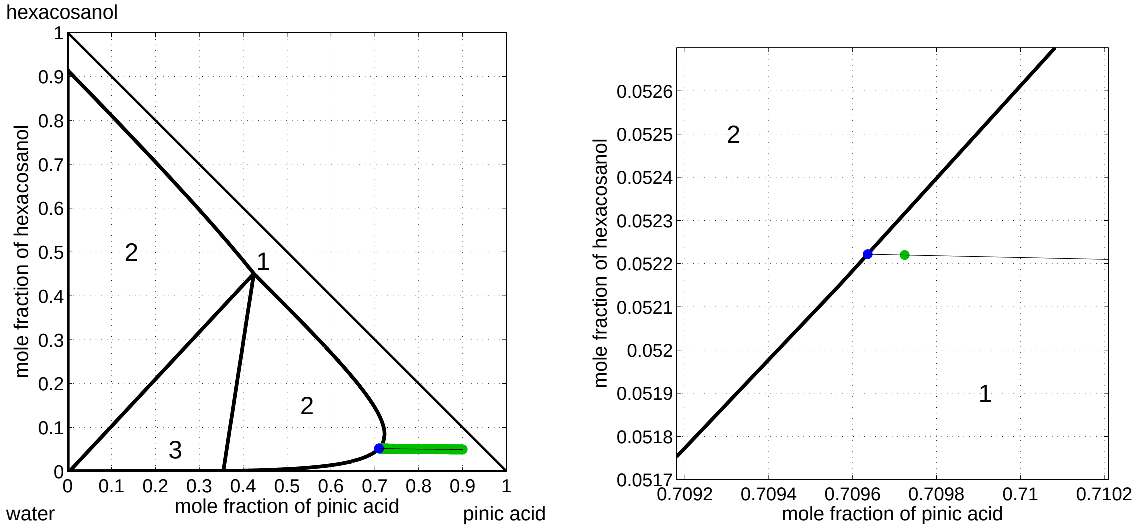
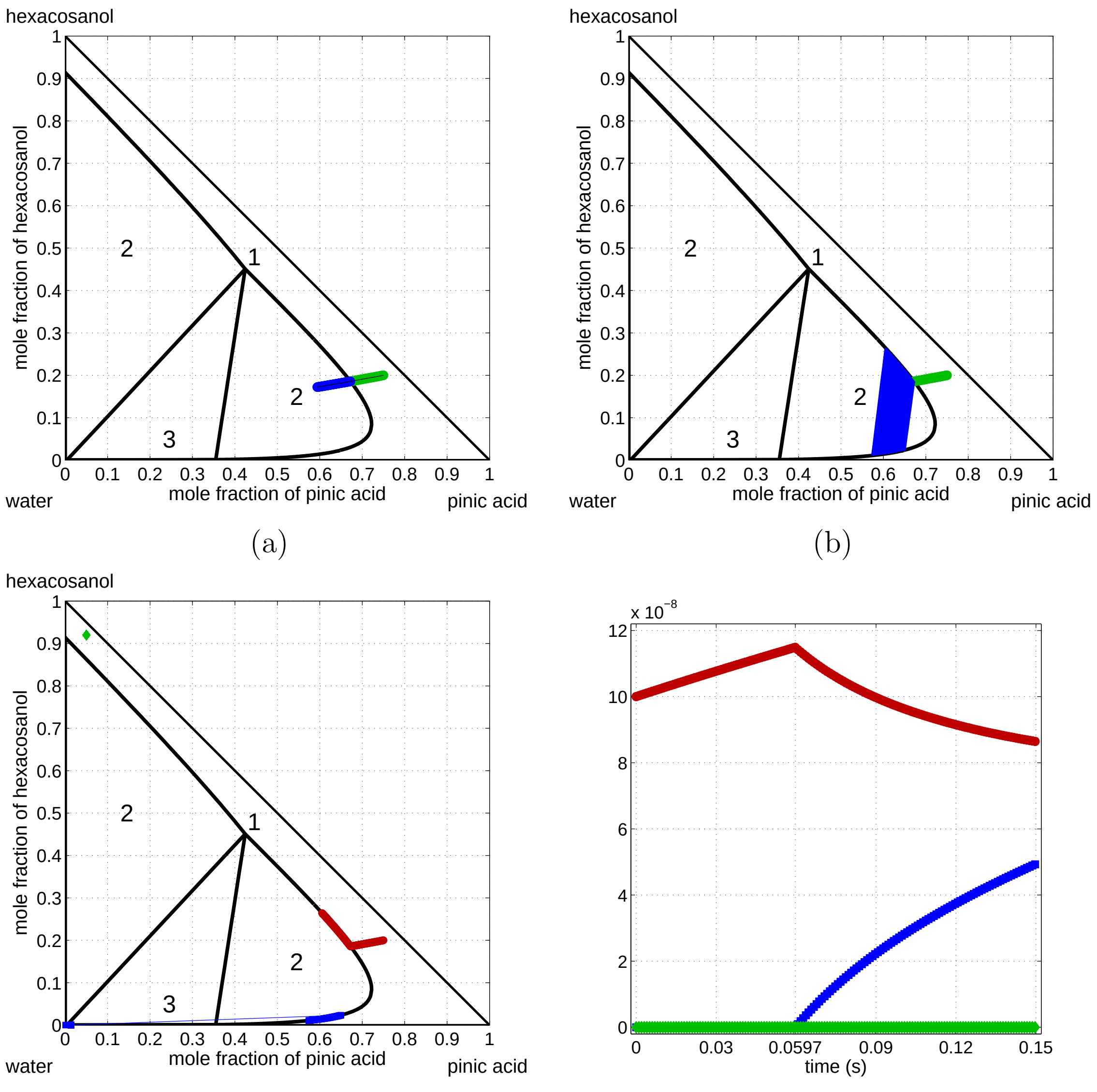


Figure 3.25: Example LL with the initial conditions $\mathbf{b}_0^T = (9.0{\cdot}10^{-8},\ 0.5{\cdot}10^{-8},\ 0.5{\cdot}10^{-8})$ mol and $\mathbf{c}_{g,0}^{\infty,T} = (4.0,\ 2.0,\ 4.0)$ mol/m^3. Left: time evolution of $\mathbf{b}^n$, $n = 0, 1, 2, \ldots$ on the phase diagram. Right: a zoomed-in view near the phase boundary between the areas 1 and 2.

The simulation succeeds and the computed deactivation time is equal to $t = 0.059677$ s. However let us make two remarks. The first remark concerns the time evolution of the mole-fraction vectors $\mathbf{x}_\alpha^n$, $\alpha = 1, 2, 3$ depicted in Figure 3.26 (c). At the deactivation point the new inactive index is $\alpha = 3$, whereas the convex area $\Delta_{2,3}$ is located at the bottom left corner of the phase diagram. The vector $\mathbf{x}_3^n$ jumps out from its convex area, which is normally forbidden by the Algorithm 3.2.1. The constraint that should deactivate is in fact $y_2 = 0$, but the point $\mathbf{x}_2^n$ remains at its initial position during the whole simulation. Hence, the simulation succeeds, nevertheless it does not have the expected behaviour. The second remark is about the behaviour of y_1^n, $n = 0, 1, 2, \ldots$ after the deactivation of y_3^n: the approximations y_1^n change completely their evolution at the discontinuity time and start to decrease.

In conclusion these examples highlight limitations of the optimization-based numerical method on a non classical phase diagram. The first difficulty comes from the algorithm that computes the values of $\mathbf{x}_{\bar{\alpha}}$, $\bar{\alpha} \in \mathcal{A}$. This algorithm should ensure the convergence of $\mathbf{x}_{\bar{\alpha}}^{\mathcal{A}}$ to the minimizer of the distance in its convex area, but $\mathbf{x}_{\bar{\alpha}}^{\mathcal{A}}$ may converge to a global minimizer if the convex area that contains this minimizer is common to $\Delta'_{s,\bar{\alpha}}$. The point $\mathbf{x}_{\bar{\alpha}}^{\mathcal{A}}$ may also jump out from $\Delta'_{s,\bar{\alpha}}$. The second limitation is due to the interior-point algorithm presented in [4, 5] that cannot keep a new constraint $\bar{\alpha}$ inactive if the associated vector $\mathbf{x}_{\bar{\alpha}}$ is to close to another vector $\mathbf{x}_\alpha$, $\alpha \in \mathcal{I}$. The constraint $\bar{\alpha}$ is activated and the simulation cannot restart correctly from the deactivation.



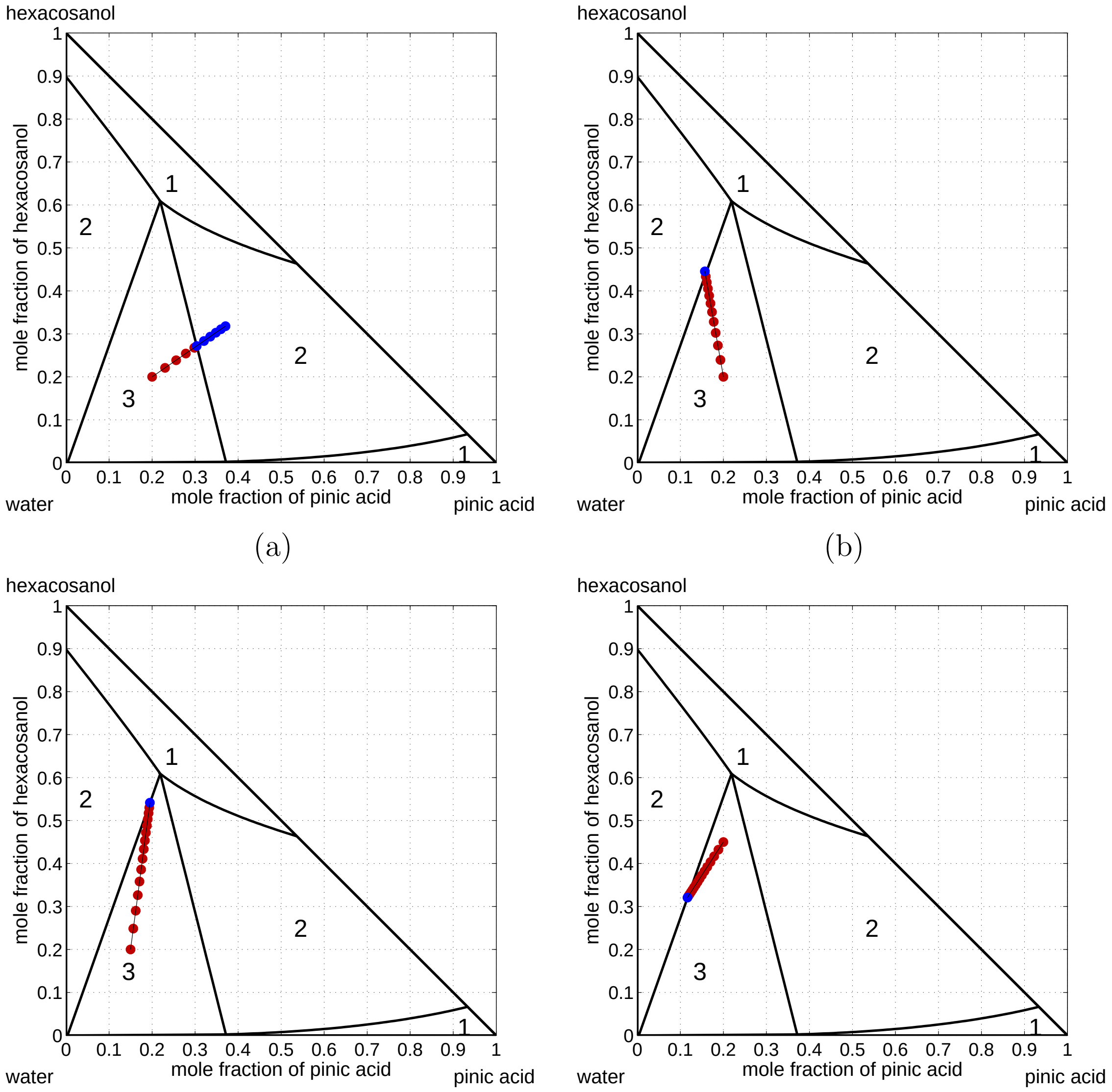
(a) (b)

(c) (d)

Figure 3.26: Example LL with the initial conditions $\mathbf{b}_0^T = (7.5{\cdot}10^{-8},\ 2.0{\cdot}10^{-8},\ 0.5{\cdot}10^{-8})$ mol and $\mathbf{c}_{g,0}^{\infty,T} = (3.0,\ 3.0,\ 4.0)$ mol/m^3. Time evolution of (a) $\mathbf{b}^n$, $n = 0, 1, 2, \ldots$; (b) the phase simplices; (c) the molar-fraction vectors $\mathbf{x}_\alpha^n$, $\alpha = 1, 2, 3$ and $n = 0, 1, 2, \ldots$; (d) the number of moles y_α^n, $\alpha = 1, 2, 3$ and $n = 0, 1, 2, \ldots$.

Theoretical example

In Section 3.4 we have seen that when all constraints are inactive the exact solution $\mathbf{b}$ and the exact time of activation t^* when an inequality constraint is activated are known and a priori error estimates can be established. Let us consider four different trajectories on the phase diagram VL to illustrate the error on the computation of discontinuity points between the approximated and exact solutions for each example. The trajectories are depicted in Figure 3.27.



(a) (b) (c) (d)

Figure 3.27: Four trajectories on the phase diagram VL starting from area 3 and evoluting to one of the areas 2 with different angles.

Figure 3.28 illustrates for each trajectory the error on the computation of the discontinuity time and point in a log-log scale. The graph on the left presents the error between the exact discontinuity t^* and the computed discontinuity time t^{n+1}. Each colored plot refers to a trajectory on the phase diagram whereas the black line stands for the function h^2. Then one can deduce that the error in time is of second order.

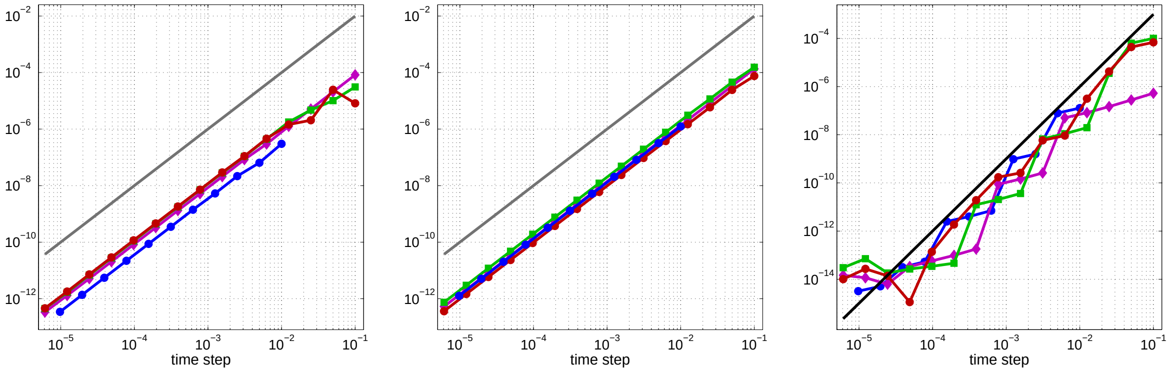


Figure 3.28: Error on the computation of the discontinuity point in the case of an activation in log-log scale. Left: error on the activation times $|t^* - t^{n+1}|$. Middle: error on the reduced composition-vector at t^{n+1} $\|\mathbf{d}(t^{n+1}) - \mathbf{d}^{n+1}\|_2$. Right: error on the activation points $\|\mathbf{d}(t^*) - \mathbf{d}^{n+1}\|_2$.

Figure 3.28 (middle) depicts the error between the computed discontinuity point and the exact trajectory $\mathbf{b}$ taken at the computed discontinuity time t^{n+1} in the Euclidean norm. As for the previous graph the black line stands for the function h^2. The conclusion is the same as before: the error is of second order.

The last graph in Figure 3.28 is concerned by the error between the computed discontinuity point and the exact discontinuity point at the exact discontinuity time in the Euclidean norm. For this graph the black line depicts the function h^3. This last error is thus of order 3 and one can conclude that the computation of the activation is well adapted to the Crank-Nicolson method since the errors are of second order.

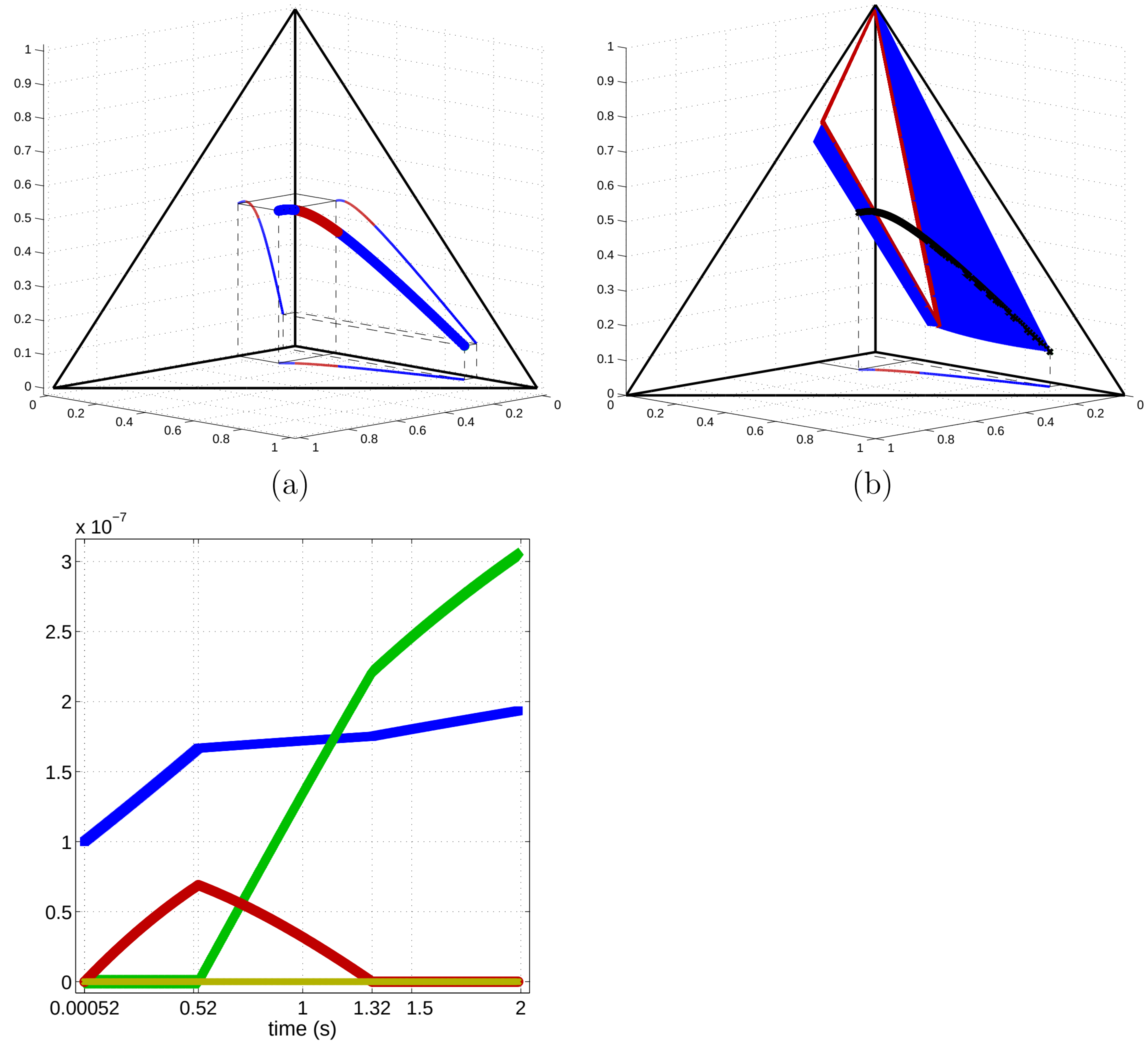
3.5.3 Numerical results in higher dimensions

For gas-aerosol systems with $s > 3$, no phase diagram is available. However let us consider first two examples with $s = 4$ that can be represented on a tetrahedron. For the numerical results of this section the colored repartition is the same as before with a yellow color for the phase simplex whose dimension is equal to 4 or the fourth inequality constraint y_4.

The gas-aerosol system that is considered is made of pinic acid, 1-hexacosanol, water and n-propanol, and the interaction parameters that define the energy function g are taken from [51]. The first example is presented in Figure 3.29 and the corresponding initial conditions are

- composition-vector: $\mathbf{b}_0^T = (0.5 \cdot 10^{-8},\ 7.5 \cdot 10^{-8},\ 1 \cdot 10^{-8},\ 1 \cdot 10^{-8})$ mol,
- gas concentration-vector: $\mathbf{c}_{g,0}^{\infty,T} = (4.0,\ 0.5,\ 4.0,\ 1.5)$ mol/m^3,
- time step: $h = 0.005$ s.

The initial-composition vector $\mathbf{b}_0$ is situated at the bottom right corner of the tetrahedron presented in Figure 3.29. The resolution of the PEP at $\mathbf{b}_0$ gives a single-phase point. A first deactivation is computed at $t^1 = 0.000523$ s. A second deactivation is detected at the time step 136 but with the backwards check the simulation has to go back to the time step 105 to compute the deactivation time. The calculated discontinuity time is then equal to $t^2 = 0.521991$ s. Finally an activation is detected and it occurs at $t^3 = 1.318174$ s.



(a)

(b)

(c)

Figure 3.29: Example whose initial conditions are $\mathbf{b}_0^T = (0.5 \cdot 10^{-8},\ 7.5 \cdot 10^{-8},\ 1 \cdot 10^{-8},\ 1 \cdot 10^{-8})$ mol and $\mathbf{c}_{g,0}^{\infty,T} = (4.0,\ 0.5,\ 4.0,\ 1.5)$ mol/m^3. Time evolution of (a) $\mathbf{b}^n$, $n = 0, 1, 2, \ldots$; (b) the phase simplices and (c) y_α^n, $\alpha = 1, 2, 3, 4$ and $n = 0, 1, 2, \ldots$.

Illustration of the results are given in Figure 3.29 where the time evolution of $\mathbf{b}^n$, $n = 0, 1, 2, \ldots$, the phase simplices and y_α^n, $\alpha = 1, 2, 3, 4$ and $n = 0, 1, 2, \ldots$, are represented on each graph respectively. For a better understanding of the trajectory of $\mathbf{b}^n$ on the tetrahedron, the projection of the trajectory is given on each face of the tetrahedron. Concerning the evolution of the phase simplices in Figure 3.29 (b), the vectors $\mathbf{b}^n$, $n = 0, 1, 2, \ldots$, are illustrated by black circles when they are not single-phase points. Hence one can observe the continuous time evolution of the phase simplices on the tetrahedron which allows us to accept the numerical results. Finally the graph of Figure 3.29 (c) shows the loss of regularity of y^n_α, $\alpha \in \mathcal{I}$, at each discontinuity time.

The second example is presented in Figure 3.30 and the corresponding initial conditions are

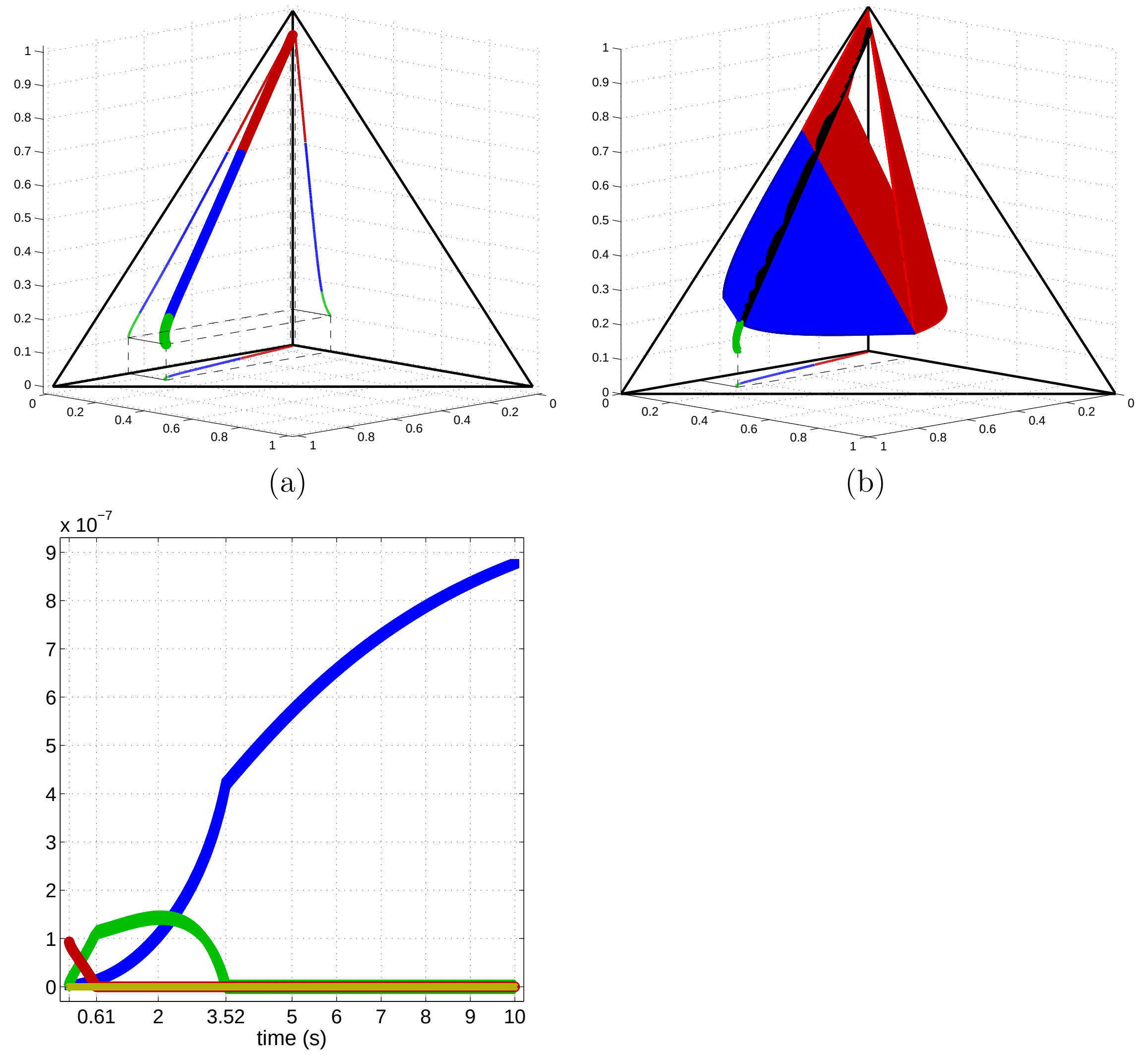
- composition-vector: $\mathbf{b}_0^T = (0.1 \cdot 10^{-8},\ 0.1 \cdot 10^{-8},\ 9.3 \cdot 10^{-8},\ 0.5 \cdot 10^{-8})$ mol,
- gas concentration-vector: $\mathbf{c}_{g,0}^{\infty,T} = (7.0,\ 2.0,\ 0.5,\ 0.5)$ mol/m^3,
- time step: $h = 0.005$ s.

This second example has an initial vector $\mathbf{b}_0$ situated near the top of the tetrahedron and the initial phase equilibrium is given by a phase simplex of dimension 3. During the simulation two activations occur: at $t^1 = 0.613664$ s and then at $t^2 = 3.517124$ s. The numerical results of this example are represented in Figure 3.30. As for the previous example, the evolution of the phase simplices is continuous and allows to conclude the accuracy of the simulation.

For gas-aerosol system with $s > 4$ the representation of the results on the phase diagram is not easy. However one can observe the CPU times in order to evaluate the efficiency of the optimization-based numerical method. In Table 3.3 the CPU times of 2 examples with $s = 18$ and $h = 0.01$ s are proposed. In the first example the solution is a single-phase point at $t = 0$ s and then a deactivation occurs at $t = 0.0743$ s. The initial solution of the second example has four inactive inequality constraints. A first inequality constraint is activated at $t = 0.0067$ s. Then a second activation occurs at $t = 0.7847$ s.

ex. 1	code	fixed point	mindist	act.	deact.	backwards	total disc.
time [s]	1218.43	732.71	18.26		0.38	0.73	1.11
%		60.1	1.5		0	0.1	0.1
ex. 2	code	fixed point	mindist	act.	deact.	backwards	total disc.
time [s]	1474.36	901.44	14.13	0.89			0.89
%		61.1	1.	0.1			0.1

Table 3.3: Computational cost of the numerical resolution for 2 examples with $s = 18$. Legend is as follows: code: total time; fixed point: time for the fixed-point algorithm; mindist: time for the computation of $\mathbf{x}_\alpha$, $\alpha \in \mathcal{A}$; act.: time for the computation of the activation; deact.: time for the computation of the deactivation; backwards: time spent in going backwards in the trajectory; total disc.: total time for the computation of the discontinuities (i.e act.+deact.+backwards).



(a)

(b)

(c)

Figure 3.30: Example whose initial conditions are $\mathbf{b}_0^T = (0.1\cdot 10^{-8},\ 0.1\cdot 10^{-8},\ 9.3\cdot 10^{-8},\ 0.5\cdot 10^{-8})$ mol and $\mathbf{c}_{g,0}^{\infty,T} = (7.0,\ 2.0,\ 0.5,\ 0.5)$ mol/m^3. Time evolution of (a) $\mathbf{b}^n$, $n = 0, 1, 2, \ldots$; (b) the phase simplices and (c) y_α^n, $\alpha = 1, 2, 3, 4$ and $n = 0, 1, 2, \ldots$.

The CPU times presented in Table 3.3 illustrate respectively the total time of execution, the time for the fixed-point algorithm, the time for the computation of $\mathbf{x}_\alpha$, $\alpha \in \mathcal{A}$, the time for the computation of the activation time and points, the time for the computation of the deactivation time and points, the time spent in going backwards in the trajectory and the total time for the computation of the discontinuity time and points (i.e the addition of the last three times). This table shows that 60% of the CPU time is spent in the fixed-point algorithm and the others CPU times are very short in comparison to the time dedicated to the fixed-point method. The time spent in the fixed-point algorithm is mainly due ($\sim$ 60%)

to the computation of $\theta_\alpha = g(\mathbf{x}_\alpha) + \boldsymbol{\lambda}^T\mathbf{x}_\alpha$, $\alpha \in \mathcal{A}$ in the active set identification procedure. This procedure allows to update the active set in the primal-dual interior-point method and is also a criterion for the detection of a deactivation since $\theta_\alpha^{\mathcal{A}}$ represents the distance between $(\mathbf{x}_\alpha^{\mathcal{A}}, g(\mathbf{x}_\alpha^{\mathcal{A}}))$ and the supporting tangent plane. Let us remind that the vector $\mathbf{x}_\alpha^{\mathcal{A}}$ is fixed in the compuation of $\theta_\alpha^{\mathcal{A}}$. The value of $\mathbf{x}_\alpha^{\mathcal{A}}$ is updated at each time step of the numerical method via the Algorithm 3.2.1. Following the computation of $\mathbf{x}_\alpha^{\mathcal{A}}$, the distance between $(\mathbf{x}_\alpha^{\mathcal{A}}, g(\mathbf{x}_\alpha^{\mathcal{A}}))$ and the supporting tangent plane is checked. Thus the criterion in the active set identification procedure is similar to this last criterion and may be omitted.

The examples 1 and 2 are solved without the computation of θ_α, $\alpha \in \mathcal{A}$ in the active set identification procedure. The solution for each example is similar to the previous simulation. The CPU times for the new simulation are summarized in Table 3.4. The CPU times in the fixed-point algorithm are considerably decreased and the total CPU times for the simulation is less than 5 minutes, which is relatively short for examples with $s = 18$. Furthermore the total disc. CPU time in Table 3.4 proves the efficiency of the technique that computes the discontinuity time and points.

ex. 1	code	fixed point	mindist	act.	deact.	backwards	total disc.
time [s] %	232.15	100.34 43.2	18.48 8.		0.06 0.	0.1 0.	0.16 0.
ex. 2	code	fixed point	mindist	act.	deact.	backwards	total disc.
time [s] %	273.95	125.92 46.	14.11 5.2	0.12 0.			0.12 0.

Table 3.4: Computational cost of the numerical resolution for 2 examples with $s = 18$. Legend is as follows: code: total time; fixed point: time for the fixed-point algorithm; mindist: time for the computation of $\mathbf{x}_\alpha$, $\alpha \in \mathcal{A}$; act.: time for the computation of the activation; deact.: time for the computation of the deactivation; backwards: time spent in going backwards in the trajectory; total disc.: total time for the computation of the discontinuities (i.e act.+deact.+backwards).

In Table 3.5 the CPU times for the previous examples with $s = 3$ and $s = 4$ are presented. The CPU times for the whole simulation is still short with less than 2 s for examples with $s = 2$ and than 1 minute for examples with $s = 4$. Furthermore the time dedicated to the computation of the activation/deactivation time and points is a short percentage of the whole simulation. Hence the goal to build a fast technique to model the gas-aerosol system is reached. Furthermore the numerical resolution proved itself to be efficient on classical phase diagrams. When the phase diagram contains common convex areas, the method encounters some difficulties to deactivate inactive inequalities and restart the simulation due to the numerical method for the optimization problems, namely the interior-point method, and the initialization of the variables $\mathbf{x}_\alpha$, $\alpha \in \mathcal{A}$, in the computation of the minimal distance criterion.

ex. Figure 3.18	code	fixed point	mindist	act.	deact.	backwards	total disc.
time [s]	0.50	0.22	0.01	0.	0.	0.0	0.01
%		44.5	2	0.1	0.3	0.1	0.5
ex. Figure 3.21	code	fixed point	mindist	act.	deact.	backwards	total disc.
time [s]	1.24	0.66	0.03	0.	0.	0.01	0.01
%		52.9	2.4	0.1	0.2	0.5	0.8
ex. Figure 3.29	code	fixed point	mindist	act.	deact.	backwards	total disc.
time [s]	21.59	9.75	0.35	0.01	0.01	1.98	2.
%		45.2	1.6	0.	0.1	9.2	9.3
ex. Figure 3.30	code	fixed point	mindist	act.	deact.	backwards	total disc.
time [s]	50.27	27.46	1.40	0.01			0.01
%		54.8	2.8	0.			0.

Table 3.5: Computational cost of the numerical resolution for examples with $s = 3, 4$. Legend is as follows: code: total time; fixed point: time for the fixed-point algorithm; mindist: time for the computation of $\mathbf{x}_\alpha$, $\alpha \in \mathcal{A}$; act.: time for the computation of the activation; deact.: time for the computation of the deactivation; backwards: time spent in going backwards in the trajectory; total disc.: total time for the computation of the discontinuities (i.e act.+deact.+backwards).

Chapter 4

A numerical method based on differential algebraic systems

Let us recall the set of equations that models the gas-aerosol system: find $\mathbf{b}, \mathbf{x}_\alpha : (0,T) \to \mathbb{R}^s_+$ and $R, y_\alpha : (0,T) \to \mathbb{R}_+$, $\alpha = 1, \ldots, p$ satisfying

$$\begin{aligned}
\frac{d}{dt}\mathbf{b}(t) &= \mathbf{j}\left(\mathbf{b}(t), \mathbf{x}_\alpha^{\mathcal{I}}(t), R(t)\right), \qquad \mathbf{b}(0) = \mathbf{b}_0 \\
R(t) &= \left(\frac{3}{4\pi}\sum_{i=1}^{s}\frac{m_{c,i}b_i(t)}{\rho_i}\right)^{\frac{1}{3}}, \\
\{\mathbf{x}_\alpha(t), y_\alpha(t)\}_{\alpha=1}^p &= \underset{\{\bar{\mathbf{x}}_\alpha, \bar{y}_\alpha\}_{\alpha=1}^p}{\operatorname{argmin}} \sum_{\alpha=1}^{p} \bar{y}_\alpha\, g(\bar{\mathbf{x}}_\alpha) \qquad (4.0.1)\\
&\qquad \text{s.t.} \sum_{\alpha=1}^{p}\bar{y}_\alpha\bar{\mathbf{x}}_\alpha = \mathbf{b}(t), \\
&\qquad \mathbf{e}^T\bar{\mathbf{x}}_\alpha = 1,\ \bar{\mathbf{x}}_\alpha > 0,\ \bar{y}_\alpha \geq 0,\ \alpha = 1, \ldots, p,
\end{aligned}$$

where T is the final time of integration, $\mathbf{b}_0$ is a given initial composition-vector and the flux $\mathbf{j}$ is defined by

$$\mathbf{j}(\mathbf{b}(t), \mathbf{x}_\alpha^{\mathcal{I}}(t), R(t)) = \mathbf{H}(R(t))\left(\mathbf{b}^{\text{tot}} - N\mathbf{b}(t) - \frac{1}{\mathcal{R}_c T}\exp\left(\boldsymbol{\nabla} g(\mathbf{x}_\alpha^{\mathcal{I}}(t)) + \ln(\mathbf{p}_g^o)\right)\right).$$

Let us remind that the exponent $\mathcal{I}$ is added to specify that $\alpha \in \mathcal{I}$ is such that $y_\alpha > 0$ (inactive constraint).

In Chapter 3 the system (4.0.1) has been solved with a fixed-point approach which coupled the Crank-Nicolson scheme for the ordinary differential part and a primal-dual interior-point method for the minimization problem. This first method consists in a dynamic motion of the phase equilibrium problem and follows the chemical meanings of the variables $\mathbf{b}$, R, $\mathbf{x}_\alpha$ and y_α, $\alpha = 1, \ldots, p$, by enforcing these latter to remain positive. A new approach is developed here, based on the mathematical observation: if the number of inactive inequality constraints is fixed, the set (4.0.1) can be treated as a differential algebraic system of equations (DAE).

The presentation of this new approach follows the same structure as in Chapter 3, namely

(i) the resolution of the DAE stemmed from (4.0.1) when the number of inactive inequality constraints is fixed;

(ii) the development of criteria for the detection of the discontinuities generated by the activation or deactivation of an inequality constraint;

(iii) the computation of the discontinuity time and points;

(iv) the definition of the new DAE in accordance with the number of inactive constraints.

Hence let us begin with the formulation and the resolution of the differential algebraic system under the assumption that the number of inactive constraints is fixed.

4.1 Numerical method for a fixed number of inactive constraints

This section is devoted to the resolution of (4.0.1) with a fixed number of inactive inequality constraints. In this case, the regularity of the variables $\mathbf{x}_\alpha$ and y_α, $\alpha = 1, \ldots, p$ is guaranteed.

4.1.1 The differential algebraic equations

The minimization problem in (4.0.1) consists in the computation of the convex envelope (see Section 1.7). Therefore if a constraint $\bar{\alpha}$ is active (*i.e.* if $y_{\bar{\alpha}}(t) = 0$), then the variables $y_{\bar{\alpha}}$ and $\mathbf{x}_{\bar{\alpha}}$ can be removed from the optimization algorithm without affecting the solution. When considering only the inactive constraints, (4.0.1) includes an optimization problem with equality constraints only:

$$\begin{aligned}
\frac{d}{dt}\mathbf{b}(t) &= \mathbf{j}\left(\mathbf{b}(t), \mathbf{x}_\alpha^{\mathcal{I}}(t), R(t)\right), \qquad \mathbf{b}(0) = \mathbf{b}_0 \\
R(t) &= \left(\frac{3}{4\pi}\sum_{i=1}^{s}\frac{m_{c,i}b_i(t)}{\rho_i}\right)^{\frac{1}{3}}, \\
\{y_\alpha(t), \mathbf{x}_\alpha(t)\}_{\alpha\in\mathcal{I}(t)} &= \underset{\{\bar{y}_\alpha, \bar{\mathbf{x}}_\alpha\}_{\alpha\in\mathcal{I}(t)}}{\arg\min} \sum_{\alpha\in\mathcal{I}(t)} \bar{y}_\alpha\, g(\bar{\mathbf{x}}_\alpha) \\
&\qquad\qquad \text{s.t.} \sum_{\alpha\in\mathcal{I}(t)} \bar{y}_\alpha \bar{\mathbf{x}}_\alpha = \mathbf{b}(t), \\
&\qquad\qquad\qquad \mathbf{e}^T\bar{\mathbf{x}}_\alpha = 1,\ \bar{\mathbf{x}}_\alpha > 0,\ \bar{y}_\alpha \geq 0,\ \alpha \in \mathcal{I}(t).
\end{aligned} \tag{4.1.1}$$

The solution of (4.0.1) is then equivalent to the solution of (4.1.1), together with $y_\alpha(t) = 0,\ \forall\alpha \in \mathcal{A}(t)$. This implies that the variables $\mathbf{x}_\alpha^{\mathcal{A}}$ do not appear in (4.1.1) and therefore are not updated in the computation of the convex envelope (since the supporting tangent plane is not tangent to the energy function at those points). The sole condition on $\mathbf{x}_\alpha^{\mathcal{A}}$ is the normalization constraint $\mathbf{e}^T\mathbf{x}_\alpha^{\mathcal{A}} = 1$.

By replacing the minimization problem by the KKT conditions, (4.1.1) becomes

$$\begin{aligned}
\frac{d}{dt}\mathbf{b}(t) &= \mathbf{j}\left(\mathbf{b}(t), \mathbf{x}_\alpha^{\mathcal{I}}(t), R(t)\right) && \\
\mathbf{0} &= y_\alpha(t)\left(\boldsymbol{\nabla} g(\mathbf{x}_\alpha(t)) + \boldsymbol{\lambda}(t)\right) + \zeta_\alpha(t)\mathbf{e}, && \alpha \in \mathcal{I}(t), \\
0 &= g(\mathbf{x}_\alpha(t)) + \boldsymbol{\lambda}^T(t)\,\mathbf{x}_\alpha(t), && \alpha \in \mathcal{I}(t), \\
\mathbf{0} &= \sum_{\alpha\in\mathcal{I}(t)} y_\alpha(t)\,\mathbf{x}_\alpha(t) - \mathbf{b}(t), && \\
0 &= \mathbf{e}^T\mathbf{x}_\alpha(t) - 1, && \alpha \in \mathcal{I}(t), \\
0 &= R(t) - \left(\frac{3}{4\pi}\sum_{i=1}^{s}\frac{m_{c,i}b_i(t)}{\rho_i}\right)^{\frac{1}{3}}, &&
\end{aligned} \tag{4.1.2}$$

where $\boldsymbol{\lambda} \in \mathbb{R}^s$ and $\zeta_\alpha \in \mathbb{R}$, $\alpha \in \mathcal{I}(t)$, are the Lagrange multipliers associated to the equality constraints.

Multiplying the second equation of (4.1.2) by $\mathbf{x}_\alpha^T(t)$ and using the homogeneous property of g ($\mathbf{x}^T\boldsymbol{\nabla} g(\mathbf{x}) = g(\mathbf{x})$), the second equation becomes

$$0 = y_\alpha(t)\left(g(\mathbf{x}_\alpha(t)) + \mathbf{x}_\alpha^T(t)\boldsymbol{\lambda}(t)\right) + \zeta_\alpha(t)\,\mathbf{x}_\alpha^T(t)\mathbf{e}, \quad \forall\alpha \in \mathcal{I}(t).$$

Then the third and fifth equations of (4.1.2) imply

$$0 = y_\alpha(t)\cdot 0 + \zeta_\alpha(t), \quad \forall\alpha \in \mathcal{I}(t).$$

In conclusion the variable ζ_α equals to 0 when $\alpha \in \mathcal{I}(t)$ and can be removed from (4.1.2). Finally the second equation is divided by $y_\alpha(t)$ since $y_\alpha(t) > 0,\ \forall\alpha \in \mathcal{I}(t)$. The system (4.1.2) is then reformulated as

$$\begin{aligned}
\frac{d}{dt}\mathbf{b}(t) &= \mathbf{j}\left(\mathbf{b}(t), \mathbf{x}_\alpha^{\mathcal{I}}(t), R(t)\right) && \\
\mathbf{0} &= \boldsymbol{\nabla} g(\mathbf{x}_\alpha(t)) + \boldsymbol{\lambda}(t), && \alpha \in \mathcal{I}(t), \\
\mathbf{0} &= \sum_{\alpha\in\mathcal{I}(t)} y_\alpha(t)\,\mathbf{x}_\alpha(t) - \mathbf{b}(t), && \\
0 &= \mathbf{e}^T\mathbf{x}_\alpha(t) - 1, && \alpha \in \mathcal{I}(t), \\
0 &= R(t) - \left(\frac{3}{4\pi}\sum_{i=1}^{s}\frac{m_{c,i}b_i(t)}{\rho_i}\right)^{\frac{1}{3}}, &&
\end{aligned} \tag{4.1.3}$$

the third equation of (4.1.2) being redundant. The second equation means that the gradient of g at the points $\mathbf{x}_\alpha$, $\alpha \in \mathcal{I}$, are all equal to $-\boldsymbol{\lambda}$ and consequently $\boldsymbol{\nabla} g(\mathbf{x}_\alpha) = \boldsymbol{\nabla} g(\mathbf{x}_\beta)$, $\forall \alpha, \beta \in \mathcal{I}$.

Let $\mathbf{Y}^T(t) = \left(\mathbf{b}^T(t), \mathbf{x}_1^{\mathcal{I},T}(t), \ldots, \mathbf{x}_{p^\mathcal{I}}^{\mathcal{I},T}(t), y_1^\mathcal{I}(t), \ldots, y_{p^\mathcal{I}}^\mathcal{I}(t), \boldsymbol{\lambda}^T(t), R(t)\right)$ be a $\mathcal{N}$-vector, $\mathcal{N} = s + sp^\mathcal{I} + p^\mathcal{I} + s + 1$, that contains all the unknowns of (4.1.3). The system (4.1.3) can be written as

$$M \frac{d\mathbf{Y}}{dt}(t) = \mathbf{F}(\mathbf{Y}(t)), \tag{4.1.4}$$

where the function $\mathbf{F}$ is the right hand side of (4.1.3), namely

$$\mathbf{F}(\mathbf{Y}(t)) = \left(\begin{array}{c} \mathbf{j}\left(\mathbf{b}(t), \mathbf{x}_\alpha^\mathcal{I}(t), R(t)\right) \\ \boldsymbol{\nabla} g(\mathbf{x}_1^\mathcal{I}(t)) + \boldsymbol{\lambda}(t) \\ \vdots \\ \boldsymbol{\nabla} g(\mathbf{x}_{p^\mathcal{I}}^\mathcal{I}(t)) + \boldsymbol{\lambda}(t) \\ \displaystyle\sum_{\alpha \in \mathcal{I}(t)} y_\alpha(t)\, \mathbf{x}_\alpha(t) - \mathbf{b}(t), \\ \mathbf{e}^T \mathbf{x}_1^\mathcal{I}(t) - 1 \\ \vdots \\ \mathbf{e}^T \mathbf{x}_{p^\mathcal{I}}^\mathcal{I}(t) - 1 \\ R(t) - \left(\dfrac{3}{4\pi} \displaystyle\sum_{i=1}^{s} \dfrac{m_{c,i} b_i(t)}{\rho_i} \right)^{\frac{1}{3}} \end{array} \right).$$

and the matrix $M = \begin{pmatrix} \mathbf{I}_s & \mathbf{0} \\ \mathbf{0} & 0 \end{pmatrix}$ with $\mathbf{I}_s$, the $s \times s$ identity matrix.

The system (4.1.4) is completed by the initial condition $\mathbf{Y}(0) = \mathbf{Y}_0$. The first s components of $\mathbf{Y}_0$ (related to the variable $\mathbf{b}$) are given by the initial condition $\mathbf{b}_0$ in (4.0.1). The initial value of the (algebraic) variables $\mathbf{x}_\alpha^\mathcal{I}$, $y_\alpha^\mathcal{I}$, $\boldsymbol{\lambda}$ and R must satisfy the *consistency conditions* $\mathbf{F}_a(\mathbf{Y}_0) = \mathbf{0}$, where $\mathbf{F}_a$ is the subvector of $\mathbf{F}$ defined by the $\mathcal{N} - s$ last components of $\mathbf{F}$. In particular, the value of R at $t = 0$ is immediately given by the last equation of the DAE system. For the other variables, the solution of the condition $\mathbf{F}_a(\mathbf{Y}_0) = \mathbf{0}$ corresponds to the solution of the minimization problem in (4.0.1) for a given concentration-vector $\mathbf{b}_0$. The primal-dual interior-point method proposed in Chapter 2 allows to determine $\boldsymbol{\lambda}$, $\mathbf{x}_\alpha$ and y_α, $\alpha \in \mathcal{I}$ for given $\mathbf{b}_0$. Hence there exist consistent initial values that solve $\mathbf{F}_a(\mathbf{Y}_0) = \mathbf{0}$, for given $\mathbf{b}_0$.

The system (4.1.4) is a system of differential algebraic equations of index one, that couples the differential variable $\mathbf{b}$ and the algebraic variables $(\mathbf{x}_\alpha^\mathcal{I}, y_\alpha^\mathcal{I}, \boldsymbol{\lambda}, R)$. Such systems are widely studied in the literature (see *e.g.* [16, 47, 49, 50]). A 3-stage implicit Runge-Kutta method RADAU5 of order 5 [49, 50] is used here for the resolution of (4.1.4). A short presentation of the Runge-Kutta methods, and especially the RADAU5 method, is addressed in next section.

4.1.2 Runge-Kutta methods and the RADAU5 method

For this section let us consider first the system of ordinary differential equations

$$\mathbf{y}^{'}(x) \;=\; \mathbf{f}(x, \mathbf{y}(x)), \tag{4.1.5}$$

and apply then the theory of the Runge-Kutta methods to the differential algebraic equations of (4.1.4). Note that the notations used for this first part are local.

A q-stage implicit Runge-Kutta method to approximate $\mathbf{y}^{n+1}$ is the one step method defined by

$$\mathbf{Z}^i \;=\; \mathbf{y}^n + h\sum_{j=1}^{q} a_{ij}\,\mathbf{f}(x^n + c_j h, \mathbf{Z}^j), \quad i = 1, \ldots, q, \tag{4.1.6}$$

$$\mathbf{y}^{n+1} \;=\; \mathbf{y}^n + h\sum_{j=1}^{q} b_j\,\mathbf{f}(x^n + c_j h, \mathbf{Z}^j), \tag{4.1.7}$$

where $\{a_{ij}\}_{i,j=1}^q$, $\{b_j\}_{j=1}^q$ and $\{c_j\}_{j=1}^q$ are given coefficients, $h = x^{n+1} - x^n$ is the step size and $\mathbf{y}^n$ is an approximation of the solution $\mathbf{y}(x^n)$ at time x^n. Relation (4.1.6) forms a nonlinear system of equations for the internal stages values $\mathbf{Z}^i$, $i = 1, \ldots, q$.

According to the values of the coefficients $\{a_{ij}\}_{i,j=1}^q$, $\{b_j\}_{j=1}^q$ and $\{c_j\}_{j=1}^q$, different Runge-Kutta methods are defined. Butcher in [15] has introduced the representation of these coefficients in a table

$$\begin{array}{c|ccc} c_1 & a_{11} & \ldots & a_{1q} \\ \vdots & \vdots & \ddots & \vdots \\ c_q & a_{q1} & \ldots & a_{qq} \\ \hline & b_1 & \ldots & b_q \end{array}$$

If the coefficients of a q-stage implicit Runge-Kutta method satisfy

$$a_{qj} = b_j,\ j = 1, \ldots, q,$$

the method is called *stiffly accurate* [87]. In ordinary differential equation theory, in order to determine if a method is stable, one needs to study the Dahlquist test equation

$$\mathbf{y}^{'} = \lambda\mathbf{y},\ \mathbf{y}^n = 1.$$

Setting $z = h\lambda$ and using the Runge-Kutta scheme (4.1.6)-(4.1.7) for the test equation, one can write one step of the method as

$$\mathbf{y}^{n+1} \;=\; R(h\lambda)\mathbf{y}^n,$$

where the function R is called the *stability function.* Moreover let us define the set

$$S \;=\; \{z \in \mathbb{C} \text{ s.t. } |R(z)| \leq 1\},$$

called the *stability domain.*

Definition 4.1.1. *A method whose stability domain satisfies*

$$S \supset \mathbb{C}^{-} = \{z \,|\, \mathcal{R}e\, z \leq 0\}$$

is called A-stable.

One can also define the L-stability

Definition 4.1.2. *A method is called* L-stable *if it is A-stable and if in addition*

$$\lim_{z \to \infty} R(z) = 0.$$

Butcher gives conditions on the coefficients $\{a_{ij}\}_{i,j=1}^{q}$, $\{b_j\}_{j=1}^{q}$ and $\{c_j\}_{j=1}^{q}$ for the method to be of order q [15, 16, 49].

Theorem 4.1.1. *If the coefficients* $\{a_{ij}\}_{i,j=1}^{q}$*,* $\{b_j\}_{j=1}^{q}$ *and* $\{c_j\}_{j=1}^{q}$ *of a Runge-Kutta method satisfy*

$$\begin{array}{lll} B(q): & \sum_{i=1}^{q} b_i c_i^{k-1} = \frac{1}{k}, & k = 1, \ldots, q; \\ C(\eta): & \sum_{j=1}^{q} a_{ij} c_j^{k-1} = \frac{c_i^k}{k}, & i = 1, \ldots, q,\; k = 1, \ldots, \eta; \\ D(\zeta): & \sum_{i=1}^{q} b_i c_i^{k-1} a_{ij} = \frac{b_j}{k}(1 - c_j^k), & j = 1, \ldots, q,\; k = 1, \ldots, \zeta; \end{array}$$

with $q \leq \eta + \zeta + 1$ *and* $q \leq 2\eta + 2$*, then the method is of order* q*.*

Ehle introduced in [31] methods based on the Radau quadrature formulas to determine $\{a_{ij}\}_{i,j=1}^{q}$, $\{b_j\}_{j=1}^{q}$ and $\{c_j\}_{j=1}^{q}$. The Radau methods are quadrature formulas of maximal order with one endpoint as a prescribed node [49, 50]. The formulas for the interval $[0, 1]$ with q stages and a fixed right endpoint have nodes c_j, $j = 1, \ldots, q$, which are zeros of

$$\frac{\mathrm{d}^{q-1}}{\mathrm{d}x^{q-1}} \left(x^{q-1}(x-1)^q \right).$$

Hence the coefficients $\{c_j\}_{j=1}^{q}$ are determined. The coefficients $\{b_j\}_{j=1}^{q}$ are chosen such that the quadrature formula satisfies $B(q)$.

Butcher named this method as process of type II. Furthermore with the help of the conditions $C(\eta)$ and $D(\zeta)$ he was able to give values to the coefficients $\{a_{ij}\}_{i,j=1}^{q}$ and to construct Runge-Kutta methods of order $2q$. Unfortunately none of his methods turned out to be A-stable. Ehle [31] (and independently Axelsson [8]) took up the ideas of Butcher and found A-stable methods by imposing condition $C(q)$. These methods are called RadauIIA and the following theorem holds

Theorem 4.1.2. *The* q*-stage RadauIIA method is A-stable and of order* $2q - 1$*.*

The proof of this theorem can be found in [49]. The coefficients for the 3-stage RadauIIA method are

$$\begin{array}{c|ccc} \frac{4-\sqrt{6}}{10} & \frac{88-7\sqrt{6}}{360} & \frac{296-169\sqrt{6}}{1800} & \frac{-2+3\sqrt{6}}{225} \\ \frac{4+\sqrt{6}}{10} & \frac{296+169\sqrt{6}}{1800} & \frac{88+7\sqrt{6}}{360} & \frac{-2-3\sqrt{6}}{225} \\ 1 & \frac{16-\sqrt{6}}{36} & \frac{16+\sqrt{6}}{36} & \frac{1}{9} \\ \hline & \frac{16-\sqrt{6}}{36} & \frac{16+\sqrt{6}}{36} & \frac{1}{9} \end{array} \tag{4.1.8}$$

The stability function for the 3-stage RadauIIA method is given by

$$R(z) = \frac{1 + \frac{2}{5}z + \frac{1}{20}z^2}{1 - \frac{3}{5}z + \frac{3}{20}z^2 - \frac{1}{60}z^3}.$$

Then $\lim_{z\to\infty} R(z) = 0$ and one deduces the L-stability of the method. Finally the 3-stage RadauIIA method is also stiffly accurate since

$$a_{3j} = b_j, \; j = 1, \ldots, 3.$$

With all these properties Hairer and Wanner show in [49] that the RadauIIA is an efficient implicit Runge-Kutta method to solve the ordinary differential system (4.1.5). They have implemented this method of order 5 with in addition a step size control. Their code is called *RADAU5* [49, 50]. The resolution method of RADAU5 consists first in substituting the internal variables $\mathbf{Z}^i$ by

$$\mathbf{z}^i \;=\; \mathbf{Z}^i - \mathbf{y}^n,$$

in order to reduce the influence of round-off errors. Then the relation (4.1.6) becomes with $q = 3$

$$\mathbf{z}^i \;=\; h \sum_{j=1}^{3} a_{ij} f(x^n + c_j h, \mathbf{y}^n + \mathbf{z}^j), \quad i = 1, \ldots, 3. \tag{4.1.9}$$

The relation (4.1.7) for the stiffly accurate method RADAU5 reads

$$\mathbf{y}^{n+1} \;=\; \mathbf{z}^3 + \mathbf{y}^n.$$

For the resolution of (4.1.9) Hairer and Wanner suggest the simplified Newton method. Details of the method can be found in [49], Section IV.8, and is not presented here.

Finally since the Runge-Kutta method is a collocation method, it provides a cheap numerical approximation to $\mathbf{y}(t^n + \theta h)$ for the whole integration interval $0 \le \theta \le 1$. The *dense output approximation* (collocation polynomial) computed at the n^{th} step t^n is denoted by $\mathbf{U}^n(t^n + \theta h^n)$. The collocation method based on Radau points is of order $2q - 1$, and the dense output of order q. The error between $\mathbf{U}^n(t^n + \theta h^n)$ and $\mathbf{y}(t^n + \theta h^n)$ is therefore composed of the global error at t^n plus the local error contribution which is bounded by $\mathcal{O}((h^n)^{q+1})$.

4.1.3 Numerical method for differential algebraic equations

In the previous subsection, Runge-Kutta methods were presented for the resolution of a system of ordinary differential equations. Here is developed how these methods can also be applied for the resolution of differential algebraic equations such as in (4.1.4). The idea is to apply the Runge-Kutta method to differential equations of singular perturbation type and then to consider in the resulting formulas the limit $\varepsilon \to 0$, if ε expresses the perturbation [50, 82]. Hence let us begin this section by considering the following singular perturbation problem

$$\begin{aligned} \mathbf{y}^{'} &= \mathbf{f}(\mathbf{y},\mathbf{w}), && (4.1.10)\\ \varepsilon\,\mathbf{w}^{'} &= \mathbf{g}(\mathbf{y},\mathbf{w}). && (4.1.11) \end{aligned}$$

The corresponding reduced problem is the differential algebraic equations

$$\begin{aligned} \mathbf{y}^{'} &= \mathbf{f}(\mathbf{y},\mathbf{w}), && (4.1.12)\\ \mathbf{0} &= \mathbf{g}(\mathbf{y},\mathbf{w}). && (4.1.13) \end{aligned}$$

The q-stage implicit Runge-Kutta methods applied to the system (4.1.10)-(4.1.11) yields

$$\begin{aligned} \mathbf{Z}^i &= \mathbf{y}^n + h\sum_{j=1}^{q} a_{ij}\,\mathbf{f}(\mathbf{Z}^j,\mathbf{W}^j), \quad i=1,\dots,q, && (4.1.14)\\ \varepsilon\mathbf{W}^i &= \varepsilon\mathbf{w}^n + h\sum_{j=1}^{q} a_{ij}\,\mathbf{g}(\mathbf{Z}^j,\mathbf{W}^j),\ i=1,\dots,q, && (4.1.15)\\ \mathbf{y}^{n+1} &= \mathbf{y}^n + h\sum_{i=1}^{q} b_i\,\mathbf{f}(\mathbf{Z}^i,\mathbf{W}^i), && (4.1.16)\\ \varepsilon\mathbf{w}^{n+1} &= \varepsilon\mathbf{w}^n + h\sum_{i=1}^{q} b_i\,\mathbf{g}(\mathbf{Z}^i,\mathbf{W}^i). && (4.1.17) \end{aligned}$$

Now let us suppose that the matrix (a_{ij}) is invertible (which is the case for the RadauIIA method). From (4.1.15) one gets for $i=1,\dots,q$

$$h\,\mathbf{g}(\mathbf{Z}^i,\mathbf{W}^i) = \varepsilon\sum_{j=1}^{q}\omega_{ij}(\mathbf{W}^j-\mathbf{w}^n).$$

where ω_{ij} are the elements of the inverse of (a_{ij}). Inserting this relation into (4.1.17) gives

$$\varepsilon\mathbf{w}^{n+1} = \varepsilon\mathbf{w}^n + \varepsilon\sum_{i,j=1}^{q} b_i\,w_{ij}\,(\mathbf{W}^j-\mathbf{w}^n).$$

The parameter ε can be simplified and the definition of $\mathbf{w}^{n+1}$ becomes independent of ε. Then by taking the limit as ε tends to 0, the system (4.1.14)-(4.1.17) with the new definition of $\mathbf{w}^{n+1}$ becomes

$$\begin{aligned}
\mathbf{Z}^i &= \mathbf{y}^n + h\sum_{j=1}^{q} a_{ij}\,\mathbf{f}(\mathbf{Z}^j,\mathbf{W}^j),\ i=1,\dots,q, && (4.1.18)\\
\mathbf{0} &= \mathbf{g}(\mathbf{Z}^i,\mathbf{W}^i), \qquad\qquad i=1,\dots,q, && (4.1.19)\\
\mathbf{y}^{n+1} &= \mathbf{y}^n + h\sum_{i=1}^{q} b_i\,\mathbf{f}(\mathbf{Z}^i,\mathbf{W}^i), && (4.1.20)\\
\mathbf{w}^{n+1} &= \left(1-\sum_{i,j=1}^{q} b_i\,\omega_{ij}\right)\mathbf{w}^n + \sum_{i,j=1}^{q} b_i\,\omega_{ij}\,\mathbf{W}^j. && (4.1.21)
\end{aligned}$$

Then the solution of the DAE (4.1.12)-(4.1.13) is given by the solution of the above system. Furthermore for the stiffly accurate method such as RadauIIA, the numerical solution $(\mathbf{y}^{n+1},\mathbf{w}^{n+1})$ satisfies the equation (4.1.13). The RADAU5 method of Hairer and Wanner also contains the resolution of the differential algebraic system.

Now let us go back to the system (4.1.4). Thanks to the definition of the matrix M, this system can be written as (4.1.12)-(4.1.13) with $\mathbf{y}=\mathbf{b}$, $\mathbf{w}^T=(\mathbf{x}_1^T,\dots,\mathbf{x}_{p^{\mathcal{I}}}^T,y_1,\dots,y_{p^{\mathcal{I}}},\boldsymbol{\lambda}^T,R)$, $\mathbf{f}(\mathbf{y},\mathbf{w})=\mathbf{j}(\mathbf{b},\mathbf{x}_\alpha^{\mathcal{I}},R)$ and $\mathbf{g}=\mathbf{F}_a$. Hence the RADAU5 method can be used for the numerical solution of (4.1.4).

In the sequel, the dense output formula for specific components of $\mathbf{Y}$ are used and the corresponding component is specified by its index. For instance, the dense output for the variables $y_{\bar{\alpha}}$ at t^n is denoted by $\mathbf{U}^n_{y_{\bar{\alpha}}}(t^n+\theta h^n)$ for $\theta\in[0,1]$.

As soon as the set of inactive constraints is fixed, the RADAU5 algorithm is used. The coupling of this algorithm with an efficient procedure to compute any change in the set of inactive constraints allows to track the activation/deactivation of constraints that correspond to losses of regularity of the trajectories.

4.2 Detection of discontinuity times and points

Unlike in the optimization-based numerical method, the detection of the discontinuity points, or equivalently of the activation/deactivation of a constraint, is achieved only at each time step t^n and all detection criteria are based on checking the sign of a particular quantity. In the sequel, the cases of an activation ($y_\alpha(t)>0\rightarrow y_\alpha(t)=0$) and of a deactivation ($y_\alpha(t)=0\rightarrow y_\alpha(t)>0$) of a constraint are distinguished.

4.2.1 Activation of an inequality constraint

Let us remind that the activation of an inequality constraint corresponds to the minimal time t such that the transition $y_\alpha(t)>0\rightarrow y_\alpha(t)=0$ occurs.

When the number of active constraints is fixed and (4.1.4) is solved with the RADAU5 method, all physical and chemical senses of the variables are omitted. Therefore the variables y_α may take negative values (which is a nonsense from a chemical point of view since the quantity y_α represents a number of moles). The criterion to detect the presence of the activation of an inequality constraint is therefore to check at each time step t^{n+1} if

$$\exists\, \bar{\alpha} \in \mathcal{I}(t^{n+1}) \quad \text{such that} \quad y_{\bar{\alpha}}^{n} > 0 \quad \text{and} \quad y_{\bar{\alpha}}^{n+1} < 0. \tag{4.2.1}$$

In that case, the conclusion is that there exists a time $\tau \in (t^n, t^{n+1})$ for which the inequality constraint $y_{\bar{\alpha}}(\tau) = 0$ is activated.

4.2.2 Deactivation of an inequality constraint

A deactivation occurs when there exists an index $\bar{\alpha} \in \mathcal{A}$ such that $y_{\bar{\alpha}}(t) = 0 \to y_{\bar{\alpha}}(t) > 0$. However, the variables $y_{\bar{\alpha}}$ and $\mathbf{x}_{\bar{\alpha}}$, $\bar{\alpha} \in \mathcal{A}$, do not appear in (4.1.3) or (4.1.4) (the only condition on $\mathbf{x}_{\bar{\alpha}}$ is the normalization condition $\mathbf{e}^T\mathbf{x}_{\bar{\alpha}} = 1$). The criterion to "add" such variables into (4.1.4) for the next time step is therefore independent of the resolution of the differential algebraic system at the previous time step as it was the case for the optimization-based numerical method. Thus the detection criterion is similar and consists in computing at each time step the vector $\mathbf{x}_\alpha$, $\alpha \in \mathcal{A}$, as the point in $\Delta'_{s,\alpha}$ that minimizes the distance between the supporting tangent plane and $(\mathbf{x}_\alpha, g(\mathbf{x}_\alpha))$. If we denote by $d^n(\mathbf{x})$ the signed distance between $(\mathbf{x}, g(\mathbf{x}))$ and the supporting tangent plane at time t^n, then the criterion to detect the presence of the deactivation of an inequality constraint is to check at each time step t^{n+1} if

$$\exists\, \bar{\alpha} \in \mathcal{A}(t^{n+1}) \quad \text{such that} \quad d^n(\mathbf{x}_{\bar{\alpha}}^{n}) > 0 \quad \text{and} \quad d^{n+1}(\mathbf{x}_{\bar{\alpha}}^{n+1}) < 0, \tag{4.2.2}$$

where $\mathbf{x}_{\bar{\alpha}}^{n}$, $\mathbf{x}_{\bar{\alpha}}^{n+1} \in \Delta'_{s,\bar{\alpha}}$ are the point that respectively minimize $d^n(\cdot)$ and $d^{n+1}(\cdot)$ in the convex area $\Delta'_{s,\bar{\alpha}}$.

In that case, there exists a time $\tau \in (t^n, t^{n+1})$ for which the inequality constraint $y_{\bar{\alpha}}(\tau) = 0$ is deactivated.

For a given supporting tangent plane, the Algorithm 3.2.1 presented in Subsection 3.2.2 is used to determine the signed distance, together with the point $\mathbf{x}_\alpha^{\mathcal{A}}$ that satisfies the minimal distance.

4.3 Computation of the discontinuity times and points

Let us assume in the following that an inequality constraint is activated/deactivated in the time interval $[t^n, t^{n+1}]$. The computation of the exact time of discontinuity follows [47, 62] and introduces the partial time step as an (unknown) additional variable, together with the additional event function equation.

Let us denote by W the function describing the event location. This function depends directly on the dense output $\mathbf{U}^n$ defined on the interval $[t^n, t^{n+1}]$. Let us denote by $\tau \in [t^n, t^{n+1}]$ the time $\tau = t^n + h^n$, which is the root of the function W. The problem corresponds therefore to finding $(\mathbf{Y}^{n+1}, h^n)$, satisfying:

$$\begin{aligned} M(\mathbf{Z}^i - \mathbf{Y}^n) &= h^n \sum_{j=1}^{q} a_{ij}\,\mathbf{F}(\mathbf{Z}^j), \quad \forall i = 1, \ldots, q, && (4.3.1)\\ M(\mathbf{Y}^{n+1} - \mathbf{Y}^n) &= h^n \sum_{j=1}^{q} b_j\,\mathbf{F}(\mathbf{Z}^j), && (4.3.2)\\ W(\mathbf{U}^n(t^n + h^n)) &= 0. && (4.3.3) \end{aligned}$$

Following [47], a *splitting algorithm* is advocated, that couples the RADAU5 algorithm together with a bisection method. It is summarized as follows.

Algorithm 4.3.1. *At each time step t^n such that an activation/deactivation is detected in $[t^n, t^{n+1}]$, consider the system (4.3.1)-(4.3.3) and solve it as follows:*

- *(i) compute $h_0^n = \theta h^n$ as the root of $W(\mathbf{U}^n(t^n + \theta h^n)) = 0$, where $\mathbf{U}^n(t)$ is the dense output obtained from the solution of (4.3.1)-(4.3.2);*
- *(ii) for $k = 0, 1, \ldots$ until convergence*
  - *(a) solve (4.3.1)-(4.3.2) with $h^n = h_k^n$; this yields a dense output $\mathbf{U}_k^n(t^n + \theta h_k^n)$ for $\theta \in [0, 1]$;*
  - *(b) with $\mathbf{U}^n$ replaced by $\mathbf{U}_k^n$ compute h_{k+1}^n with a bisection method applied to (4.3.3);*
- *(iii) terminate the iterations with a step of (4.3.1)-(4.3.2).*

The convergence criterion is based on the difference between two successive step lengths h_k^n, i.e. $|h_{k+1}^n - h_k^n| < \varepsilon_c$, where ε_c is a given prescribed tolerance.

The addition of the time step as an unknown in (4.3.1)-(4.3.3) [47] allows to avoid the numerical error due to the dense output formula and to recover the full accuracy of the method. Furthermore the choice of the splitting algorithm for the resolution of (4.3.1)-(4.3.3) allows for a simple implementation, because the subsystem (4.3.1)-(4.3.2) is solved anyway at each step of the time stepping procedure.

The event function W is defined explicitly for the case of an activation and a deactivation.

The case of the activation of a constraint

A constraint is activated if there exists $\bar{\alpha} \in \mathcal{I}(t^n)$ such that corresponds to $y_{\bar{\alpha}}^n > 0$ and $y_{\bar{\alpha}}^{n+1} < 0$. Hence a natural definition for W is to set $W(\mathbf{U}^n(t^n + h^n)) = \mathbf{U}_{y_{\bar{\alpha}}}^n(t^n + h^n)$ where $\mathbf{U}_{y_{\bar{\alpha}}}^n$ is the component of $\mathbf{U}^n$ relative to the variable $y_{\bar{\alpha}}$.

The case of the deactivation of a constraint

When there exists $\bar{\alpha} \in \mathcal{A}(t^n)$ such that the distance between $(\mathbf{x}_{\bar{\alpha}}^{n+1}, g(\mathbf{x}_{\bar{\alpha}}^{n+1}))$ and the supporting tangent plane defined by the normal vector $\boldsymbol{\lambda}^{n+1}$ is negative, set

$$\begin{aligned} W(\mathbf{U}^n(t^n+h^n)) &= g(\mathbf{x}_{\bar{\alpha}}^{\boldsymbol{\lambda}}(t^n+h^n)) + \mathbf{U}_{\boldsymbol{\lambda}}^{n,T}(t^n+h^n)\,\mathbf{x}_{\bar{\alpha}}^{\boldsymbol{\lambda}}(t^n+h^n) \\ \text{with} \quad & \mathbf{e}^T\mathbf{x}_{\bar{\alpha}}^{\boldsymbol{\lambda}}(t^n+h^n) - 1 = 0, \end{aligned} \tag{4.3.4}$$

where $\mathbf{U}_{\boldsymbol{\lambda}}^n$ is the subvector of $\mathbf{U}^n$ relative to the variable $\boldsymbol{\lambda}$ and $\mathbf{x}_{\bar{\alpha}}^{\boldsymbol{\lambda}}$ is the point that minimizes the distance to the supporting tangent plane defined by the normal vector $\mathbf{U}_{\boldsymbol{\lambda}}^n(t^n+h^n)$. The expression of W resumes the definition of the distance d between the graph of the energy function and the tangent plane but, unlike in (3.2.4), $\mathbf{U}_{\boldsymbol{\lambda}}^n(t^n+h^n)$ is also an unknown in (4.3.4). Hence during the bisection steps of the Algorithm 4.3.1 for each $\mathbf{U}_{\boldsymbol{\lambda}}^n(t^n+\theta h_k^n)$ the minimization problem (3.2.4) is solved with $\mathbf{U}_{\boldsymbol{\lambda}}^n(t^n+\theta h_k^n)$ instead of $\boldsymbol{\lambda}^{n+1}$ in order to determine $\mathbf{x}_{\bar{\alpha}}^{\boldsymbol{\lambda}}$.

After the computation of the activation or deactivation time, all variables in $\mathbf{Y}$ are reinitialized to their value at time $t = \tau$ thanks to (iii) in Algorithm 4.3.1. The differential algebraic system (4.1.3) (or (4.1.4)) is then updated by moving the index $\bar{\alpha}$ from the set $\mathcal{I}(\tau)$ into the set $\mathcal{A}(\tau)$ or vice-versa. The complete algorithm is summarized as follows

Algorithm 4.3.2 (Summary of Complete Algorithm)**.** *For a fixed number of inactive inequality constraints, solve (4.1.4) with the RADAU5 algorithm. At each time step* t^{n+1}*:*

(i) Verify if an inactive constraint has to be activated by checking the sign of $y_{\bar{\alpha}}^{n+1}$ *for all* $\bar{\alpha} \in \mathcal{I}(t^{n+1})$*. If so, stop RADAU5 and compute the time of activation* τ *with the Algorithm 4.3.1, and the new size of (4.1.4). Restart the time-discretization scheme RADAU5 at* $t = \tau$ *without checking for a deactivation of constraints.*

(ii) Verify if an active constraint has to be deactivated by computing $\forall \alpha \in \mathcal{A}(t^{n+1})$, $\mathbf{x}_{\alpha}^{n+1}$ *as the minimizer of the distance to the supporting tangent plane with the Algorithm 3.2.1 and checking if the distance is negative. If so, stop RADAU5 and compute the time of deactivation* τ *with the Algorithm 4.3.1, and the new size of (4.1.4). Restart the time-discretization scheme RADAU5 at* $t = \tau$.

The adaptive time-step procedure avoids that both activation and deactivation happen during one time step.

Remark 4.3.1. *When* $r > 1$*, the computation of the point satisfying the minimal distance to the supporting tangent plane in Algorithm 3.2.1 strongly depends on the topology of the energy function* g*. In order to improve the robustness of the algorithm and avoid to miss a time of deactivation, the number of inactive constraints obtained by the RADAU5 algorithm may be compared with the number of actual inactive constraints computed by using the interior-point method described in Chapter 2. The robust version of the algorithm returns back a few time steps when a mismatch is detected.*

4.4 Numerical Results

The same numerical examples as for the optimization-based numerical method are considered. The graphical representation and the associated color code of the results are identical to the one presented in Chapter 3. The numerical parameters typically used are as follows: $\varepsilon_c = 10^{-7}$ and for the RADAU5 method the absolute and relative error tolerances are respectively equal to 10^{-13} and 10^{-7}.

4.4.1 Numerical results in one dimension

The chemical system composed of pinic acid ($C_9H_{14}O_4$) and 1-hexacosanol ($C_{26}H_{54}O$) at temperature 298.15 K and pressure 1 atm is considered ($s = 2$).

Figure 4.1 (left) shows the time evolution of the vector $\mathbf{b}$ on the phase diagram Δ_1. For more visibility the approximations $\mathbf{b}^n$ are lying on an axis situated just above the phase diagram. The phase diagram is given by

$$\Delta_1 = [0,\, 0.06656724] \cup]0.06656724,\, 0.463349192[\cup [0.463349192,\, 1],$$

The initial point $\mathbf{b}^0$ is situated in the left convex region of the phase diagram and one constraint is inactive ($y_1 > 0$ and $y_2 = 0$), then $\mathbf{b}^n$ moves from left to right. The corresponding iterates $g(\mathbf{b}^n)$, moving on the convex envelope of g, and the corresponding supporting tangent planes are also represented.

The time evolution of the $\mathbf{b}^n$, $n = 0, 1, \ldots$, with the color distinction follows the phase diagram correctly. First approximations are single-phase points, and the corresponding tangent planes are tangent to the curve g at only one point and lie below g. When $\mathbf{b}$ comes closer to the deactivation, the tangent planes come near a second contact point with g. At the moment of the deactivation the supporting plane is tangent to g at 2 points ($x_{1,1} = 0.0665672398$ and $x_{2,1} = 0.463349192$). These two points are accurate approximations of the points situated at the boundaries of the area on Δ_1 where both constraints are inactive. A zoomed-in view of the deactivation on g is proposed in Figure 4.1 (middle). After the deactivation, the points $g(\mathbf{b}^n)$ follow the convex envelope of g. Furthermore the tangent planes touch g at two points and are confused with the convex envelope of g. Figure 4.1 (right) illustrates the time evolution of y_1^n and y_2^n, $n = 0, 1, \ldots$, and exhibits a discontinuity of the derivatives at time $t = 0.372487544$ s when the second inequality constraint is deactivated, for *both* of the variables.

Figure 4.2 uses the same notations as in Figure 4.1 to illustrate the time evolution of the $\mathbf{b}^n$ (left), and y_1^n and y_2^n (right) when one inequality constraint is activated, namely when $\mathbf{b}^n$ moves from the middle of Δ_1 to the extreme right of the phase diagram. The color of the diamonds representing the approximations $\mathbf{b}^n$ follows the phase diagram. Furthermore the point $\mathbf{b}^n$ for which the activation occurs is located on the frontier of Δ_1 between the area 2 and the area 1. After the activation, the tangent planes get released from g and remains below g.

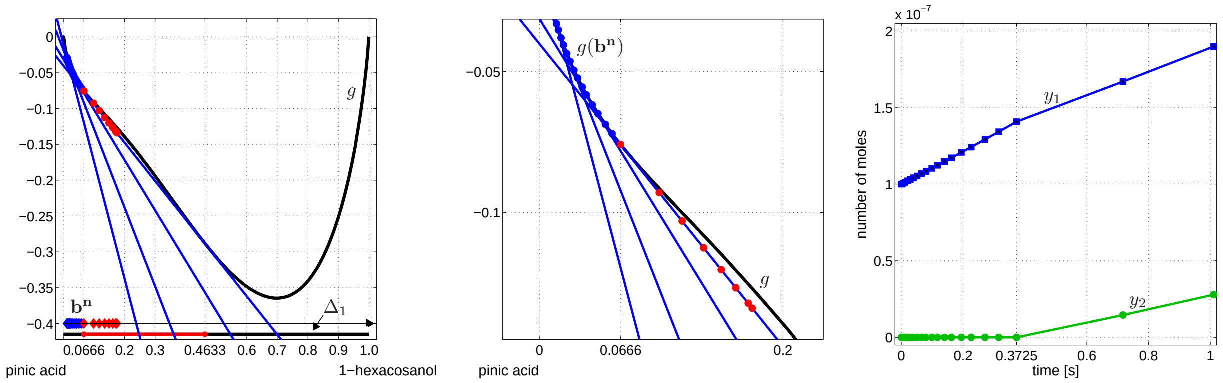


Figure 4.1: Organic aerosol made of 1-hexacosanol and pinic acid with initial composition-vector $\mathbf{b}_0^T = (0.01 \cdot 10^{-7},\, 0.99 \cdot 10^{-7})$ mol. Left: evolution of $\mathbf{b}^n$, $n = 0, 1, 2, \ldots$, the corresponding supporting tangent plane evolves until making contact with the graph of g. Middle: zoomed-in view of the deactivation. Right: time evolution of y_1^n and y_2^n, $n = 0, 1, 2, \ldots$.

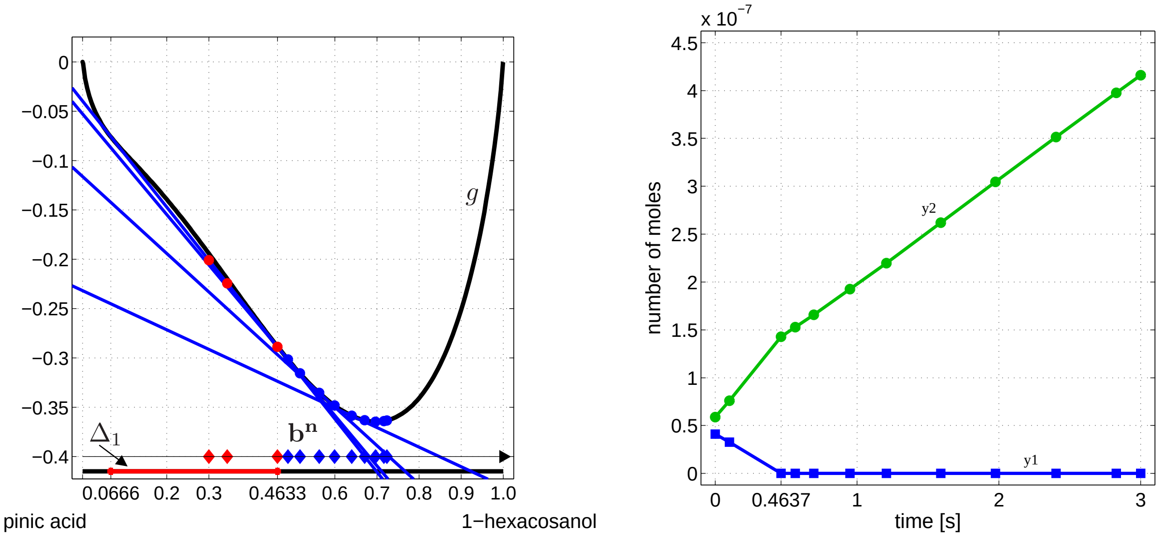


Figure 4.2: Organic aerosol made of 1-hexacosanol and pinic acid with initial composition-vector $\mathbf{b}_0^T = (3 \cdot 10^{-8},\, 7 \cdot 10^{-8})$ mol. Left: evolution of $\mathbf{b}^n$, $n = 0, 1, 2, \ldots$, the corresponding supporting tangent plane evolves after leaving the contact with the left convex region on the graph of g. Right: time evolution of y_1^n and y_2^n, $n = 0, 1, 2, \ldots$.

In Figures 4.3 and 4.4 two other examples are illustrated. The difficulty of the first example is the small size of the left area 1 on the phase diagram which is given by

$$\Delta_1 = [0,\, 0.206110058 \cdot 10^{-11}] \cup]0.206110058 \cdot 10^{-11},\, 0.898380832[\cup [0.898380832,\, 1].$$

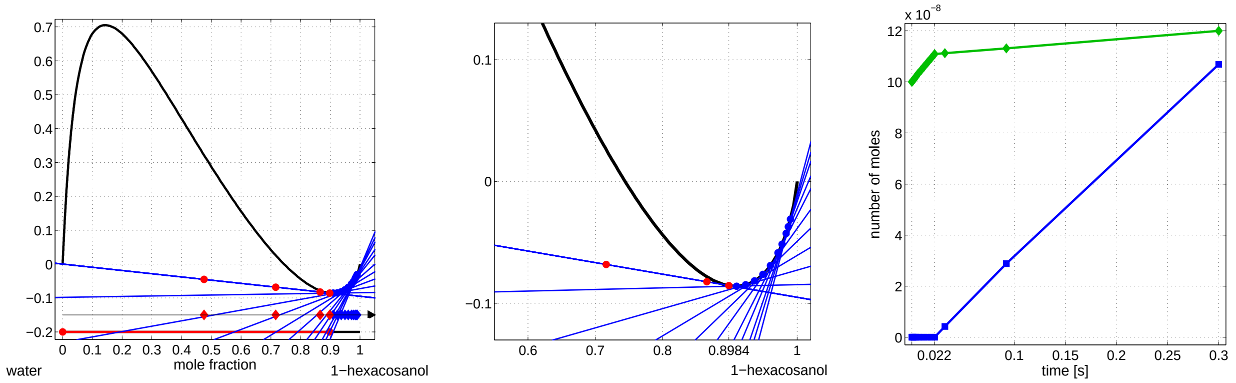


Figure 4.3: Organic aerosol made of water and 1-hexacosanol with initial composition-vector $\mathbf{b}_0^T = (9.9 \cdot 10^{-8},\, 0.1 \cdot 10^{-8})$ mol. The deactivation occurs at $t = 0.0220509482$ s. Left: evolution of $\mathbf{b}^n$, $n = 0, 1, 2, \ldots$ and the corresponding supporting tangent planes. Middle: zoomed-in view of the deactivation. Right: evolution of y_1^n and y_2^n, $n = 0, 1, 2, \ldots$.

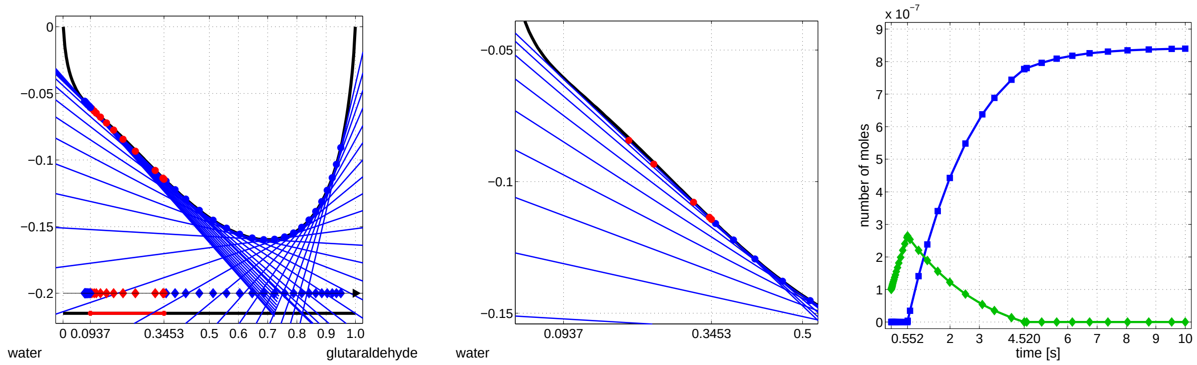


Figure 4.4: Organic aerosol made of water and glutaraldehyde with initial composition-vector $\mathbf{b}_0^T = (9.5 \cdot 10^{-8},\, 0.5 \cdot 10^{-8})$ mol. The deactivation occurs at $t = 0.542032077$ s and the activation at $t = 4.51698170$ s. Left: evolution of $\mathbf{b}^n$, $n = 0, 1, 2, \ldots$ and the corresponding supporting tangent planes. Middle: zoomed-in view of the deactivation. Right: evolution of y_1^n and y_2^n, $n = 0, 1, 2, \ldots$.

The second example consists of an energy function g whose convex envelope is very close to g. The corresponding phase diagram is

$$\Delta_1 = [0,\, 0.0936539076] \cup]0.0936539076,\, 0.345326525[\cup [0.345326525,\, 1].$$

In both figures the initial composition-vector is located in the right extremity of Δ_1 and the approximations $\mathbf{b}^n$, $n = 0, 1, 2, \ldots$, move to the left during the simulation. The approximations $\mathbf{b}^n$ follow the phase diagram correctly as well as their images $g(\mathbf{b}^n)$ with the convex envelope of g. Moreover the Gibbs tangent plane criterion is satisfied for the whole simulation of each example.

Let $\mathbf{z}_b$ denote the points on the phase diagram situated at the frontier between the areas 1 and 2, and call these points *boundary points*. For instance the boundary points of the phase diagram Δ_1 associated to the first example are 0.06656724 and 0.463349192. Moreover let $\mathbf{x}_b$ denote the point whose image through the projection P is equal to $\mathbf{z}_b$, namely: $\mathbf{z}_b = P\mathbf{x}_b$. The point $\mathbf{x}_b$ is also called *boundary point*.

In Table 4.1 the computed discontinuity points are listed for each example. Since in the fourth example a deactivation and an activation occur, two discontinuity points are given. The exact discontinuity point (i.e. those from the exact solution) hit one of the boundary point of the phase diagram, depending on the trajectory of $\mathbf{b}$. Hence the error between the computed and exact discontinuity points can be estimated by calculating the discrepancy between the computed discontinuity point and the appropriate boundary point. This error (in the Euclidean norm) is listed in the third column of Table 4.1. The range of the discrepancy is from 10^{-13} to 10^{-10}. Thus the computation of the discontinuity point is done accurately in all examples. Furthermore the evolution of $\mathbf{b}^n$ is correct and the Gibbs tangent plane criterion is always satisfied.

Example	**discontinuity point d**	**error $\|\mathbf{b} - \mathbf{x}_b\|_2$**
Figure 4.1	0.0665672395	$7.06{\cdot}10^{-11}$
Figure 4.2	0.463349193	$6.66{\cdot}10^{-10}$
Figure 4.3	0.898380832	$4.84{\cdot}10^{-13}$
Figure 4.4	0.345326525	$2.85{\cdot}10^{-11}$
	0.0936539078	$2.87{\cdot}10^{-10}$

Table 4.1: The reduced discontinuity point $\mathbf{d} = P\mathbf{b}$ and the error between the computed discontinuity point $\mathbf{b}$ and the boundary point $\mathbf{x}_b$ in the Euclidian norm.

In conclusion the method based on differential algebraic systems is more accurate than the optimization-based method for the case $s = 2$. Another comparison between both methods can be established by analyzing the number of time steps needed for the simulation. This number is illustrated in Table 4.2 for each method and each example. The step size selection of RADAU5 significantly decreases the number of time steps and makes the second approach more efficient.

Example	**opt.-based method**	**DAE-based method**
Figure 4.1	100	19
Figure 4.2	30	11
Figure 4.3	30	14
Figure 4.4	100	42

Table 4.2: Number of time steps executed by the optimization-based resolution method and the method based on the differential algebraic systems on each example.

4.4.2 Numerical results in two dimensions

The chemical system composed of pinic acid ($C_9H_{14}O_4$), 1-hexacosanol ($C_{26}H_{54}O$) and water (H_2O) at temperature 298.15 K and pressure 1 atm is considered ($s = 3$). The solution $\mathbf{b}$ and its numerical approximation are represented on a two-dimensional simplex Δ_2 [3, 52, 55]. As in Section 3.5 two classes of interaction parameters are considered, leading to two different phase diagrams depicted in Figure 3.12. The regions of the simplices with respectively one, two or three deactivated constraints are numbered by $1, 2, 3$ on the simplices. Let us begin with the numerical examples on the phase diagram VL generated by the interaction parameters of [51].

Examples on the phase diagram VL

Figure 4.5 illustrates the solution of the initial value problem whose initial conditions are given by

- composition-vector: $\mathbf{b}_0^T = (1.5 \cdot 10^{-8},\ 8.0 \cdot 10^{-8},\ 0.5 \cdot 10^{-8})$ mol,
- gas concentration-vector: $\mathbf{c}_{g,0}^{\infty,T} = (3.5,\ 3,\ 3.5)$ mol/m^3,
- initial time step: $h_0 = 0.1$ s.

Figure 4.5 (left) shows two simulated trajectories of $\mathbf{b}$, one with tracking of discontinuities (colored line) and the other without the tracking (black line). The colored trajectory undergoes two deactivations and one activation of constraints, whereas the black one stands for approximations $\mathbf{b}^n$ that wrongly remain single-phase points during the whole simulation.

Figure 4.5 (left) demonstrates that the tracking of such events strongly influences the solution of the initial value problem. Figure 4.5 (middle) is a zoomed-in view on the phase diagram that illustrates how the trajectories move away from each other after the first deactivation. Eventually both trajectories converge to the unique stationary solution of the closed gas-aerosol system. Figure 4.5 (middle) emphasizes the importance to detect and compute the discontinuity points accurately.

Figure 4.5 (right) illustrates the evolution of y_α^n, $\alpha = 1, 2, 3$ and $n = 0, 1, 2, \ldots$, the number of moles relative to each liquid phase $\mathbf{x}_\alpha^n$ present in the aerosol. At $t = 0$ s, $y_1^0 = y_2^0 = 0$ and $y_3^0 > 0$, and two constraints are activated (*i.e.* the particle only contains the third liquid phase). Then constraints are activated/deactivated and the trajectory of y_α^n, $\alpha = 1, 2, 3$ and $n = 0, 1, 2, \ldots$, present jumps of the derivatives at each event.

Figures 4.6 and 4.7 illustrate two other examples on the phase diagram VL. The trajectory of $\mathbf{b}^n$ for the first example evolves close to the phase boundary between the areas 1 and 2, whereas the second example is concerned with a trajectory that crosses the region on the phase diagram situated at the intersection between the areas 1, 2 and 3. The initial conditions for each example are given by

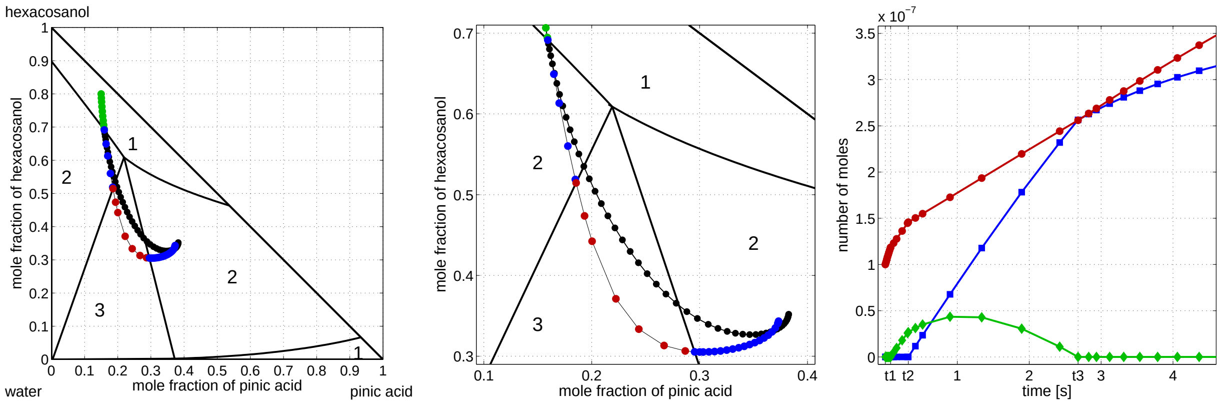
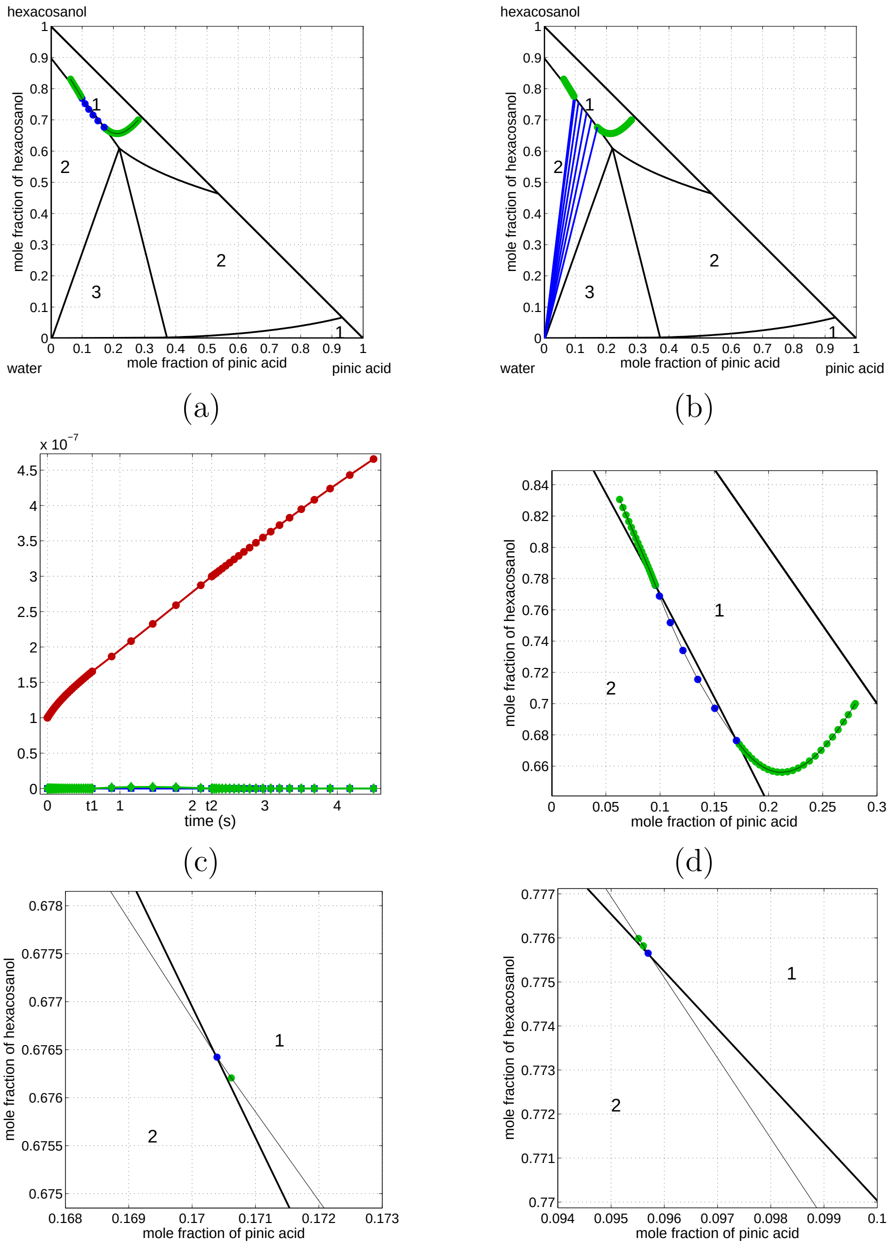


Figure 4.5: Left: evolution of $\mathbf{b}^n$, $n = 0, 1, 2, \ldots$, on the phase diagram of the particle without the tracking of the discontinuity points (black line) and with the tracking (colored line). Middle: zoomed-in view. Right: time evolution of the number of moles relative to each liquid phase present in the particle.

Figure 4.6		
$\mathbf{b}_0$	$\mathbf{c}_{g,0}^{\infty}$	h_0
$2.8 \cdot 10^{-8}$	0.02	
$7.0 \cdot 10^{-8}$	7.3	1.0
$0.2 \cdot 10^{-8}$	1.68	

Figure 4.7		
$\mathbf{b}_0$	$\mathbf{c}_{g,0}^{\infty}$	h_0
$1.5 \cdot 10^{-8}$	2.85	
$8.0 \cdot 10^{-8}$	4.2	0.01
$0.5 \cdot 10^{-8}$	1.95	

As for the optimization-based method, the numerical results for both examples are consistent with the respective phase diagrams. Moreover comparing these numerical results with those of the first method, one can deduce a reduction in the number of time steps needed in the simulation for the second method.



(a) (b)

(c) (d)

(e) (f)

Figure 4.6: Example VL with the initial conditions $\mathbf{b}_0^T = (2.8 \cdot 10^{-8},\ 7.0 \cdot 10^{-8},\ 0.2 \cdot 10^{-8})$ mol and $\mathbf{c}_{g,0}^{\infty,T} = (0.02,\ 7.3,\ 1.68)$ mol/m^3. A deactivation occurs at $t^1 = 0.61742613$ s and an activation at $t^2 = 2.26623802$ s. Time evolution of (a) $\mathbf{b}^n$, $n = 0, 1, 2, \ldots$ on the phase diagram, (b) the phase simplices on the phase diagram; (c) the number of moles y_α^n, $\alpha = 1, 2, 3$ and $n = 0, 1, 2, \ldots$. The graphs (d), (e) and (f) are zoomed-in views of the trajectory of the $\mathbf{b}^n$, $n = 0, 1, 2, \ldots$ near the phase boundary between the areas 2 and 1.

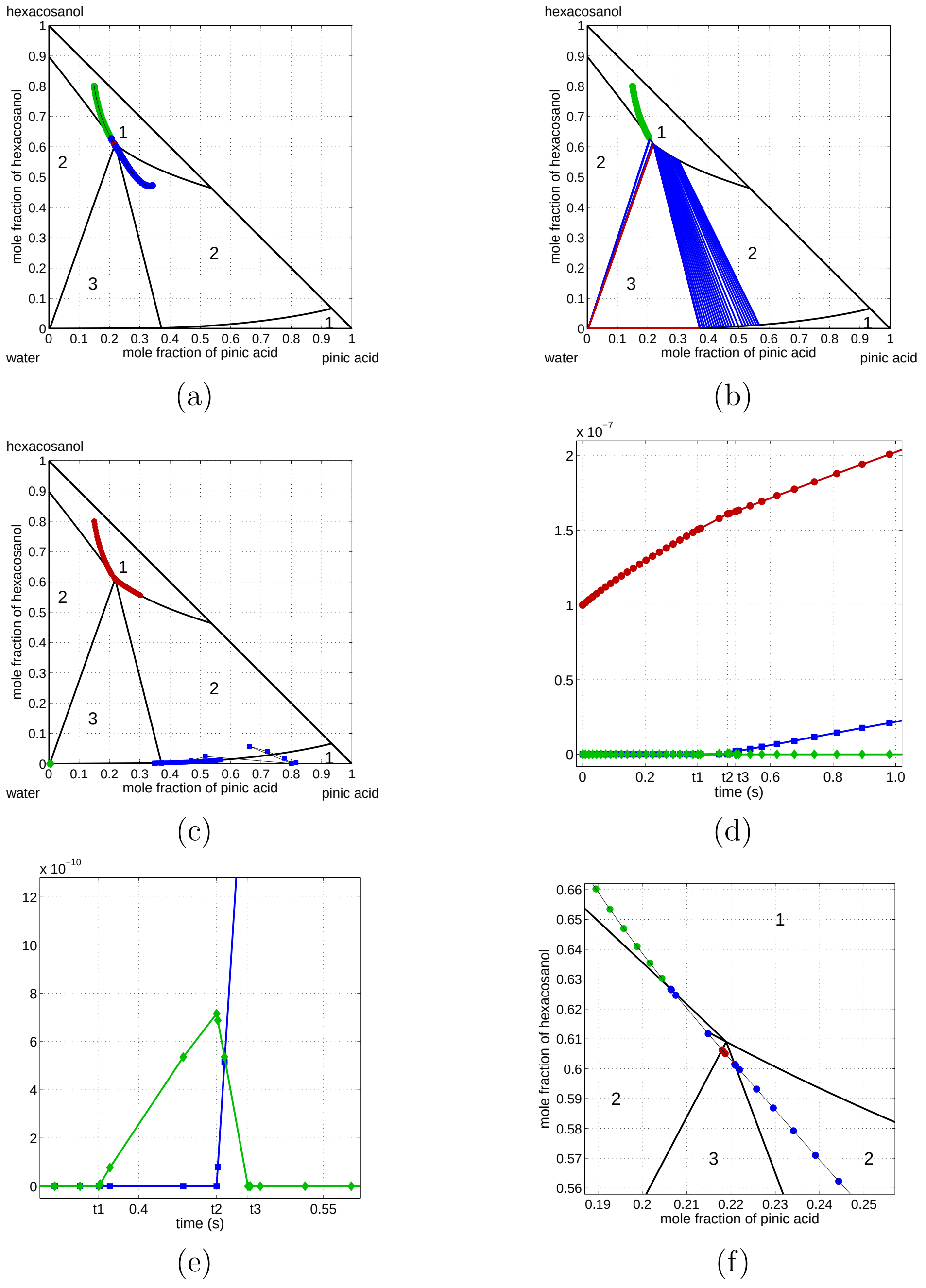


Figure 4.7: Example VL with the initial conditions $\mathbf{b}_0^T = (1.5{\cdot}10^{-8},\ 8.0{\cdot}10^{-8},\ 0.5{\cdot}10^{-8})$ mol and $\mathbf{c}_{g,0}^{\infty,T} = (2.85,\ 4.2,\ 1.95)$ mol/m^3. The deactivations occur at $t^1 = 0.367736798$ s and $t^2 = 0.463307013$ s, and the activation at $t^3 = 0.488743211$ s. Time evolution of (a) $\mathbf{b}^n$, $n = 0, 1, 2, \ldots$ on the phase diagram, (b) the phase simplices on the phase diagram; (c) the mole-fraction vectors $\mathbf{x}_\alpha^n$, $\alpha = 1, 2, 3$ and $n = 0, 1, 2, \ldots$, on the phase diagram; (d) the number of moles y_α^n, $\alpha = 1, 2, 3$ and $n = 0, 1, 2, \ldots$. The graphs (e) and (f) are respectively zoomed-in views of (e) the number of moles y_α^n, $\alpha = 1, 2, 3$ and $n = 0, 1, 2, \ldots$, and (f) the trajectory of $\mathbf{b}^n$, $n = 0, 1, 2, \ldots$.

Examples on the phase diagram LL

Let us consider some examples on the phase diagram LL. The initial conditions of the first example are

- composition-vector: $\mathbf{b}_0^T = (1.0 \cdot 10^{-8},\, 4.0 \cdot 10^{-8},\, 5.0 \cdot 10^{-8})$ mol,
- gas concentration-vector: $\mathbf{c}_{g,0}^{\infty,T} = (2.0,\, 1.0,\, 7.0)$ mol/m^3,
- initial time step: $h = 0.001$ s.

The simulation starts from the left area 2 and enters in the area 3 as it is depicted in Figure 4.8 (a). However the color of the trajectories $\mathbf{b}^n$, $n = 0, 1, 2, \ldots$ in Figure 4.8 (a) and the evolution of the phase simplices in Figure 4.8 (b) show that the deactivation of the constraint 1 is not detected. The evolution of $\mathbf{x}_1^n$, $n = 0, 1, 2, \ldots$, is represented in Figure 4.8 (c) and one can observe that $\mathbf{x}_1^n$ tends to $\mathbf{x}_2^n$ instead of the activation point situated at the bottom left corner of the phase simplex of dimension 3. Consequently no deactivation can be detected and the method based on differential algebraic systems fails on this example as the first method did.

Let us change the initialization of $\mathbf{x}_1$ as it has been done for the first method, namely

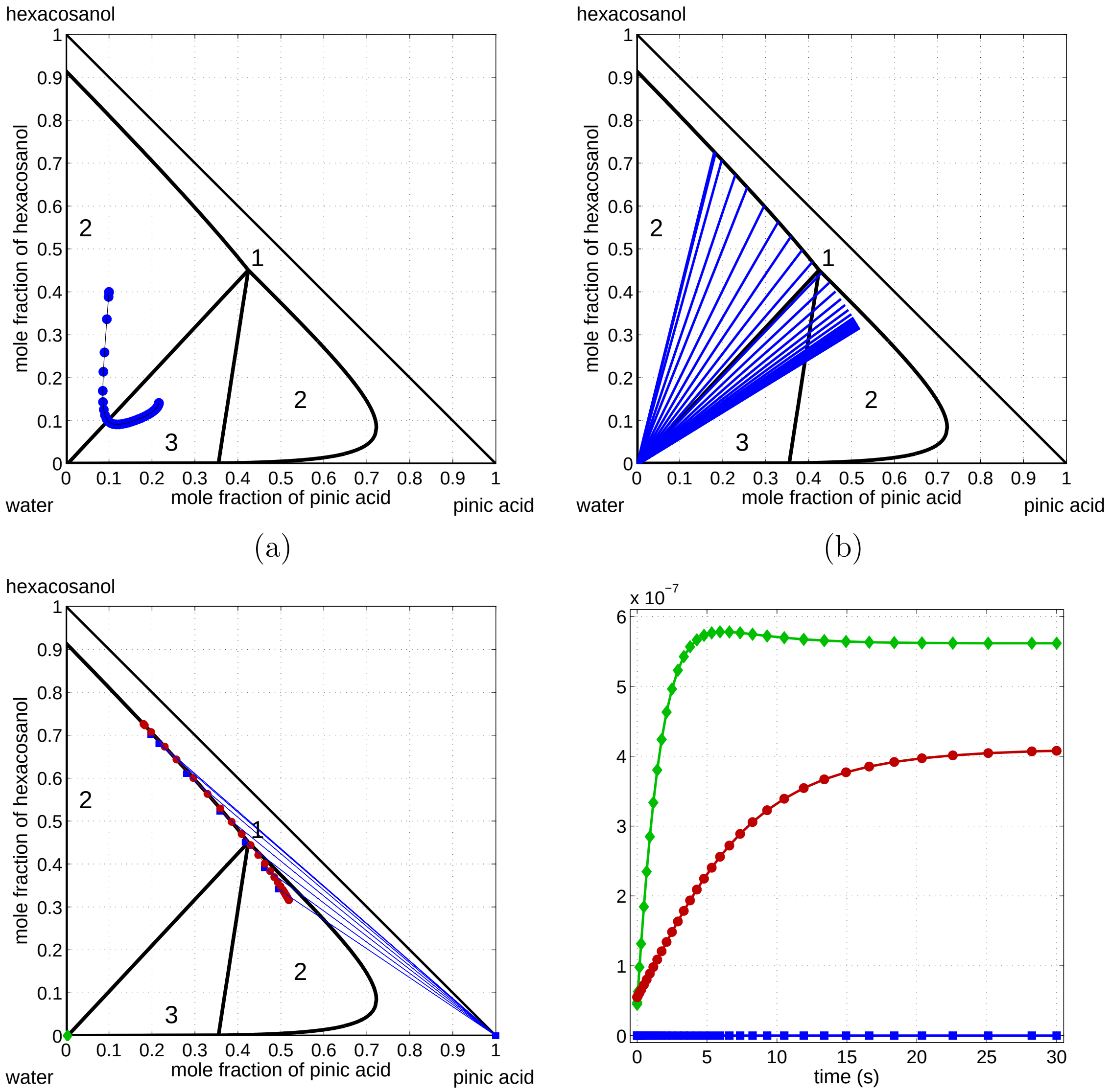
$$\mathbf{x}_1^{0,T} = (0.7,\, 3.0 \cdot 10^{-8},\, 0.29999997),$$

and run the simulation again. The results are given in Figure 4.9. With this new initialization the deactivation is detected and correctly computed.

Now let us study a second example. The initial conditions are

- composition-vector: $\mathbf{b}_0^T = (9.0 \cdot 10^{-8},\, 0.5 \cdot 10^{-8},\, 0.5 \cdot 10^{-8})$ mol,
- gas concentration-vector: $\mathbf{c}_{g,0}^{\infty,T} = (4.0,\, 2.0,\, 4.0)$ mol/m^3,
- initial time step: $h = 0.001$ s.

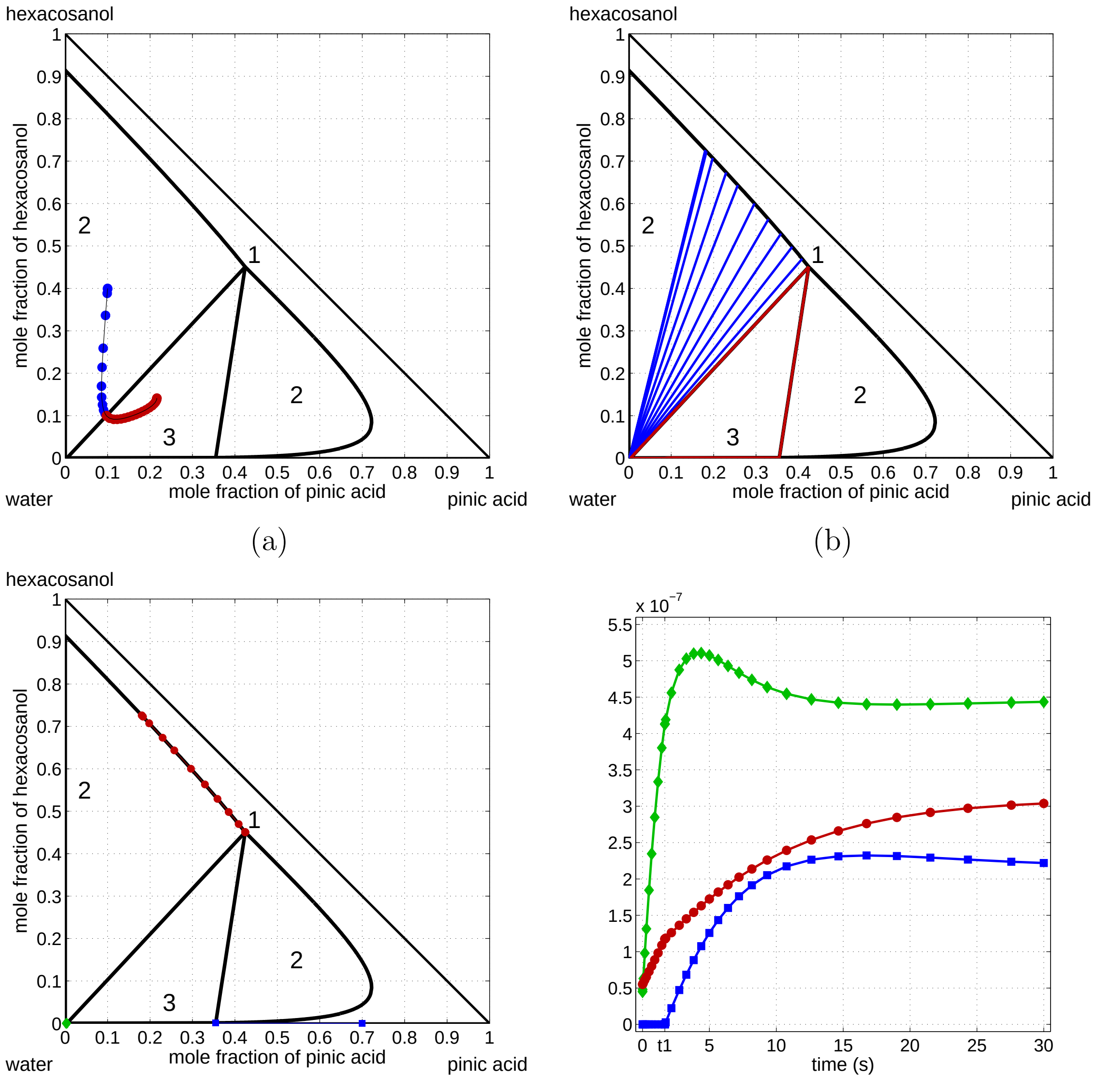
The optimization-based method succeeded in the computation of the deactivation but failed in the restart of the simulation, the interior-point method activating the freshly inactive inequality constraint. The numerical results with the second method are illustrated in Figure 4.10. The deactivation is correctly detected and computed. Furthermore the simulation restarts and is consistent with the phase diagram. The method is then more efficient on such example, namely when the solution $\mathbf{b}^n$ starts in area 1 and comes right in front and nearly perpendicular to the phase simplices of area 2. Finally let us also note the extremely large slope of the trajectories y_1^n and y_2^n, $n = 0, 1, 2, \ldots$ after the deactivation time.



(a) (b)

(c) (d)

Figure 4.8: Example LL with the initial conditions $\mathbf{b}_0^T = (1.0{\cdot}10^{-8},\ 4.0{\cdot}10^{-8},\ 5.0{\cdot}10^{-8})$ mol and $\mathbf{c}_{g,0}^{\infty,T} = (2.0,\ 1.0,\ 7.0)$ mol/m^3 when $\mathbf{x}_\alpha^0$, $\alpha = 1, 2, 3$ are initialized in the corners of the phase diagram. Time evolution of (a) $\mathbf{b}^n$, $n = 0, 1, 2, \ldots$; (b) the phase simplices; (c) the molar-fraction vectors $\mathbf{x}_\alpha^n$, $\alpha = 1, 2, 3$ and $n = 0, 1, 2, \ldots$; (d) the number of moles y_α^n, $\alpha = 1, 2, 3$ and $n = 0, 1, 2, \ldots$.



(a) (b) (c) (d)

Figure 4.9: Example LL with the initial conditions $\mathbf{b}_0^T = (1.0{\cdot}10^{-8},\ 4.0{\cdot}10^{-8},\ 5.0{\cdot}10^{-8})$ mol and $\mathbf{c}_{g,0}^{\infty,T} = (2.0,\ 1.0,\ 7.0)$ mol/m^3 when $\mathbf{x}_3^0$ is initialized closer to the bottom left vertex of the area 3. Time evolution of (a) $\mathbf{b}^n$, $n = 0, 1, 2, \ldots$; (b) the phase simplices; (c) the molar-fraction vectors $\mathbf{x}_\alpha^n$, $\alpha = 1, 2, 3$ and $n = 0, 1, 2, \ldots$; (d) the number of moles y_α^n, $\alpha = 1, 2, 3$ and $n = 0, 1, 2, \ldots$.

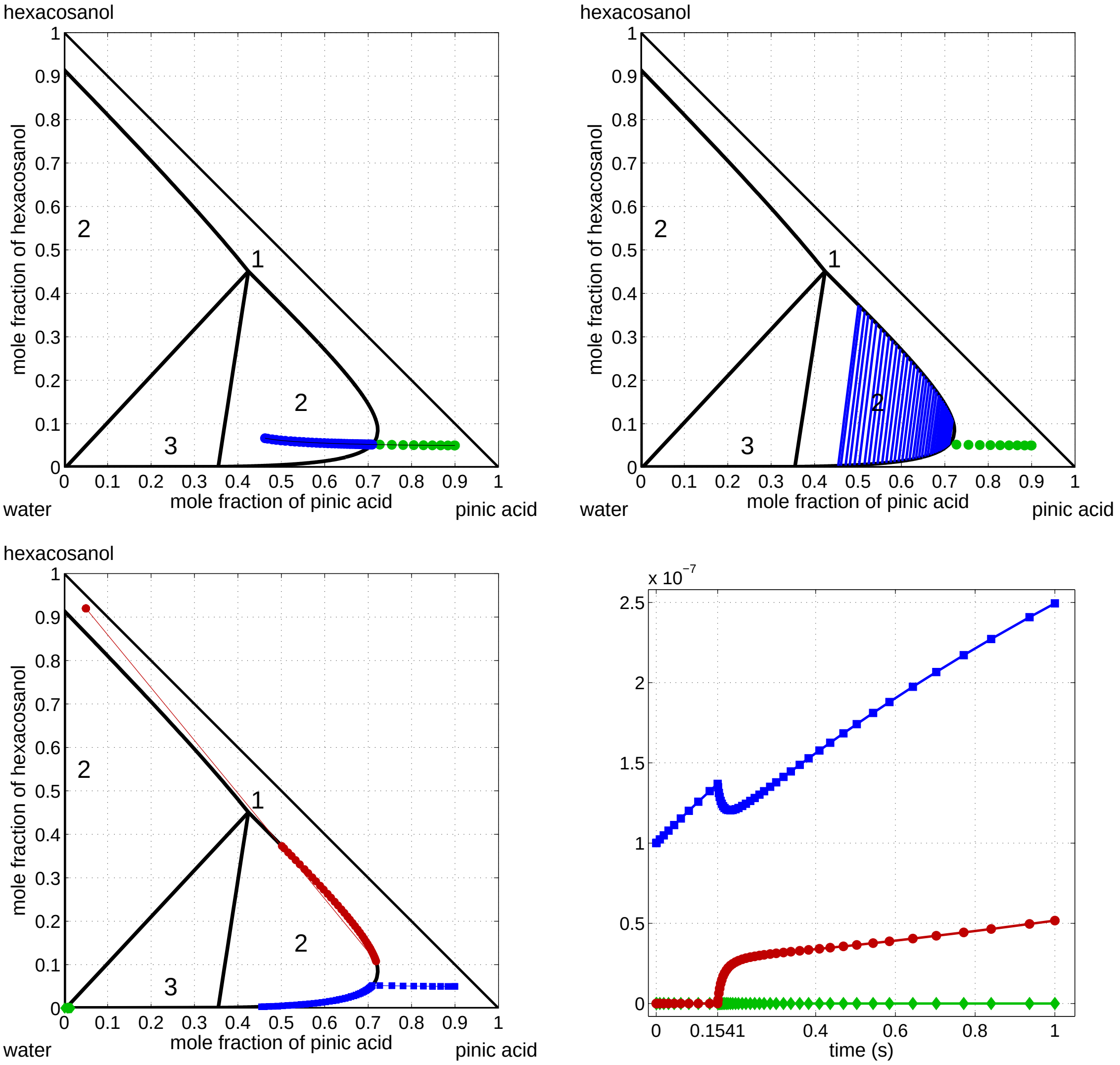


Figure 4.10: Example LL with the initial conditions $\mathbf{b}_0^T = (9.0 \cdot 10^{-8},\ 0.5 \cdot 10^{-8},\ 0.5 \cdot 10^{-8})$ and $\mathbf{c}_{g,0}^{\infty,T} = (4.0,\ 2.0,\ 4.0)$. Time evolution of (a) $\mathbf{b}^n$, $n = 0, 1, 2, \ldots$; (b) the phase simplices; (c) the molar-fraction vectors $\mathbf{x}_\alpha^n$, $\alpha = 1, 2, 3$ and $n = 0, 1, 2, \ldots$; (d) the number of moles y_α^n, $\alpha = 1, 2, 3$ and $n = 0, 1, 2, \ldots$.

Theoretical example

In the particular case when all y_α are strictly positive, the exact solution $\mathbf{b}$ and the exact time of activation t^* when an inequality constraint is activated, are known [18, 19]. Four different examples are considered starting all from the area 3 and going to one of the areas 2. All trajectories are represented in Figure 3.27 of Section 3.5. Figure 4.11 illustrates the error on the computation of activation points between the approximated and exact solutions for each example. It shows that the error on both the time and location of the activation is negligible, up to machine precision and algorithm tolerance, and validate the accuracy of this second method based on differential algebraic systems.

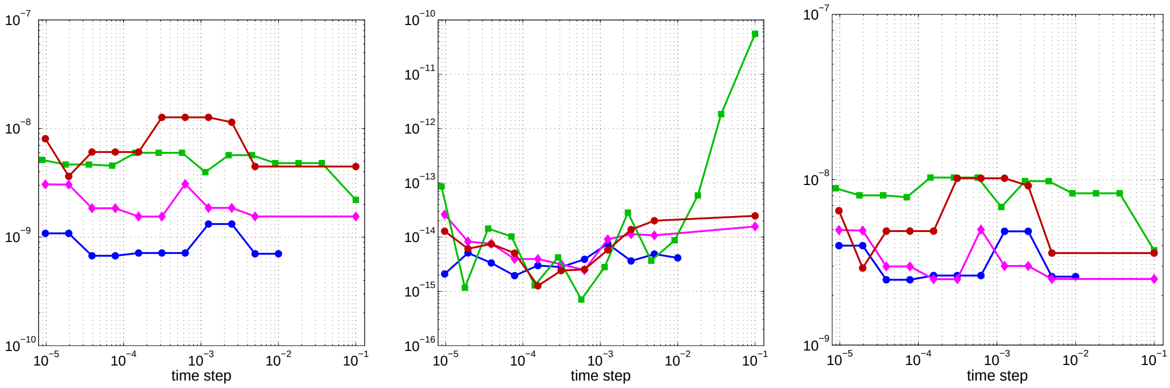


Figure 4.11: Error on the computation of the activation/deactivation of inequality constraints: the case of the activation of a constraint. Error on the time of activation $|t^\star - t^{n+1}|$ (left); $\|\mathbf{d}(t^n) - \mathbf{d}^n\|_2$ (middle) and error on the location of activation $\|\mathbf{d}(t^\star) - \mathbf{d}^n\|_2$ (right).

4.4.3 Numerical results in higher dimensions

When $s = 4$, no phase diagram is available. However the numerical results may still be visualized on a tetrahedron. Let us consider a gas-aerosol system made of pinic acid, 1-hexacosanol, water and n-propanol. In Section 3.5 two examples have been presented. Let us reconsider both examples and add a new one. The numerical results are respectively given in Figure 4.12, Figure 4.13 and Figure 4.14, and the corresponding initial conditions are

Figure 4.12		
$\mathbf{b}_0$	$\mathbf{c}^{\infty}_{g,0}$	h_0
$0.5 \cdot 10^{-8}$	0.02	1.0
$7.5 \cdot 10^{-8}$	0.5	
$1.0 \cdot 10^{-8}$	4.0	
$1.0 \cdot 10^{-8}$	1.5	

Figure 4.13		
$\mathbf{b}_0$	$\mathbf{c}^{\infty}_{g,0}$	h_0
$0.1 \cdot 10^{-8}$	7.0	1.0
$0.1 \cdot 10^{-8}$	2.0	
$9.3 \cdot 10^{-8}$	0.5	
$0.5 \cdot 10^{-8}$	0.5	

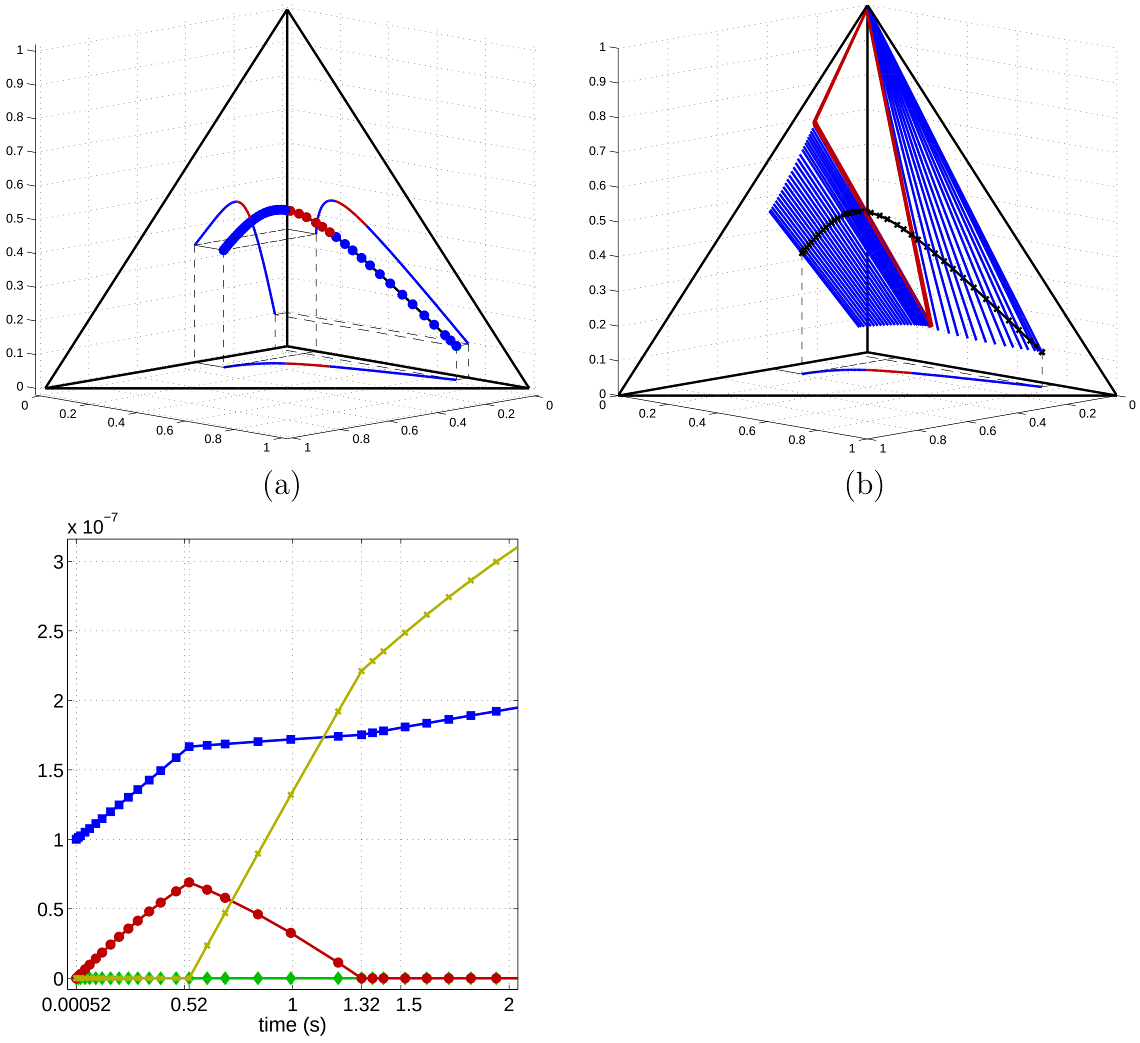
Figure 4.14		
$\mathbf{b}_0$	$\mathbf{c}^{\infty}_{g,0}$	h_0
$8.5 \cdot 10^{-8}$	0.5	1.0
$0.5 \cdot 10^{-8}$	4.5	
$0.5 \cdot 10^{-8}$	4.0	
$0.5 \cdot 10^{-8}$	1.0	

For each example the time evolution of $\mathbf{b}^n$, y_α^n, $\alpha = 1, 2, 3, 4$ and $n = 0, 1, 2, \ldots$, and the phase simplices look consistent with the results obtained in Chapter 3 and the activations/deactivations are correctly computed. The last example was not presented in Section 3.5. The reason is that the optimization-based method could not restart from the first deactivation because of the interior-point method that directly activates the new inactive inequality constraint.

When s is greater than 5, the phase diagrams can no more be visualized. In the following let us compare the CPU times for different values of s. Table 4.3 summarizes the computational times for several examples that were run with an Intel processor of 2.4 GHz and with 2 GB of RAM memory. The first 3 examples are the studied examples on the phase diagram VL, the 3 following examples are the above-mentioned examples with $s = 4$, and the last two are with $s = 18$ of Chapter 3. The presented CPU times illustrate respectively the total time of execution, the time for the detection of events, the time for the computation of the activation, the time spent in going backwards in the trajectory, the time for the computation of the deactivation and the total time for the detection and computation of the discontinuities.

Table 4.3 shows that the larger s, the more expensive the tracking of discontinuity points. However, the percentage of the computational cost for the tracking remains stable as s becomes larger. For all examples the number of iterations in the splitting Algorithm 4.3.1 is equal to 3 in average and the number of iterates for the bisection in the deactivation case is equal to 30 in average.

Comparing with the optimization-based method, the total CPU times is strongly reduced for this second method. This decrease essentially comes from the different methods to solve the optimization-constrained differential equations when the number of inactive constraints is fixed. Indeed the CPU time for the first method is mainly due to the fixed-point algorithm whereas the CPU time for the second method derives from the detection and the computation of the discontinuities.

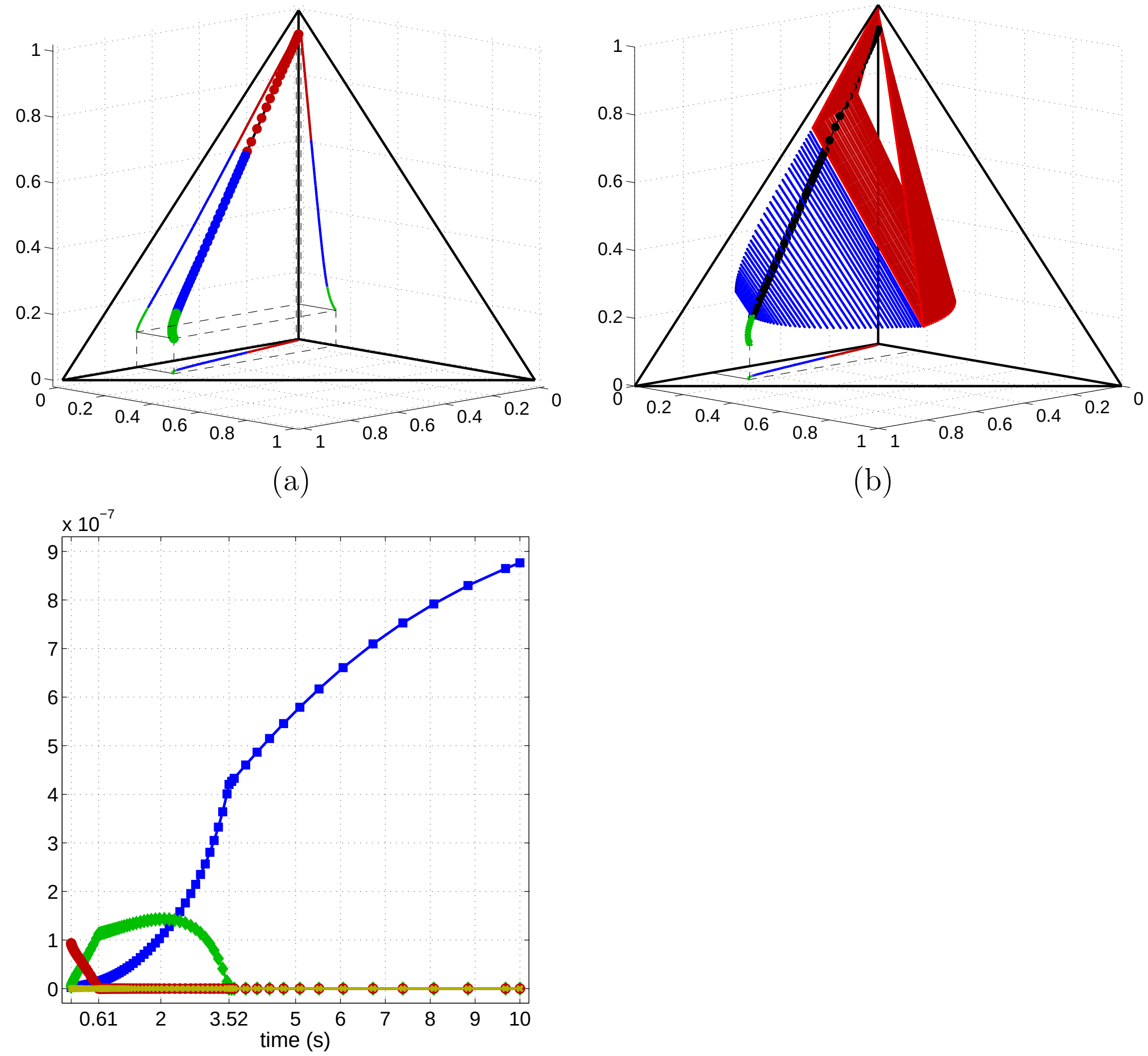


(a)

(b)

(c)

Figure 4.12: Example whose initial conditions are $\mathbf{b}_0^T = (0.5 \cdot 10^{-8},\ 7.5 \cdot 10^{-8},\ 1 \cdot 10^{-8},\ 1 \cdot 10^{-8})$ mol and $\mathbf{c}_{g,0}^{\infty,T} = (4.0,\ 0.5,\ 4.0,\ 1.5)$ mol/m^3. Time evolution of (a) $\mathbf{b}^n$, $n = 0, 1, 2, \ldots$; (b) the phase simplices and (c) y_α^n, $\alpha = 1, 2, 3, 4$ and $n = 0, 1, 2, \ldots$.



(a) (b)

(c)

Figure 4.13: Example whose initial conditions are $\mathbf{b}_0^T = (0.1\cdot 10^{-8},\ 0.1\cdot 10^{-8},\ 9.3\cdot 10^{-8},\ 0.5\cdot 10^{-8})$ mol and $\mathbf{c}_{g,0}^{\infty,T} = (7.0,\ 2.0,\ 0.5,\ 0.5)$ mol/m^3. Time evolution of (a) $\mathbf{b}^n$, $n = 0, 1, 2, \ldots$; (b) the phase simplices and (c) y_α^n, $\alpha = 1, 2, 3, 4$ and $n = 0, 1, 2, \ldots$.

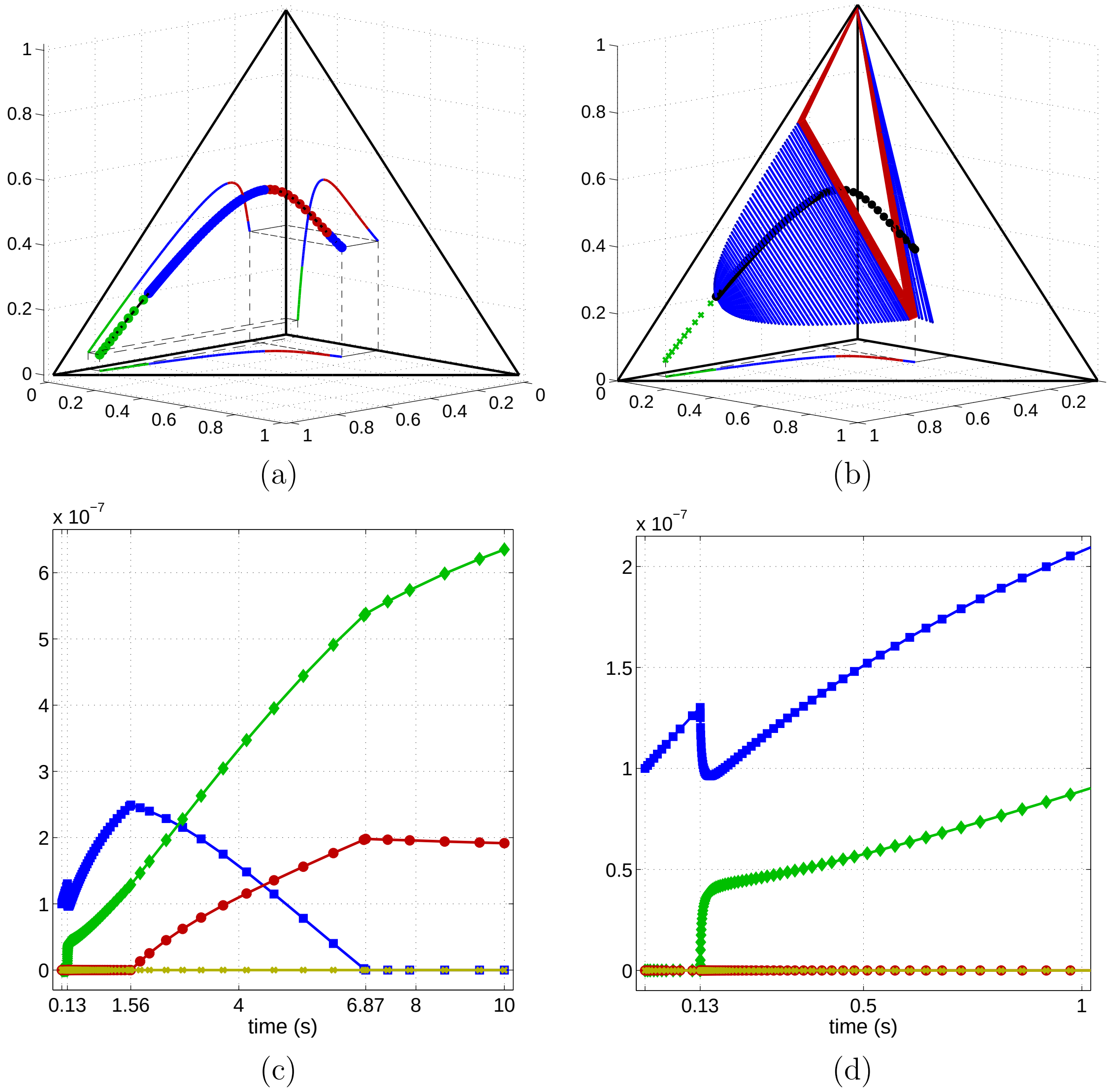


Figure 4.14: Example whose initial conditions are $\mathbf{b}_0^T = (8.5\cdot 10^{-8},\ 0.5\cdot 10^{-8},\ 0.5\cdot 10^{-8},\ 0.5\cdot 10^{-8})$ mol and $\mathbf{c}_{g,0}^{\infty,T} = (0.5,\ 4.5,\ 4.0,\ 1.0)$ mol/m^3. Time evolution of (a) $\mathbf{b}^n$, $n = 0, 1, 2, \ldots$; (b) the phase simplices and (c) y_α^n, $\alpha = 1, 2, 3, 4$ and $n = 0, 1, 2, \ldots$. The graph (d) is a zoomed-in view of y_α^n, $\alpha = 1, 2, 3, 4$ and $n = 0, 1, 2, \ldots$.

ex. on Figure 4.5	code	detect	act.	backwards	deact.	total disc.
temps [s]	0.10	0.01	0.02	0.02	0.03	0.08
%		5.3	15.2	16.0	25.3	61.8
ex. on Figure 4.6	code	detect	act.	backwards	deact.	total disc.
temps [s]	0.10	0.02	0.01	0.00	0.04	0.07
%		24.2	7.5	2.5	37.5	71.7
ex. on Figure 4.7	code	detect	act.	backwards	deact.	total disc.
temps [s]	0.10	0.02	0.02	0.01	0.01	0.06
%		19.9	19.9	13.4	8.0	61.2
ex. on Figure 4.12	code	detect	act.	backwards	deact.	total disc.
temps [s]	0.48	0.06	0.04	0.14	0.07	0.31
%		13.4	7.4	30.1	14.8	65.7
ex. on Figure 4.13	code	detect	act.	backwards	deact.	total disc.
temps [s]	0.54	0.08	0.12	0.05		0.25
%		14.4	21.8	9.2		45.4
ex. on Figure 4.14	code	detect	act.	backwards	deact.	total disc.
temps [s]	0.46	0.09	0.07	0.04	0.13	0.33
%		19.4	14.3	9.4	28.6	71.7
ex. 1, $s = 18$	code	detect	act.	backwards	deact.	total disc.
temps [s]	7.15	3.05		0.27	1.84	5.16
%		42.7		3.8	25.8	72.3
ex. 2, $s = 18$	code	detect	act.	backwards	deact.	total disc.
temps [s]	13.00	5.34	5.74			11.08
%		41.1	44.2			85.3

Table 4.3: Computational cost of the algorithm for system with $s = 3, 4, 18$. Legend is as follows: code: total time; detect: time for the detection of events; act.: computation of activation time; backwards: time spent in going backwards in the trajectory for checking purposes; deact.: computation of deactivation time; total disc.: total time for detection and computation of events.

Conclusions

This thesis is concerned by the modeling and the numerical simulation of a system composed by organic aerosol particles surrounded by a gas phase. The aerosol particles have been supposed to be all identical and the chemical species constituting the particles have been assumed to be present in the surrounding gas. This gas-aerosol system is closed and the processes between the gas and the particles are the evaporation and the condensation solely. The time evolution of the composition of the aerosol particles, their size and repartition in liquid phases, as well as the concentration of the gas at the surface of the particles and far from the particles, have been studied.

The modeling of the gas-aerosol system has led to optimization-constrained differential equations. The differential part comes from the modeling of the mass flux between the aerosol particles and the surrounding gas. The optimization problem is the mathematical transcription of the phase equilibrium problem which determines the phases repartition in each aerosol particle. This minimization problem contains mixed constraints and a nonconvex nonlinear objective function. The objective function is the molar Gibbs free energy of a particle.

The coupling of the differential equations and the minimization problem induces losses of regularity in the primal variables of the optimization problem. These losses occur when an inequality constraint is activated or deactivated. Consequently, in addition to the development of a numerical method that solves the optimization-constrained differential equations, techniques for the detection and computation of the times at which the losses of regularity occurred (discontinuity times) and the solution defined at this discontinuity time (discontinuity points) have been established. Two approaches have been proposed. Both methods follow the same strategy: at each time step t^n of the simulation

1. solve the optimization-constrained differential equations over the time interval $[t^n, t^{n+1}]$ for a fixed number of inactive inequality constraints in the optimization problem;

2. check if an inequality constraint has to be activated or deactivated in $[t^n, t^{n+1}]$ with detection criteria. If a detection criterion is satisfied,

    (a) compute the discontinuity time and points,

(b) adapt the optimization-constrained differential equations with the new number of inactive constraints.

The procedure is executed until the final time of integration or the concentration equilibrium between the gas and the particle is reached.

Both approaches are sharing the same detection criteria. The activation of an inequality constraint is detected by observing the time evolution of the variables $y_\alpha(t)$, $\alpha \in \mathcal{I}(t)$; whereas for the deactivation of a constraint, the observed quantity is the distance between the supporting tangent plane and the points $(\mathbf{x}_\alpha(t), g(\mathbf{x}_\alpha(t)))$, $\alpha \in \mathcal{A}(t)$. A slight difference in the detection criteria exists between the 2 approaches. For the first approach, the detection of an activation has been tested at each inner iteration of the resolution of the optimization problems. For the second approach the satisfaction of this criterion has been checked only at each time step. Let us note that the variables $y_\alpha(t)$, $\alpha \in \mathcal{I}(t)$ have been directly computed by both approaches, whereas the variables $\mathbf{x}_\alpha(t)$, $\alpha \in \mathcal{A}(t)$ have required an additional work. This work has been executed at each time step and consisted in setting $\mathbf{x}_\alpha(t)$, $\alpha \in \mathcal{A}(t)$ as the minimizer of the distance between the supporting tangent plane and the point $(\mathbf{x}_\alpha(t), g(\mathbf{x}_\alpha(t)))$.

The first approach is in the line with the resolution of the optimization problem proposed by Amundson et al. in [4, 5]. In order to keep the efficient primal-dual interior-point method of Amundson et al., the first method consists in a splitting between the differential and the optimization parts of the optimization-constrained differential equations. In the splitting, the ordinary differential equations have been solved with the Crank-Nicolson scheme. In order to decrease the computational cost, a warm-start strategy has been added for the initialization of the successive optimization problems. With this approach the meaning of the variables $\mathbf{x}_\alpha$ and y_α, $\alpha = 1, \dots, p$, has been preserved and the variables y_α, $\alpha = 1, \dots, p$, have remained nonnegative. This fact implies on the one hand that the activation of the constraint $\bar{\alpha}$ was detected if the relation $0 \leq y_\alpha(t) < \epsilon_y$ held, where ϵ_y is a given threshold. On the other hand the continuous nonnegativeness of y_α, $\alpha = 1, \dots, p$, prevents from using interpolation techniques for the computation of the discontinuity times. An extrapolation technique based on the works of Esposito and Kumar [33] has been employed to determine the fractional time step needed to reach the discontinuity.

In the extrapolation technique the time derivative of the variable $y_{\bar{\alpha}}$ is required, if $\bar{\alpha}$ denoted the activating constraint. The variable $y_{\bar{\alpha}}$ has been the result of the optimization problem. Hence the derivative $\frac{\mathrm{d}}{\mathrm{d}t} y_{\bar{\alpha}}$ has been first transformed via the chain rule in order to come out the partial derivatives $\frac{\partial y_{\bar{\alpha}}}{\partial b_i}$ and $\frac{\mathrm{d}}{\mathrm{d}t} b_i$, $i = 1, \dots, s$. Second a sensitivity analysis of the optimization problem defined at the time step just before the activation has allowed to know the exact value of the partial derivatives $\frac{\partial y_{\bar{\alpha}}}{\partial b_i}$, $i = 1, \dots, s$ [36]. The derivatives $\frac{\mathrm{d}}{\mathrm{d}t} b_i$, $i = 1, \dots, s$ have been approached by a linear interpolation. Thus the main work in the extrapolation technique has consisted in the computation of the derivatives $\frac{\partial y_{\bar{\alpha}}}{\partial b_i}$, $i = 1, \dots, s$. Thanks to the sensitivity analysis the computational cost of this work is not high (in terms of CPU time). For the computation of the deactivation time the extrapolation technique with the sensitivity analysis has also been employed. The extrapolated function in that case was the distance between the supporting tangent plane and the minimum of the energy function.

This first approach is not provided with a step size selection and uses second order schemes for the resolution of the ordinary differential equations and the computation of the discontinuity time and points. Nevertheless the method is fast (less than 5 minutes for the simulation of a gas-aerosol system made of 18 chemical components). The warm-start strategy in addition with the backwards check have considerably increased the efficiency of this first method. Difficulties have however been encountered for gas-aerosol system whose phase diagram was not classical, namely when some convex areas $\Delta'_{s,\alpha}$, $\alpha = 1, \ldots, p$, of the phase diagram are similar. In that case some simulations have not succeeded because of one of the following reasons

- the deactivation of the constraint $\bar{\alpha}$ is missed because of the convergence of $\mathbf{x}_{\bar{\alpha}}$ to the inactive vector $\mathbf{x}_\alpha$, $\alpha \in \mathcal{I}$, whose convex area $\Delta'_{s,\alpha}$ is identical to $\Delta'_{s,\bar{\alpha}}$. The vector $\mathbf{x}_{\bar{\alpha}}$ should converge to the global minimum of the distance function.
- The deactivation of the constraint $\bar{\alpha}$ is correctly computed but the simulation does not restart since the primal-dual interior-point method has activated the freshly inactive constraint $\bar{\alpha}$.

In the second approach the physical and chemical senses of the variables have been left asides and the optimization-constrained differential equations is considered as a differential algebraic system, after replacement of the optimization problem by its KKT conditions. The fifth-order implicit Runge-Kutta method, RADAU5, has been employed to solve the differential algebraic system. The detection criteria are similar to the first approach with the slight difference that now y_α may become negative. The computation of the discontinuity time is different. The technique consists to add the fractional time step needed to reach the discontinuity as an unknown in the differential algebraic system and solve the resulting system with a splitting idea in order to use the RADAU5 method again.

This second approach uses a fifth-order scheme to solve the differential algebraic system and computes the discontinuity exactly. In comparison to the first approach, this method is more efficient. The second approach is also faster. Indeed the CPU times of all numerical examples are smaller with this approach (only 13 s for the example with 18 chemical components). However the method has encountered similar difficulties for gas-aerosol system whose phase diagram was not classical by missing the deactivation. This fact is normal since the detection criteria are similar. But unlike for the first approach, the method restarts after the computation of the discontinuity points and time.

In conclusion the second approach is more efficient than the first one. The first approach can be improved by adding a step size selection, but it will be difficult to increase its order. Indeed to increase its order, the order of the method that computes the discontinuity time has to be increased. This requires to approximate the derivatives of y_α of higher order, which is very difficult. The second approach can also be improved by replacing the bisection method in the algorithm that computes the discontinuity time. Indeed after one step of the algorithm, one knows that the discontinuity time is in the neighborhood of the right extremity of the considered time interval. The bisection method forces to move in the middle of the time interval and then converges to the right extremity. An improvement could be to use the bisection method once and then use the Newton method with the solution of the bisection as the initialization.

The Achilles' heel of each approach is the detection of the deactivations on non classical phase diagrams. The deactivations may be missed because the variables $\mathbf{x}_{\bar{\alpha}}$, $\bar{\alpha} \in \mathcal{A}$ do not realize the minimum of the distance between the supporting tangent plane and $(\mathbf{x}_{\bar{\alpha}}, g(\mathbf{x}_{\bar{\alpha}}))$. In the computation of $\mathbf{x}_{\bar{\alpha}}$, $\bar{\alpha} \in \mathcal{A}$, the iterates in the algorithm converge to the inactive vector $\mathbf{x}_{\alpha}$, $\alpha \in \mathcal{I}$ instead of converging to the global minimizer of the distance function. In the numerical examples concerned by this difficulty one has observed that if the initialization of the algorithm is chosen more appropriately, the iterates converge to the global minimizer and the deactivation is detected. But since no information on the topology of the energy function and the phase diagram is a priori known, the initialization is either the last iterate of the previous time step or in a corner of the phase diagram. Consequently we are sure that each vector $\mathbf{x}_{\alpha}$ belongs to its corresponding convex area $\Delta'_{s,\alpha}$ on the phase diagram. An idea to improve the efficiency of the detection criterion is to determine roughly the topology of the phase diagram before the simulation. The approximation of the local minimizers of the energy function defines the new initialization of the vector $\mathbf{x}_{\alpha}$, $\alpha = 1, \ldots, p$. This technique must not be time or memory-consuming. Our first idea is to use global optimization such as particle swarm optimizers [14, 25, 103]. The goal of global optimization is to determine the global minimum of a minimization problem, whereas our goal is to determine the local minima of the energy function g on the phase diagram. Then adaptations in the parameters should be done.

In the gas-aerosol system, all organic aerosol particles are considered identical. The next step in the modeling is to integrate a population of aerosol particles in the system. Caboussat and Leonard in [20, 21] work in this direction. They present a model that follows the first approach by using a fixed-point algorithm. Following the results of this thesis, a resolution with the second approach shall be investigated.

Bibliography

[1] A. G. Allen, R. M. Harrison, and J.-W. Erisman. Field measurements of the dissociation of ammonium nitrate and ammonium chloride aerosols. *Atmospheric Environment (1967)*, 23(7):1591 – 1599, 1989.

[2] N. R. Amundson, A. Caboussat, J. W. He, C. Landry, and J. H. Seinfeld. A dynamic optimization problem related to organic aerosols. *C. R. Acad. Sci.*, 344(8):519–522, 2007.

[3] N. R. Amundson, A. Caboussat, J. W. He, C. Landry, C. Tong, and J. H. Seinfeld. A new atmospheric aerosol phase equilibrium model (UHAERO): organic systems. *Atmos. Chem. Phys.*, 7:4675–4698, 2007.

[4] N. R. Amundson, A. Caboussat, J. W. He, and J. H. Seinfeld. An optimization problem related to the modeling of atmospheric organic aerosols. *C. R. Acad. Sci.*, 340(10):765–768, 2005.

[5] N. R. Amundson, A. Caboussat, J. W. He, and J. H. Seinfeld. Primal-dual interior-point algorithm for chemical equilibrium problems related to modeling of atmospheric organic aerosols. *J. Optim. Theory Appl.*, 130(3):375–407, 2006.

[6] H. Antil, R. H. W. Hoppe, and C. Linsenmann. Path-following primal-dual interior-point methods for shape optimization. *Journal of Numerical Mathematics*, 15(2):81–100, 2007.

[7] U. M. Ascher and L. R. Petzold. *Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations.* Society for Industrial and Applied Mathematics, 1998.

[8] O. Axelsson. A class of *A*-stable methods. *Nordisk Tidskr. Informationsbehandling (BIT)*, 9:185–199, 1969.

[9] H. Y. Benson and D. F. Shanno. An exact primal-dual penalty method approach to warmstarting interior-point methods for linear programming. *Comput. Optim. Appl.*, 38(3):371–399, 2007.

[10] H. Y. Benson and D. F. Shanno. Interior-point methods for nonconvex nonlinear programming: regularization and warmstarts. *Comput. Optim. Appl.*, 40(2):143–189, 2008.

[11] F. M. Bowman, J. R. Odum, J. H. Seinfeld, and S. N. Pandis. Mathematical model for gas-particle partitioning of secondary organic aerosols. *Atmospheric Environment*, 31(23):3921–3931, 1997.

[12] S. Boyd and L. Vandenberghe. *Convex optimization.* Cambridge University Press, Cambridge, 2004.

[13] M. Brauer, C. Avila-Casado, T. I. Fortoul, S. Vedal, B. Stevens, and A. Churg. Air pollution and retained particles in the lung. *Environmental Health Perspectives*, 109(10):1039–1043, 2001.

[14] R. Brits, A. P. Engelbrecht, and F. van den Bergh. Locating multiple optima using particle swarm optimization. *Appl. Math. Comput.*, 189(2):1859–1883, 2007.

[15] J. C. Butcher. Implicit Runge-Kutta processes. *Math. Comp.*, 18:50–64, 1964.

[16] J. C. Butcher. *The numerical analysis of ordinary differential equations: Runge-Kutta and general linear methods.* Wiley, Chichester, 1987.

[17] D. W. Byun and J. K. S. Ching. Science algorithms of the EPA models-3 community multiscale air quality (CMAQ) modeling system. Technical Report EPA/600/R-99/030, USA EPA, 1999.

[18] A. Caboussat and C. Landry. Dynamic optimization and event location in atmospheric chemistry. *Proc. Appl. Math. Mech. (PAMM)*, 7(1):2020035–2020036, 2007.

[19] A. Caboussat and C. Landry. A second order scheme for solving optimization-constrained differential equations with discontinuities. In *Numerical Mathematics and Advanced Applications*, pages 761–768. Springer Verlag, Berlin, 2008. Proceedings of Enumath 2007.

[20] A. Caboussat and A. Leonard. Numerical method for a dynamic optimization problem arising in the modeling of a population of aerosol particles. *C. R. Math. Acad. Sci. Paris*, 346(11-12):677–680, 2008.

[21] A. Caboussat and A. Leonard. Numerical solution and fast-slow decomposition of a population of weakly coupled systems. *DCDS Supplements*, 2009. (submitted).

[22] G. Caloz and J. Rappaz. *Numerical Analysis for Nonlinear and Bifurcation Problems*, volume V of *Handbook of Numerical Analysis (P.G. Ciarlet, J.L. Lions eds)*, pages 487–637. Elsevier, Amsterdam, 1997.

[23] F. E. Cellier. *Combined Continuous/Discrete System Simulation by Use of Digital Computers*. PhD thesis, Swiss Federal Institute of Technology, ETH, Zurich, 1979.

[24] S. L. Clegg, J. H. Seinfeld, and P. Brimblecombe. Thermodynamic modelling of aqueous aerosols containing electrolytes and dissolved organic compounds. *Journal of Aerosol Science*, 32(6):713–738, 2001.

[25] M. Clerc. *Particle swarm optimization*. ISTE, London, 2006.

[26] D. S. Covert and J. Heintzenberg. Measurement of the degree of internal/external mixing of hygroscopic compounds and soot in atmospheric aerosols. *Science of The Total Environment*, 36:347 – 352, 1984.

[27] B. Dahneke. Simple kinetic theory of Brownian diffusion in vapors and aerosols. In R. E. Meyer, editor, *Theory of Dispersed Multiphase Flow*, pages 97–133. Academic Press, New York, 1983.

[28] K. Denbigh. *The principles of chemical equilibrium : with applications in chemistry and chemical engineering*. Cambridge University Press, Cambridge, 1997.

[29] P. Deuflhard. *Newton methods for nonlinear problems: affine invariance and adaptive alorithms*. Springer Verlag, Berlin, 2004.

[30] D. B. Van Dongen, M. F. Doherty, and J. R. Haight. Material stability of multicomponent mixtures and the multiplicity of solutions to phase-equilibrium equations. 1. nonreacting mixtures. *Industrial and Engineering Chemistry Fundamentals*, 22(4):472–485, 1983.

[31] B. E. Ehle. On Padé approximations to the exponential function and A-stable methods for the numerical solution of initial value problems. Technical Report CSRR 2010, Dept. AACS, University of Waterloo, Ontario, Canada, 1969.

[32] G. B. Erdakos and J. F. Pankow. Gas/particle partitioning of neutral and ionizing compounds to single- and multi-phase aerosol particles. 2. phase separation in liquid particulate matter containing both polar and low-polarity organic compounds. *Atmospheric Environment*, 38(7):1005–1013, 2004.

[33] J. M. Esposito and V. Kumar. A state event detection algorithm for numerically simulating hybrid systems with model singularities. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 17:1–22, 2007.

[34] F. Facchinei, A. Fischer, and C. Kanzow. On the accurate identification of active constraints. *SIAM J. Optim.*, 9(1):14–32, 1999.

[35] B. Faugeras, J. Pousin, and F. Fontvieille. An efficient numerical scheme for precise time integration of a diffusion-dissolution/precipitation chemical system. *Math. Comp.*, 75(253):209–222, 2005.

[36] A. V. Fiacco and G. P. McCormick. *Nonlinear programming : sequential unconstrained minimization techniques.* Wiley, New York, 1968.

[37] C. A. Floudas. *Deterministic global optimization*, volume 37 of *Nonconvex Optimization and its Applications.* Kluwer Academic Publishers, Dordrecht, 2000. Theory, methods and applications.

[38] A. Forsgren, P. E. Gill, and M. H. Wright. Interior Methods for Nonlinear Optimization. *SIAM Review*, 44(4):525–597, 2002.

[39] A. Fredenslund, J. Gmehling, and P. Rasmussen. *Vapor-Liquid Equilibrium Using UNIFAC.* Elsevier, Amsterdam, 1977.

[40] N. A. Fuchs. *The Mechanics of Aerosols.* Pergamon Press, 1964.

[41] N. A. Fuchs and A. G. Sutugin. High dispersed aerosols. In G. M. Hidy and J. R. Brock, editors, *Topics in Current Aerosol Research (Part 2)*, pages 1–200. Pergamon Press, 1971.

[42] C. W. Gear and O. Østerby. Solving ordinary differential equations with discontinuities. *ACM Trans. Math. Software*, 10(1):23–44, 1984.

[43] F. Gelbard and J. H. Seinfeld. The general dynamic equation for aerosols. theory and application to aerosol formation and growth. *Journal of Colloid and Interface Science*, 68(2):363–382, 1979.

[44] J. Gondzio. Warm start of the primal-dual method applied in the cutting-plane scheme. *Math. Programming*, 83(1, Ser. A):125–143, 1998.

[45] J. Gondzio and A. Grothey. A new unblocking technique to warmstart interior point methods based on sensitivity analysis. *SIAM Journal on Optimization*, 19(3):1184–1210, 2008.

[46] N. I. M. Gould. On practical conditions for the existence and uniqueness of solutions to the general equality quadratic programming problem. *Mathematical Programming*, 32:90–99, 1985.

[47] N. Guglielmi and E. Hairer. Computing breaking points in implicit delay differential equations. *Advances in Computational Mathematics*, 29(3):229–247, 2008.

[48] E. Hairer, S. P. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations I: Nonstiff Problems. 2nd Edition*, volume 8 of *Springer Series in Computational Mathematics.* Springer-Verlag, 1993.

[49] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II. Stiff and Differential-Algebraic Problems. 2nd Edition*, volume 14 of *Springer Series in Computational Mathematics.* Springer-Verlag, 1996.

[50] E. Hairer and G. Wanner. Stiff differential equations solved by Radau methods. *J. Comput. Appl. Math.*, 111:93–111, 1999.

[51] H. K. Hansen, P. Rasmussen, A. Fredenslund, M. Schiller, and J. Gmehling. Vapor-liquid equilibria by UNIFAC group contribution. 5. revision and extension. *Industrial and Engineering Chemistry Process Design and Development*, 30(10):2352–2355, 1991.

[52] G. A. Iglesias-Silva, A. Bonilla-Petriciolet, P. T. Eubank, J. C. Holste, and K. R. Hall. An algebraic method that includes Gibbs minimization for performing phase equilibrium calculations for any number of components or phases. *Fluid Phase Equilibria*, 210:229–245, 2003.

[53] M. Z. Jacobson. *Fundamentals of atmospheric modeling.* Cambridge University Press, 2005.

[54] M. Z. Jacobson, A. Tabazadeh, and R. P. Turco. Simulating equilibrium within aerosols and nonequilibrium between gases and aerosols. *Journal of Geophysical Research*, 101(D4):9079–9091, 1996.

[55] Y. Jiang, G. R. Chapman, and W. R. Smith. On the geometry of chemical reaction and phase equilibria. *Fluid Phase Equilibria*, 118(1):77–102, 1996.

[56] Y. Jiang, W. R. Smith, and G. R. Chapman. Global optimality conditions and their geometric interpretation for the chemical and phase equilibrium problem. *SIAM Journal on Optimization*, 5(4):813–834, 1995.

[57] E. John and E. A. Yıldırım. Implementation of warm-start strategies in interior-point methods for linear programming in fixed dimension. *Computational Optimization and Applications*, 41:151–183, 2008.

[58] V. R. Kelly, G. M. Lovett, K. C. Weathers, and G. E. Likens. Trends in atmospheric concentration and deposition compared to regional and local pollutant emissions at a rural site in southeastern New York, USA. *Atmos. Environ.*, 36(10):1569–1575, 2002.

[59] S. Kinne, U. Lohmann, J. Feichter, M. Schulz, C. Timmreck, S. Ghan, R. Easter, M. Chin, P. Ginoux, T. Takemura, I. Tegen, D. Koch, M. Herzog, J. Penner, G. Pitari, B. Holben, T. Eck, A. Smirnov, O. Dubovik, I. Slutsker, D. Tanre, O. Torres, M. Mishchenko, I. Geogdzhayev, D. A. Chu, and Y. Kaufman. Monthly averages of aerosol properties: A global comparison among models, satellite data, and AERONET ground data. *J. Geophys. Res.*, 108(D20, 4634), 2003.

[60] B. Koo, T. M. Gaydos, and S. N. Pandis. Evaluation of the equilibrium, dynamic, and hybrid aerosol modeling approaches. *Aerosol Science and Technology*, 37(1):53 – 64, 2003.

[61] P. Kunkel and V. Mehrmann. *Differential-Algebraic Equations, Analysis and Numerical Solution.* EMS Textbooks in Mathematics. European Mathematical Society, 2006.

[62] C. Landry, A. Caboussat, and E. Hairer. Solving optimization-constrained differential equations with discontinuity points. *SIAM Journal on Scientific Computing*, 2009. (submitted).

[63] Y.L. Lee and R. Sequeira. Water-soluble aerosol and visibility degradation in Hong Kong during autumn and early winter, 1998. *Environmental Pollution*, 116(2):225–233, 2002.

[64] T. Magnussen, P. Rasmussen, and A. Fredenslund. UNIFAC parameter table for prediction of liquid-liquid equilibria. *Industrial and Engineering Chemistry Process Design and Development*, 20:331–339, 1981.

[65] W. C. Malm, J. F. Sisler, D. Huffman, R. A. Eldred, and T. A. Cahill. Spatial and seasonal trends in particle concentration and optical extinction in the United States. *J. Geophys. Res.*, 99:1347–1370, 1994.

[66] G. Mao and L. R. Petzold. Efficient integration over discontinuities for differential-algebraic systems. *Comput. Math. Appl.*, 43(1-2):65–79, 2002.

[67] C. Marcolli, B. P. Luo, Th. Peter, and F. G. Wienhold. Internal mixing of the organic aerosol by gas phase diffusion of semivolatile organic compounds. *Atmospheric Chemistry and Physics*, 4:2593–2599,, 2004.

[68] C. M. McDonald and C. A. Floudas. GLOPEQ: A new computational tool for the phase and chemical equilibrium problem. *Computers and Chemical Engineering*, 21(1):1–23, 1996.

[69] C.M. McDonald and C.A. Floudas. Global optimization and analysis for the Gibbs free energy function for the UNIFAC, Wilson, and ASOG equations. *Industrial and Engineering Chemistry Research*, 34:1674–1687, 1995.

[70] P. H. McMurry and M. R. Stolzenburg. On the sensitivity of particle size to relative humidity for Los Angeles aerosols. *Atmospheric Environment (1967)*, 23(2):497 – 507, 1989.

[71] Z. Meng and J. H. Seinfeld. Time scales to achieve atmospheric gas-aerosol equilibrium for volatile species. *Atmospheric Environment*, 30(16):2889–2900, 1996.

[72] C. Mészáros. On numerical issues of interior point methods. *SIAM J. Matrix Anal. Appl.*, 30(1):223–235, 2008.

[73] A. S. Nemirovski and M. J. Todd. Interior-point methods for optimization. *Acta Numer.*, 17:191–234, 2008.

[74] A. Nenes, S. N. Pandis, and C Pilinis. ISORROPIA: A new thermodynamic equilibrium model for multiphase multicomponent inorganic aerosols. *Aquatic Geochemistry*, 4:123–152, 1998.

[75] J. Nocedal and S. J. Wright. *Numerical optimization.* Springer Series in Operations Research and Financial Engineering. Springer, New York, second edition, 2006.

[76] W. B. Norris, S. F. Mueller, and J. E. Langstaff. Estimates of sulfate deposition in the middle eastern United States: 1975, 1990, and 2010. *J. Air Waste Manage. Assoc.*, 49(6):655–668, 1999.

[77] G. Oberdorster. Pulmonary, effects of inhaled ultrafine particles. *International Archives of Occupational and Environmental Health*, 74(1):1–8, 2001.

[78] Intergovernmental Panel on Climate Change. *Climate change 2001: the scientific basis.* Cambridge Univ. Press, New York, 2001.

[79] J. F. Pankow. Gas/particle partitioning of neutral and ionizing compounds to single and multi-phase aerosol particles. 1. unified modeling framework. *Atmospheric Environment*, 37(24):3323–3333, 2003.

[80] R. J. Park, D. J. Jacob, B. D. Field, R. M. Yantosca, and M. Chin. Natural and transboundary pollution influences on sulfate-nitrate-ammonium aerosols in the United States: Implications for policy. *Journal of Geophysical Research*, 109:D15204, 2004.

[81] T. Park and P. I. Barton. State event location in differential-algebraic models. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 6:137–165, 1996.

[82] L. R. Petzold. Order results for implicit Runge-Kutta methods applied to differential/algebraic systems. *SIAM J. Numer. Anal.*, 23(4):837–852, 1986.

[83] C. Pilinis, K. P. Capaldo, A. Nenes, and S. N. Pandis. MADM - a new multicomponent aerosol dynamics model. *Aerosol Science and Technology*, 2000.

[84] C. Pilinis and J. H. Seinfeld. Development and evaluation of an Eulerian photochemical gas-aerosol model. *Atmospheric Environment*, 22:1985–2001, 1988.

[85] C. Pilinis, J. H. Seinfeld, and C. Seigneur. Mathematical modeling of the dynamics of multicomponent atmospheric aerosols. *Atmospheric Environment*, 21(4):943–955, 1987.

[86] C. Pilinis and J.H. Seinfeld. Continued development of a general equilibrium model for inorganic multicomponent atmospheric aerosols. *Atmospheric Environment*, 21:2453–2466, 1987.

[87] A. Prothero and A. Robinson. On the stability and accuracy of one-step methods for solving stiff systems of ordinary differential equations. *Math. Comp.*, 28:145–162, 1974.

[88] H. R. Pruppacher and J. D. Klett. *Microphysics of Clouds and Precipitation*, volume 18 of *Atmospheric and Oceanographic Sciences Library.* Kluwer Academic Publishers, 2nd rev edition edition, 1996.

[89] P. J. Rabier and A. Griewank. Generic aspects of convexification with applications to thermodynamic equilibrium. *Archive for Rational Mechanics and Analysis*, 118(4):349–397, 1992.

[90] V. Ramanathan, M. V. Ramana, G. Roberts, D. Kim, C. Corrigan, C. Chung, and D.Winker. Warming trends in Asia amplified by brown cloud solar absorption. *Nature*, 448(7153):575–578, 2007.

[91] J. Rappaz. Numerical approximation of PDEs and Clément's interpolation. In *Partial Differential Equations and Functional Analysis, The Philippe Clément Festschrift*, volume 168 of *Operator Theory: Advances and Applications*, pages 237–250. Birkhäuser Verlag, 2006.

[92] R. A. Reck. The role of aerosols in the climate system: Results of numerical experiments in climate models. *Advances in Space Research*, 2(5):11 – 18, 1982.

[93] R. T. Rockafellar. *Convex analysis.* Princeton Landmarks in Mathematics. Princeton University Press, Princeton, NJ, 1997. Reprint of the 1970 original,Princeton Paperbacks.

[94] K. N. Sartelet, H. Hayami, B. Albriet, and B. Sportisse. Development and preliminary validation of a Modal Aerosol Model for tropospheric chemistry: MAM. *Aerosol Science and Technology*, 40(2):118–127, 2006.

[95] P. Saxena, C. Seigneur, and T. W. Peterson. Modeling of multiphase atmospheric aerosols. *Atmospheric Environment (1967)*, 17(7):1315 – 1329, 1983.

[96] J. H. Seinfeld and S. N. Pandis. *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change.* Wiley, New York, 1998.

[97] L. F. Shampine, I. Gladwell, and R. W. Brankin. Reliable solution of special event location problems for ODEs. *ACM Trans. Math. Software*, 17:11–25, 1991.

[98] B. Sportisse. A review of current issues in air pollution modeling and simulation. *Computational Geosciences*, 11:159–181, 2007.

[99] B. Sportisse. *Pollution atmosphérique : des processus à la modélisation.* Springer, Paris, 2008.

[100] S. P. Tan and M. Radosz. Constructing binary and ternary phase diagrams on the basis of phase stability analysis. *Industrial and Engineering Chemistry Research*, 41(15):3722 –3730, 2002.

[101] K. Tone. An active-set strategy in an interior point method for linear programming. *Math. Programming*, 59(3, Ser. A):345–360, 1993.

[102] D. A. Vallero. *Fundamentals of air pollution*. Elsevier, Amsterdam, 4th ed. edition, 2008.

[103] F. van den Bergh and A. P. Engelbrecht. A study of particle swarm optimization particle trajectories. *Inform. Sci.*, 176(8):937–971, 2006.

[104] V. Veliov. On the time-discretization of control systems. *SIAM J. Control Optim.*, 35(5):1470–1486, 1997.

[105] S. K. Wasylkiewicz, L. N. Sridhar, M. F. Doherty, and Michael F. Malone. Global stability analysis and calculation of liquid-liquid equilibrium in multicomponent mixtures. *Industrial and Engineering Chemistry Research*, 35:1395–1408, 1996.

[106] A. S. Wexler, F. W. Lurmann, and J. H. Seinfeld. Modelling urban and regional aerosols-i. model development. *Atmospheric Environment*, 28:531–546, 1994.

[107] A. S. Wexler and J. H. Seinfeld. The distribution of ammonium salts among a size and composition dispersed aerosol. *Atmospheric Environment*, 24:1231–1246, 1990.

[108] E. A. Yıldırım and S. J. Wright. Warm-start strategies in interior-point methods for linear programming. *SIAM J. Optim.*, 12(3):782–810 (electronic), 2002.

[109] C. Zălinescu. *Convex analysis in general vector spaces*. World Scientific Publishing Co. Inc., River Edge, NJ, 2002.

[110] R. A. Zaveri, R. C. Easter, J. D. Fast, and L. K. Peters. Model for simulating aerosol interactions and chemistry (MOSAIC). *J. Geophys. Res.*, 113, 2008. D13204.

[111] K. M. Zhang and A. S. Wexler. Modeling the number distributions of urban and regional aerosols: theoretical foundations. *Atmospheric Environment*, 36(11):1863–1874, 2002.

[112] Y. Zhang, C. Seigneur, J. H. Seinfeld, M. Z. Jacobson, and F. S. Binkowski. Simulation of aerosol dynamics: A comparative review of algorithms used in air quality models. *Aerosol Science and Technology*, 1999.

Curriculum Vitae

I was born on December 27th, 1980 in Saint-Maurice, Switzerland. I have done my primary school in Vernayaz, my secondary school in Saint-Maurice and finished the high school at the Collège de l'Abbaye de Saint-Maurice in 2000 where I obtained the *maturité de type C* (scientific option). I was then admitted in the mathematics section at the Swiss Federal Institute of Technology (EPFL) in Lausanne. In March 2005 I received a master of mathematical sciences after having completed my master thesis under the supervision of Prof. Jacques Rappaz and Dr. Michel Flueck. From April to October 2005 I have worked on the air quality project, called *UHAERO*, with the Profs Jiwen He and Alexandre Caboussat at the University of Houston, USA. Since October 2005 I have been pursuing my research on the UHAERO project at EPFL and working as an assistant in the Chair of Numerical Analysis and Simulation for Prof. Jacques Rappaz. My research theme is ordinary differential equations and constrained optimization problems.