DOKTORANDSKÉ DNY 2007

sborník workshopu doktorandů FJFI
oboru Matematické inženýrství

16. a 23. listopadu 2007

P. Ambrož, Z. Masáková (editoři)

ISBN 978-80-01-03913-7
Tisk: Česká technika — nakladatelství ČVUT

Katedra matematiky
Fakulta jaderná a fyzikálně inženýrská
České vysoké učení technické v Praze
Trojanova 13
120 00 Praha 2

Seznam příspěvků

Předmluva

Ve dnech 16. a 23. listopadu 2007 se na katedře matematiky FJFI konají Doktorandské dny určené pro studenty doktorského studijního programu Aplikace přírodních věd oboru Matematické inženýrství. Tento obor je zajišťován katedrami matematiky a fyziky spolu s některými ústavy Akademie věd ČR a pokrývá témata od matematických modelů přírodních procesů přes otázky kvantové teorie až po databázové systémy či neuronové sítě.

Letošní Doktorandské dny jsou již druhé v řadě. V rámci tohoto setkání představí doktorandi svoji práci ostatním studentům, školitelům i všem zájemcům z řad odborné veřejnosti. Texty příspěvků jsou předkládány v tomto sborníku.

Workshop Doktorandské dny si klade za cíl umožnit doktorandům konfrontovat své výsledky na širším fóru, poskytnout prostor pro oponentní připomínky k studentově práci ze strany školitelů a přítomných odborníků, a tím tak přispět ke zkvalitnění výchovy doktorandů oboru Matematické inženýrství. Uchování příspěvků ve sborníku pak umožní sledovat postup práce jednotlivých studentů na jejich vědeckém úkolu.

Editoři

Factor Frequencies of Reversal Closed Languages

Ľubomíra Balková

3rd year of PGS, email: l.balkova@centrum.cz
Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU
advisor: Zuzana Masáková, Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU

Abstract. We study infinite words over a finite alphabet. In particular, we focus on frequencies of factors (subwords) of infinite words whose language is reversal closed, i.e. u contains with each factor also its mirror image. Crucial is the notion of Rauzy graphs associated with the infinite word. Investigation of symmetries of the reduced Rauzy graph Γ_n, $n \in \mathbb{N}$, allows us to determine a good and easily calculable upper bound on the number of different factor frequencies.

Abstrakt. Studujeme nekonečná slova nad konečnou abecedou. Speciálně se zaměřujeme na frekvence faktoru nekonečných slov, jejichž jazyk je uzavřen na reverzi, tj. s každým slovem obsahuje také slovo, které získáme, když přečteme dané slovo pozpátku. Klíčovým pojmem je Rauzyho graf přiřazený nekonečnému slovu. Zkoumání symetrií redukovaného Rauzyho grafu Γ_n, $n \in \mathbb{N}$, nám umožní nalézt dobrý a snadno vypočitatelný horní odhad na počet ruzných frekvencí faktoru nekonečného slova.

1 Introduction

Everybody who is about to study a foreign language is interested in word frequencies of this language. The reason is simple. If you start, there is no point in beginning with low-frequency words provided your aim is to manage everyday communication. Word frequencies are in focus of designers of internet search engines, but also of the one who wants to raise the visit rate of his web page. There exist so-called "stoplists" which provide frequencies of most often used words. For instance, just three words *I, and, the* account for ten percent of all words in printed English. This is "easy" to calculate. Prepare a sheet of paper, go through all printed matters in English, for each word you read, put a black tally on the sheet, and each time you see *I* or *and* or *the*, put a red tally on the sheet. At the end, divide the number of red tallies by the number of black tallies and you should obtain approximately $0,1$. In the Czech language, similar role is played by words *a, v, se, na, je, že, o* which take about 9 percent of a written text. In this paper, our point of view will not be linguistic (statistic), we will instead move to the domain of Combinatorics on Words and Graph Theory. We will turn our attention to factor frequencies in infinite words, so the number of occurrences of a factor will be possibly infinite and the definition of factor frequency will have to be generalized. We will show how to find a good upper bound on the number of different factor frequencies in infinite words which contain with every factor also its mirror image. Let us also mention that we have studied factor frequencies in several classes of infinite words (to be found in the thesis) and the results confirm accuracy of the obtained upper bound.

Having introduced notation and basic definitions, we will first recall well-known relations holding for frequencies of edges and vertices in Rauzy graphs (Kirchhoff's law). Afterwards, we will introduce a useful tool- *reduced Rauzy graph*. With this in hand, one can easily deduce the upper bound derived by Boshernitzan (Theorem 7). Knowing that for any infinite reversal closed word u, the mirror map does not change factor frequencies will allow us to improve essentially the upper bound in case of words whose language is reversal closed (Theorem 10).

2 Preliminaries

First, let us recall our "vocabulary" which will be used throughout this paper. An *alphabet* $\mathcal{A}$ is a finite set of symbols called *letters*. A concatenation of letters is a *word*. *Length* of a word w is the number of letters contained in w and is denoted $|w|$. We will also deal with right-sided infinite words $u = u_0u_1u_2....$ A finite word w is called a *factor* of the word u (finite or infinite) if there exist a finite word $w^{(1)}$ and a word $w^{(2)}$ (finite or infinite) such that $u = w^{(1)}ww^{(2)}$. An infinite word u is said to be *recurrent* if each of its factors occur infinitely many times in u and u is *uniformly recurrent* if for any $n \in \mathbb{N}$ there exists an $R(n) \in \mathbb{N}$ such that any factor of u of length $R(n)$ contains all factors of length n. An infinite word u is said to be *eventually periodic* if there exist finite words v, w such that $u = vw^\omega$, where w^ω means that w is repeated infinitely many times. A word which is not eventually periodic is called *aperiodic*. *Language* $\mathcal{L}(u)$ of an infinite word u is the set of all factors of u. A language $\mathcal{L}(u)$ is reversal closed, if for every factor $w = w_0w_1 \dots w_n$, where $w_i \in \mathcal{A},\ i \in \{0, \dots, n\}$, also its mirror image $\overline{w} = w_n \dots w_1w_0$ belongs to $\mathcal{L}(u)$. We denote by $\mathcal{L}_n(u)$ the set of factors of length n of the infinite word u. Then, we can define *complexity function* (or *complexity*) $C_u : \mathbb{N} \to \mathbb{N}$ which associates to every n the number of different factors of length n of the infinite word u, i.e.

$$C_u(n) = \#\mathcal{L}_n(u).$$

Let us mention that if there exists $n \in \mathbb{N}$ such that $C_u(n) \leq n$, then the infinite word u is eventually periodic. In other words, aperiodic words has complexity $C(n) \geq n + 1$ for all $n \in \mathbb{N}$. Aperiodic words with the lowest possible complexity are called *Sturmian*. Similarly, let us denote by $\mathcal{P}al_n(u)$ the set of palindromes of length n contained in u and let us define *palindromic complexity* $P_u : \mathbb{N} \to \mathbb{N}$ which associates to every n the number of different palindromes of length n of the infinite word u. We recall that *palindrome* is a word which is equal to its mirror image. We say that $a \in \mathcal{A}$ is *right extension* of a factor $w \in \mathcal{L}(u)$ if wa is also a factor of u. We denote by $Rext(w)$ the set of all right extensions of w in u, i.e. $Rext(w) = \{a \in \mathcal{A} \mid wa \in \mathcal{L}(u)\}$. If $\#Rext(w) \geq 2$, then the factor w is called *right special* (RS for short). Analogously, we define *left extensions, $Lext(w)$, left special factor* (LS for short). Moreover, we say that a factor w is *bispecial* (BS for short) if w is LS and RS. With this in hand, we can give a formula for the *first difference of complexity* $\Delta C_u(n) = C_u(n+1) - C_u(n)$. We leave the proof as an easy exercise.

$$\Delta C_u(n) = \sum_{w \in \mathcal{L}_n(u)} \big(\#Rext(w) - 1\big) = \sum_{w \in \mathcal{L}_n(u)} \big(\#Lext(w) - 1\big), \quad n \in \mathbb{N}. \tag{1}$$

To have everything prepared for the deduction of an improved upper bound on the number of different frequencies, it remains to define *Rauzy graph*, and, of course, *factor frequency*.

Definition 1. Rauzy graph Γ_n of an infinite word u (of order n) is a directed graph whose set of vertices is $\mathcal{L}_n(u)$ and set of edges is $\mathcal{L}_{n+1}(u)$. Let $w_0, w_1, \ldots, w_n$ be letters in $\mathcal{A}$ and let $e = w_0 w_1 \ldots w_{n-1} w_n$ be an edge of Γ_n, then e starts in the vertex $w = w_0 w_1 \ldots w_{n-1}$ and ends in the vertex $v = w_1 \ldots w_{n-1} w_n$.

Definition 2. Let w be a factor of an infinite word u over a finite alphabet $\mathcal{A}$, then (factor) frequency of w (in u) is defined as

$$\rho(w) = \lim_{|v|\to\infty, v\in\mathcal{L}(u)} \frac{\#\{\text{occurrences of } w \text{ in } v\}}{|v|}$$

if the limit exists.

3 Upper bound on the number of factor frequencies

In the sequel, let us suppose that **frequencies of all factors of $\mathcal{L}(u)$ exist**. It is not difficult to see that the frequency of a vertex w in Γ_n is equal to the sum of frequencies of the edges starting in w, or, by symmetry, the sum of frequencies of the edges ending in w. Let us formalize this observation and leave its proof as a simple exercise.

Lemma 3 (Kirchhoff's law)**.** *Let w be a factor of u, then*

$$\rho(w) = \sum_{a\in Lext(w)} \rho(aw) = \sum_{a\in Rext(w)} \rho(wa).$$

Consequently, if a factor $w \in \mathcal{L}(u)$ is neither LS nor RS, then both the frequency of the unique edge starting in w and the frequency of the unique edge ending in w is equal to $\rho(w)$. Formally rewritten, this observation has the following reading.

Corollary 4. *Let w be a factor of u which is neither LS nor RS. Let us denote by a the only left extension of w and by b its only right extension. Then, $\rho(w) = \rho(aw) = \rho(wb)$.*

We can label every edge e in the Rauzy graph Γ_n of u by $\rho(e)$. Then the number of different frequencies of factors in $\mathcal{L}_{n+1}(u)$ corresponds to the number of different edge labels in Γ_n. For a factor $w \in \mathcal{L}_n(u)$ which is neither LS nor RS, it is thus evident that the unique edge ending in w has the same label $\rho(w)$ as the unique edge starting in w. Consequently, if we are interested in the number of different edge labels, we can remove the vertex w from the graph and replace the incoming and outgoing edge with a new edge keeping the label $\rho(w)$. Repeating this procedure, we obtain the so-called reduced Rauzy graph, which has obviously the same set of edge labels. Let us give precise definitions.

Definition 5. Let Γ_n be the Rauzy graph of order n of an infinite word u. A directed path $w^{(0)} w^{(1)} \ldots w^{(m)}$ in Γ_n such that its initial vertex $w^{(0)}$ is LS or RS, its final vertex $w^{(m)}$ is also LS or RS, and the other vertices are neither LS nor RS factors is called simple. We define label of the simple path as the label of any edge of this path.

Definition 6. Reduced Rauzy graph $\tilde{\Gamma}_n$ of u (of order n) is a directed graph whose set of vertices is formed by LS and RS factors of $\mathcal{L}_n(u)$ and whose set of edges is given in the following way. Vertices w and v are connected with an edge e if there exists in Γ_n a simple path starting in w and ending in v. We assign to such an edge e the label of the corresponding simple path.

The number of different edge labels in the reduced Rauzy graph $\tilde{\Gamma}_n$ is clearly less or equal to the number of edges in $\tilde{\Gamma}_n$. Let us thus calculate the number of edges in $\tilde{\Gamma}_n$ in order to get an upper bound on the number of frequencies of factors in $\mathcal{L}_{n+1}(u)$. For every RS factor $w \in \mathcal{L}_n(u)$, it holds that $\#Rext(w)$ edges begin in w, and for every LS factor $v \in \mathcal{L}_n(u)$ which is not RS, only one edge begins in v, thus we get the following relation

$$\#\{e|\ e \text{ edge in } \tilde{\Gamma}_n\} = \sum_{w \text{ RS}} \#Rext(w) + \sum_{v \text{ LS not RS}} 1. \tag{2}$$

Using Equation 1, we deduce that

$$\#\{e|\ e \text{ edge in } \tilde{\Gamma}_n\} = \Delta C(n) + \sum_{v \text{ RS}} 1 + \sum_{v \text{ LS not RS}} 1. \tag{3}$$

The following result initially proved by Boshernitzan in [3] follows immediately.

Theorem 7 (Boshernitzan)**.** *Let u be an infinite word such that for every factor $w \in \mathcal{L}(u)$, the frequency $\rho(w)$ exists. Then for every $n \in \mathbb{N}$, it holds*

$$\#\{\rho(e) \mid e \in \mathcal{L}_{n+1}(u)\} \leq 3\Delta C(n).$$

This upper bound can be lowered for an infinite word u whose language $\mathcal{L}(u)$ is reversal closed. In this case, each factor of u has the same frequency as its mirror image.

Lemma 8. *Let u be an infinite word whose language $\mathcal{L}(u)$ is reversal closed and such that for each factor $w \in \mathcal{L}(u)$, the frequency $\rho(w)$ exists. Then $\rho(w) = \rho(\overline{w})$ holds for each factor w of $\mathcal{L}(u)$.*

Proof. Take an arbitrary factor $w \in \mathcal{L}(u)$ and let $(v^{(n)})_{n=1}^{\infty}$ be any sequence of a strictly growing length in $\mathcal{L}(u)$. Since the frequency of w exists, we can write

$$\rho(w) = \lim_{n\to\infty} \frac{\#\{\text{occurrences of } w \text{ in } v^{(n)}\}}{|v^{(n)}|}.$$

As $\mathcal{L}(u)$ is reversal closed, we get

$$\#\{\text{occurrences of } w \text{ in } v^{(n)}\} = \#\{\text{occurrences of } \overline{w} \text{ in } \overline{v^{(n)}}\}.$$

Using $|v^{(n)}| = |\overline{v^{(n)}}|$, we can then rewrite $\rho(w)$ as follows

$$\rho(w) = \lim_{n\to\infty} \frac{\#\{\text{occurrences of } \overline{w} \text{ in } \overline{v^{(n)}}\}}{|\overline{v^{(n)}}|} = \rho(\overline{w}).$$

The last equality holds thanks to the assumption that frequencies of all factors exist. □

We have now everything prepared for an improvement of the upper bound on the number of edge labels in $\tilde{\Gamma}_n$, or, equivalently, on the number of different factor frequencies of $\mathcal{L}_{n+1}(u)$ of an infinite word u whose language is reversal closed. The following lemma will play an essential role in this improvement.

Lemma 9. *Let u be an infinite word whose language $\mathcal{L}(u)$ is reversal closed and such that for each factor $w \in \mathcal{L}(u)$, the frequency $\rho(w)$ exists. Then for every $n \in \mathbb{N}$, we have*

$$\#\{\rho(e)|e \in \mathcal{L}_{n+1}\} \leq \frac{1}{2}\left(P(n) \ + \ P(n+1) \ + \ \Delta C(n) \ - \sum_{w \text{ BS in } \mathcal{L}_n} 1 \ - \sum_{w \text{ BS in } \mathcal{P}al_n} 1\right) + \sum_{w \text{ RS in } \mathcal{L}_n} 1.$$

Proof. Let Γ_n be the Rauzy graph of u of order n. Let us define a mapping f which to every vertex $w \in \mathcal{L}_n(u)$ associates the vertex $\overline{w}$, to every edge $e \in \mathcal{L}_{n+1}(u)$ associates the edge $\overline{e}$, and to every path $w^{(0)}w^{(1)}\dots w^{(m)}$ in Γ_n associates the path $\overline{w^{(m)}}\dots\overline{w^{(1)}}\ \overline{w^{(0)}}$. Then, clearly, $f^2 = Id$ and thanks to the closeness of $\mathcal{L}(u)$ under reversal, f maps the Rauzy graph Γ_n onto itself, in fact, f is an automorphism of Γ_n. Let us replace the Rauzy graph Γ_n by the reduced Rauzy graph $\tilde{\Gamma}_n$. We know already that the set of edge labels of $\tilde{\Gamma}_n$ is equal to the set of edge labels of Γ_n. Let us denote by A the number of edges e in $\tilde{\Gamma}_n$ such that e is mapped by f onto itself and by B the number of edges e in $\tilde{\Gamma}_n$ such that e is not mapped by f onto itself, then clearly, $\#\{e|\ e \text{ edge in } \tilde{\Gamma}_n\} = A + B$. To be more precise, if e is an edge in $\tilde{\Gamma}_n$ corresponding to the simple path $w^{(0)}w^{(1)}\dots w^{(m)}$ in Γ_n, then $f(e)$ is the edge in $\tilde{\Gamma}_n$ corresponding to the simple path $f(w^{(0)}w^{(1)}\dots w^{(m)}) = \overline{w^{(m)}}\dots\overline{w^{(1)}}\ \overline{w^{(0)}}$. Consequently, if e is mapped by f onto itself, then the corresponding simple path $w^{(0)}w^{(1)}\dots w^{(m)}$ satisfies that its central vertex $w^{(\frac{m}{2})}$ is a palindrome (for m even) or its central edge going from $w^{(\frac{m-1}{2})}$ to $w^{(\frac{m+1}{2})}$ is a palindrome (for m odd). On the other hand, every palindrome of length $n+1$ forms the central edge of a simple path in Γ_n which is mapped by f onto itself and every palindrome of length n forms either the central vertex of a simple path which is mapped by f on itself or is BS and thus a vertex in Γ_n. Therefore,

$$A = P(n) + P(n+1) - \#\{w \in \mathcal{L}_n | w \text{ BS in } \mathcal{P}al_n\}. \tag{4}$$

We subtract the number of palindromic BS factors of $\mathcal{L}_n(u)$ since they form vertices, not edges in $\tilde{\Gamma}_n$. Now, let us turn our attention to edges e which are not mapped by f onto themselves. If e is an edge in $\tilde{\Gamma}_n$ going from a vertex w to v, where $f(e) \neq e$, then there exists an edge e' in $\tilde{\Gamma}_n$ going from $\overline{v}$ to $\overline{w}$ with $e' \neq f(e')$, namely $e' = f(e)$. However, e and e' have the same label. (If e corresponds to the simple path $w^{(0)}w^{(1)}\dots w^{(m)}$ in Γ_n, then e' corresponds to the simple path $\overline{w^{(m)}}\dots\overline{w^{(1)}}\ \overline{w^{(0)}}$ in Γ_n. Lemma 8 implies that the label of these simple paths is the same.) These considerations lead to the following estimate

$$\#\{\rho(e)|\ e \in \mathcal{L}_{n+1}(u)\} \leq A + \frac{1}{2}B = \frac{1}{2}A + \frac{1}{2}(A+B) \tag{5}$$

Rewriting Equation (3), we obtain

$$A + B = \Delta C(n) + 2\sum_{w \text{ RS in } \mathcal{L}_n} 1 - \sum_{w \text{ BS in } \mathcal{L}_n} 1.$$

The statement follows then using Equation (4). □

Theorem 10. *Let u be an infinite word whose language $\mathcal{L}(u)$ is reversal closed and such that for every factor $w \in \mathcal{L}(u)$, the frequency $\rho(w)$ exists. Then for every $n \in \mathbb{N}$, we have*

$$\#\{\rho(e)|e \in \mathcal{L}_{n+1}\} \quad \leq \quad 2\Delta C(n)+1-\frac{1}{2}\left(\sum_{w \text{ BS in } \mathcal{P}al_n} 1+\sum_{w \text{ BS in } \mathcal{L}_n} 1\right) \quad \leq \quad 2\Delta C(n)+1.$$

The equality $\#\{\rho(e)|e \in \mathcal{L}_{n+1}(u)\} = 2\Delta C(n)+1$ holds for all sufficiently large n if and only if u is periodic.

Remark 11. To approve that the estimate from Theorem 10 cannot be easily lowered keeping its general validity, let us demonstrate that it is reached for all lengths $n \in \mathbb{N}$ in the case of Sturmian words. Thanks to [4], we know that every Sturmian word is reversal closed and all BS factors are palindromes. Moreover, since $\Delta C(n) = 1$ for all $n \in \mathbb{N}$, the upper bound on the number of different frequencies can be simplified as follows

$$\#\{\rho(e)|e \in \mathcal{L}_{n+1}(u)\} \quad \leq \quad 3 \; - \sum_{w \text{ BS in } \mathcal{L}_n} 1.$$

To see that the upper bound is reached, it suffices to recall the result of Berthé in [2]

$$\#\{\rho(e)|e \in \mathcal{L}_{n+1}(u)\} = \begin{cases} 2 & \text{if } n \text{ is the length of a BS factor,} \\ 3 & \text{otherwise.} \end{cases}$$

Proof of Theorem 10. It has been shown in [1] that

$$P(n)+P(n+1) \leq \Delta C(n)+2 \quad \text{for every } n \in \mathbb{N}. \tag{6}$$

The term $\sum_{w \text{ RS in } \mathcal{L}_n} 1$ can be bounded by $\sum_{w \text{ RS in } \mathcal{L}_n}(\#Rext(w)-1) = \Delta C(n)$. Applying these bounds on the result of Lemma 9, we obtain

$$\#\{\rho(e)|e \in \mathcal{L}_{n+1}\} \leq 2\Delta C(n)+1-\frac{1}{2}\left(\sum_{w \text{ BS in } \mathcal{P}al_n} 1+\sum_{w \text{ BS in } \mathcal{L}_n} 1\right).$$

Let us turn our attention to eventually periodic words. Since $\mathcal{L}(u)$ is reversal closed, it follows immediately that u is recurrent. If u is eventually periodic and recurrent, then u is known to be periodic. Thus, there exists a minimal period K such that $u = z^\omega$, where $|z| = K$. Then, $C(n) = K$ for every $n \geq K$ and every factor of length n occurs with frequency $\frac{1}{K}$. Thus, $\#\{\rho(e)|\ e \text{ edge in } \Gamma_n\} = 2\Delta C(n)+1 = 1$ for $n \geq K$. If u is aperiodic, then $\Delta C(n) \geq 1$ together with the fact that every LS factor is prefix of a BS factor implies that for every $N \in \mathbb{N}$, there exists a BS factor in $\mathcal{L}(u)$ of length $n \geq N$, hence $\#\{\rho(e)|\ e \text{ edge in } \Gamma_n\} \leq 2\Delta C(n)+1-\frac{1}{2}\left(\sum_{w \text{ BS in } \mathcal{P}al_n} 1+\sum_{w \text{ BS in } \mathcal{L}_n} 1\right) < 2\Delta C(n)+1$. □

For completeness' sake, let us mention another proof which will not use Equation (6), nevertheless, similar ideas as those ones occurring in [1] will be present. Going through this second version of the proof, it can be in particular noticed that Theorem 10 does not

require uniform recurrence of the infinite word u. We will keep notation from Proof of Lemma 9 and we will make use of a partial result rewritten in a different way this time:

$$\#\{\rho(e)|e \in \mathcal{L}_{n+1}(u)\} \leq A + \frac{1}{2}B = (A+B) - \frac{1}{2}B. \tag{7}$$

We want to find a lower bound on B, i.e. on the number of edges in $\tilde{\Gamma}_n$ which are not mapped by f on themselves. $\tilde{\Gamma}_n$ contains the following disjoint subgraphs (whose union comprises all vertices of $\tilde{\Gamma}_n$) of three types:

1. subgraphs containing two vertices w and $\overline{w}$, where w is RS not LS, and all edges connecting them mutually
2. subgraphs containing two vertices w and $\overline{w}$, where w is non-palindromic BS, and all edges connecting them mutually (attention! number of subgraphs of this type is just $\frac{1}{2}\#\{w \in \mathcal{L}_n(u) \mid w \text{ non-palindromic BS}\}$)
3. subgraphs containing one vertex w, where w is a palindromic BS, and eventually edges-loops starting and ending in w

Clearly, all edges in $\tilde{\Gamma}_n$ which are mapped by f on themselves are contained in the above subgraphs. Since (reduced) Rauzy graphs of infinite words are connected, each subgraph is connected with an edge to the union of the remaining subgraphs. Moreover, since the language $\mathcal{L}(u)$ is reversal closed, if an edge e starts in a subgraph Γ and ends in a subgraph Γ', then the edge $f(e)$ starts in Γ' and ends in Γ. It follows that B is greater or equal to $2\times$ the minimal number of edges which can ensure connection of the disjoint subgraphs of the graph:

$$B \geq 2 \times \text{number of subgraphs} - 2 = 2 \sum_{w \text{ RS in } \mathcal{L}_n} 1 + \sum_{w \text{ BS in } \mathcal{P}al_n} 1 - \sum_{w \text{ BS in } \mathcal{L}_n} 1 - 2. \tag{8}$$

Implanting in Equation (7) the just deduced lower bound on B together with the expression of $A + B$ derived in Proof of Lemma 9

$$A + B = \Delta C(n) + 2 \sum_{w \text{ RS in } \mathcal{L}_n} 1 - \sum_{w \text{ BS in } \mathcal{L}_n} 1,$$

and with the fact that $\sum_{w \text{ RS in } \mathcal{L}_n} 1$ can be bounded by $\sum_{w \text{ RS in } \mathcal{L}_n}(\#Rext(w) - 1) = \Delta C(n)$, we have proved the upper bound from Theorem 10

$$\#\{\rho(e)|\ e \text{ edge in } \Gamma_n\} \leq 2\Delta C(n) + 1 - \frac{1}{2}\left(\sum_{w \text{ BS in } \mathcal{P}al_n} 1 + \sum_{w \text{ BS in } \mathcal{L}_n} 1\right).$$

To conclude, let us throw in that we have studied frequencies of infinite words associated with β-integers for β being a quadratic non-simple Parry number, thus defined over a two-letter alphabet, and we have learned that the upper bound from Theorem 10 is either reached (for most of the lengths) or is only by 1 greater than the real number of factor frequencies of a given length. Another example of an infinite word, even over a k letter alphabet, where the upper bound is reached for all lengths, is the k-interval exchange word. (Description of frequencies has been recently given by Ferenczi [5].)

References

[1] P. Baláži, Z. Masáková, E. Pelantová. *Factor versus palindromic complexity of uniformly recurrent infinite words*. Theoret. Comput. Sci. vol. **380**, issue **3** (2007), 266–275.

[2] V. Berthé. *Fréquences des facteurs des suites sturmiennes*. Theoret. Comput. Sci. **165** (1996), 295–309.

[3] M. Boshernitzan. *A condition for minimal interval exchange maps to be uniquely ergodic*. Duke Math. **52** (1985), 723–752.

[4] X. Droubay, J. Justin, G. Pirillo. *Episturmian words and some constructions of de Luca and Rauzy*. Theoret. Comput. Sci. **255** (2001), 539–553.

[5] S. Ferenczi, L. Zamboni. *Combinatorial structure of symmetric k-interval exchange transformation*. Preprint available at http://iml.univ-mrs.fr/∼ferenczi/

Parallel Algorithm for Numerical Solution of Nonlinear Evolution Equations

Stanislav Brand

2nd year of PGS, email: brands1@kmlinux.fjfi.cvut.cz
Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU
advisor: Michal Beneš, Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU

Abstract. The article briefly summarizes numerical solution of Mean-Curvature Flow problem for the Level Set method and its parallelization. The numerical solution is based on spatial discretization by finite differences, and time discretization is given by the Runge-Kutta and the Runge-Kutta-Merson scheme. The algorithm is parallelized for the shared memory systems using OpenMP. The computational results demonstrate the time evolution of the initial curve under given curvature.

Abstrakt. Tento příspěvek krátce shrnuje numerické řešení rovnice Mean-Curvature Flow, s použitím Level Set methody, a možnosti paralelizace. Řešení se skládá z diskretizace prostorové oblasti pomocí konečných diferencí a následného řešení časové úlohy pomocí Runge-Kuttovy a Runge-Kutta-Mersnovy metody. Algoritmus byl paralelizován pro systémy se sdílenou pamětí pomocí OpenMP. Výpočetní výsledky ukazují časový vývoj počáteční křivky pod vlivem křivosti.

1 Introduction

The equation describes the motion of the hypersurface $\Gamma(t)$ with the velocity, which is its mean curvature. The problem can be written as

$$v_\Gamma = -K_\Gamma + F \qquad \text{on } \Gamma(t), \tag{1}$$

where v_Γ is the normal velocity, K_Γ is the mean curvature of the hypersurface $\Gamma(t)$ and F is the forcing term. Such a system of equations has been studied by many authors throughout last decade (see [2] ,[3], [4], [5], [6]).

We would like to track possible topological changes of $\Gamma(t)$ and therefore we have chosen the Level Set method. This method describes the hypersurface $\Gamma(t)$ as the zero Level Set of an auxiliary function $P : \mathbb{R}^{n+1} \times [0, \infty) \to \mathbb{R}$, i.e.

$$\Gamma(t) = \{x \in \mathbb{R}^{n+1} | P(x, t) = 0\}. \tag{2}$$

Assume that $\nabla P \neq 0$ in a neighbourhood of $\Gamma(t)$. Then the outer normal to $\Gamma(t)$, its

mean curvature and the normal velocity are given by

$$n_\Gamma = -\frac{\nabla P}{|\nabla P|}, \tag{3}$$

$$K_\Gamma = \operatorname{div}(n_\Gamma) = -\operatorname{div}\Big(\frac{\nabla P}{|\nabla P|}\Big), \tag{4}$$

$$v_\Gamma = \frac{\partial_t P}{|\nabla P|}. \tag{5}$$

Let Ω be a bounded domain in $\mathbb{R}^{n+1}$.The equation (1) yields into partial differential equation (PDE) for P.

$$\frac{\partial_t P}{|\nabla P|} = \operatorname{div}\Big(\frac{\nabla P}{|\nabla P|}\Big) + F \qquad \text{on } \Omega \times [0,\infty), \tag{6a}$$

$$P(x,0) = P_{ini}(x) \qquad \text{on } \Omega, \tag{6b}$$

$$\frac{\partial P}{\partial n_\Gamma} = 0 \qquad \text{on } \partial\Omega. \tag{6c}$$

2 Numerical algorithm

The equation (6) is highly nonlinear, degenerate parabolic PDE. We solved it by the method of lines. This technique for solving PDEs starts with discretising all but one dimension using finite difference and then solve resulting semi-discrete problem, that is a set of ordinary differential equations (ODEs).

2.1 Spatial discretization

We performed spatial discretization using two different finite differential scheme. Schemes are written using following notation.

$$h = [h_1, h_2], h_1 = \frac{L_1}{N_1}, h_2 = \frac{L_2}{N_2}, x_{i,j} = [x^1_{i,j}, x^2_{i,j}], P_{i,j} = P(x_{i,j}), \tag{7}$$

$$P_{\bar{x_1},i,j} = \frac{P_{i,j} - P_{i-1,j}}{h_1}, P_{x_1,i,j} = \frac{P_{i+1,j} - P_{i,j}}{h_1}, \tag{8}$$

$$P_{\bar{x_2},i,j} = \frac{P_{i,j} - P_{i,j-1}}{h_2}, P_{x_2,i,j} = \frac{P_{i,j+1} - P_{i,j}}{h_2}, \tag{9}$$

$$\bar{\nabla}_h P = [P_{\bar{x_1}}, P_{\bar{x_2}}], \nabla_h P = [P_{x_1}, P_{x_2}], \tag{10}$$

$$P_n = \mathcal{P}_n P. \tag{11}$$

2.1.1 Regularized scheme

This scheme is based on operator form of equation (6), where we substitute derivative operators using discrete differential operators (10), (see [4])

$$\frac{\mathrm{d}P_n}{\mathrm{d}t} = \bar{Q}\nabla_n \cdot \Big(\frac{\bar{\nabla}_n P_n}{Q(\bar{\nabla}_n P_n)}\Big) + \bar{Q}F, \tag{12}$$

where

$$Q(\nabla P) = \sqrt{\epsilon^2 + ||\nabla P||^2}, \qquad \epsilon > 0, (\epsilon = 10^{-9}), \tag{13}$$

$$\bar{Q} = \frac{1}{2}(Q(\bar{\nabla}_n P_n) + Q(\nabla_n P_n)). \tag{14}$$

2.1.2 Central difference scheme

Expanding divergence and gradient operators in the equation (6) yields

$$\begin{aligned} \partial_t P = & \frac{\partial_{x_1 x_1} P(\partial_{x_2} P)^2 + \partial_{x_2 x_2} P(\partial_{x_1} P)^2 - 2\partial_{x_1} P \partial_{x_2} P \partial_{x_1 x_2} P}{(\partial_{x_1} P)^2 + (\partial_{x_2} P)^2} \\ & + \sqrt{(\partial_{x_1} P)^2 + (\partial_{x_2} P)^2} F \end{aligned} \tag{15}$$

Substituting spatial derivatives using central differences we get following ODE

$$\begin{aligned} \partial_t P = & \frac{\frac{P_{i,j+1}-2P_{i,j}+P_{i,j-1}}{\mathrm{d}x_2^2}\left(\frac{P_{i+1,j}-P_{i-1,j}}{2\mathrm{d}x_1}\right)^2 + \frac{P_{i+1,j}-2P_{i,j}+P_{i-1,j}}{\mathrm{d}x_1^2}\left(\frac{P_{i,j+1}-P_{i,j-1}}{2\mathrm{d}x_2}\right)^2}{\left(\frac{P_{i+1,j}-P_{i-1,j}}{2\mathrm{d}x_1}\right)^2 + \left(\frac{P_{i,j+1}-P_{i,j-1}}{2\mathrm{d}x_2}\right)^2} \\ & - \frac{2\left(\frac{P_{i+1,j}-P_{i-1,j}}{2\mathrm{d}x_1}\right)\left(\frac{P_{i,j+1}-P_{i,j-1}}{2\mathrm{d}x_2}\right)\left(\frac{P_{i+1,j+1}-P_{i+1,j-1}-P_{i-1,j+1}+P_{i-1,j-1}}{4\mathrm{d}x_1\mathrm{d}x_2}\right)}{\left(\frac{P_{i+1,j}-P_{i-1,j}}{2\mathrm{d}x_1}\right)^2 + \left(\frac{P_{i,j+1}-P_{i,j-1}}{2\mathrm{d}x_2}\right)^2} \\ & + \sqrt{\left(\frac{P_{i+1,j}-P_{i-1,j}}{2\mathrm{d}x_1}\right)^2 + \left(\frac{P_{i,j+1}-P_{i,j-1}}{2\mathrm{d}x_2}\right)^2} F \end{aligned} \tag{16}$$

2.2 Time discretization

To solve this system of ODE's we use the Runge-Kutta-Merson, that is a modified Runge-Kutta method with adaptive time stepping (see [11]). The time-step length adaptivity may shorten the time needed for computation. The algorithm can be written in the following form

1. compute $k1_{ij}(\mathrm{d}t)$
2. compute $k2_{ij}(\mathrm{d}t)$
3. compute $k3_{ij}(\mathrm{d}t)$
4. compute $k4_{ij}(\mathrm{d}t)$
5. compute $k5_{ij}(\mathrm{d}t)$
6. $q = \max\{|2k1_{ij}(\mathrm{d}t) - 9k3_{ij}(\mathrm{d}t) + 8k4_{ij}(\mathrm{d}t) - k5_{ij}(\mathrm{d}t)|/30\}$
7. if($q < adaptivity$)
8. {
9. $\quad y_i j(t0 + \mathrm{d}t) = y_i j(t0) + (k1_i j(\mathrm{d}t) + 4k4_i j(\mathrm{d}t) + k5_i j(\mathrm{d}t))/6$
10. $\quad t0 = t0 + \mathrm{d}t$
11. }
12. $\mathrm{d}t = \mathrm{d}t\omega(adaptivity/q)^{0.2}$

where coefficients $k1, \ldots, k5$ are defined as follows

$$\begin{aligned} k1(\mathrm{d}t) &= \mathrm{d}t f(t0, y(t0)) \\ k2(\mathrm{d}t) &= \mathrm{d}t f(t0 + \mathrm{d}t/3, y(t0) + k1(\mathrm{d}t)/3) \\ k3(\mathrm{d}t) &= \mathrm{d}t f(t0 + \mathrm{d}t/3, y(t0) + (k1(\mathrm{d}t) + k2(\mathrm{d}t))/6) \\ k4(\mathrm{d}t) &= \mathrm{d}t f(t0 + \mathrm{d}t/2, y(t0) + 0.125k1(\mathrm{d}t) + 0.375k3(\mathrm{d}t)) \\ k5(\mathrm{d}t) &= \mathrm{d}t f(t0 + \mathrm{d}t, y(t0) + 0.5k1(\mathrm{d}t) - 1.5k3(\mathrm{d}t) + 2.0k4(\mathrm{d}t)) \end{aligned} \tag{17}$$

where f is the right hand side of equations (12) or (16). We usually choose *adaptivity* $\in [10^{-6}, 10^{-3}]$,$\omega \in [0.8, 0.9]$.

3 Stability of numerical algorithm

We do not know the analytical solution for the equation (1), so to demonstrate stability and consistency, we have to use numerical results computed on a refined grid. We linearly interpolate the solution on the finest grid and compare it with the remaining solutions (see Tables 1,2 and 5,6).

For the initial condition, where the zero levelset is the circle with the radius r_0 and the forcing term $F = 0$ the equation (1) yields

$$\frac{\mathrm{d}r}{\mathrm{d}t} = -\frac{1}{r} \tag{18}$$

This equation has the exact analytical solution

$$r(t) = \sqrt{r_0^2 - 2t}. \tag{19}$$

Tables 3,4 and 7,8 presents convergence errors and EOC coefficients computed using exact solution (19).

Mesh h	$\mathcal{L}_\infty(0,T;\mathcal{L}_2)$ error of u	$\mathcal{L}_\infty(0,T;\mathcal{L}_\infty)$ error of u
0.2040816	0.0475471	0.2709150
0.1010101	0.0138025	0.1251500
0.0502513	0.0046021	0.0533660
0.0250627	0.0013799	0.0177440

Table 1: Table of convergence errors. Space discretization: Regularized scheme, Time discretization: RK-mersn scheme, *adaptivity* = 0.00001. Computed for the hypersurface Γ.

Mesh h	EOC u L_2	EOC u L_∞
0.2040816	0.0000000	0.0000000
0.1010101	1.7586649	1.0980984
0.0502513	1.5731378	1.2207927
0.0250627	1.7314871	1.5828592

Table 2: Table of EOC coefficients. Space discretization: Regularized scheme, Time discretization: RK-mersn scheme, *adaptivity* = 0.00001. Computed for the hypersurface Γ.

Mesh h	$\mathcal{L}_\infty(0,T;\mathcal{L}_2)$ error of u
0.2083333	0.0080304
0.1020408	0.0035612
0.0505050	0.0013680
0.0251256	0.0006267
0.0125313	0.0002926

Table 3: Table of convergence errors. Space discretization: Regularized scheme, Time discretization: RK-mersn scheme, *adaptivity* = 0.00001. Computed for the zero Level Set.

Mesh h	EOC u L_2
0.2083333	0.0000000
0.1020408	1.1392145
0.0505050	1.3604080
0.0251256	1.1179657
0.0125313	1.0951178

Table 4: Table of EOC coefficients. Space discretization: Regularized scheme, Time discretization: RK-mersn scheme, *adaptivity* = 0.00001. Computed for the zero Level Set.

Mesh h	$\mathcal{L}_\infty(0,T;\mathcal{L}_2)$ error of u	$\mathcal{L}_\infty(0,T;\mathcal{L}_\infty)$ error of u
0.2040816	0.0475471	0.2709150
0.1010101	0.0121113	0.1251500
0.0502513	0.0033617	0.0533660
0.0250627	0.0010117	0.0177440

Table 5: Table of convergence errors. Space discretization: Central difference scheme, Time discretization: RK-mersn scheme, *adaptivity* = 0.00001. Computed for the hypersurface Γ.

Mesh h	EOC u L_2	EOC u L_∞
0.2040816	0.0000000	0.0000000
0.1010101	1.9445204	1.0980984
0.0502513	1.8357707	1.2207927
0.0250627	1.7261743	1.5828592

Table 6: Table of EOC coefficients. Space discretization: Central difference scheme, Time discretization: RK-mersn scheme, *adaptivity* = 0.00001. Computed for the hypersurface Γ.

Mesh h	$\mathcal{L}_\infty(0,T;\mathcal{L}_2)$ error of u
0.2083333	0.0047062
0.1020408	0.0011564
0.0505050	0.0002992
0.0251256	0.0000714
0.0125313	0.0000193

Table 7: Table of convergence errors. Space discretization: Central difference scheme, Time discretization: RK-mersn scheme, *adaptivity* = 0.00001. Computed for the zero Level Set.

Mesh h	EOC u L_2
0.2083333	0.0000000
0.1020408	1.9664470
0.0505050	1.9224476
0.0251256	2.0527963
0.0125313	1.8825815

Table 8: Table of EOC coefficients. Space discretization: Central difference scheme, Time discretization: RK-mersn scheme, *adaptivity* = 0.00001. Computed for the zero Level Set.

4 Numerical results

This section contains results for the Mean-Curvature Flow problem. Results are represented by graphs displaying the Level Set hypersurface. The solution was computed at the time interval $[0, 1.4]$ using the space domain $[-3, 3] \times [-3, 3]$ with the grid containing 200×200 points and the Neumann boundary condition.

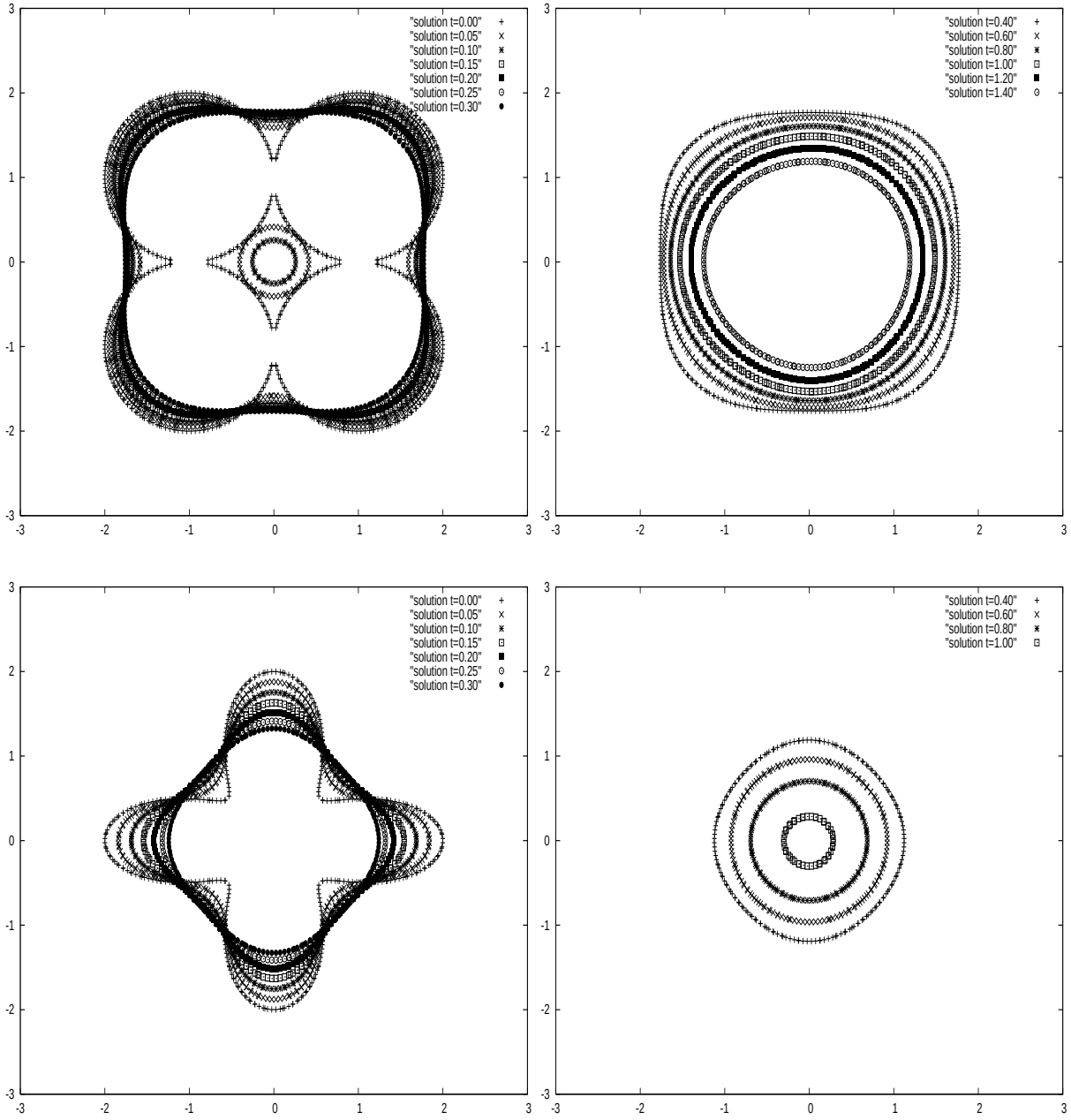


Figure 1: Solution of (6) with $f = 0$.

5 Parallelization of numerical algorithms

The main purpose of our work was to compare the efficiency of parallel algorithms for numerical solution of the Mean-Curvature Flow problem on systems with shared memory. Shared memory means that all data are saved in the memory that can be accessed by all CPUs. This concept is used by OpenMP API (see [1]).

The results of efficiency measurement are presented in Table 9. The table has the following structure. The first column contains the grid dimension. The second column contains the time of sequence program in seconds, this means the time of the computation made by only one processing unit. The remaining columns contain the time of parallel program and the efficiency of this program in the brackets. In the header of these columns there is specified how many OpenMP processes were used in the computation. The time duration of each computation was measured by the C gettimeofday() function as a difference between the start and the end time. The times listed in here are the times needed for the computation only. This means the times needed for the value initialization and result saving is excluded. The efficiency is calculated from the following formula:

$$\text{efficiency} = \frac{\text{sequence time}}{\text{parallel time} \times \text{number of processors}} \tag{20}$$

grid		OMP=2
50×50	6.100	4.714(0.647)
100×100	25.609	15.298(0.837)
150×150	62.783	35.192(0.892)
200×200	114.270	61.042(0.936)
250×250	178.611	98.789(0.904)
300×300	261.127	145.881(0.895)
350×350	357.475	194.069(0.921)
400×400	438.124	233.293(0.939)
450×450	550.205	287.163(0.958)
500×500	679.820	361.222(0.941)

Table 9: Time and efficiency of parallel program using Central difference scheme, Time discretization: RK-mersn scheme, $adaptivity = 0.001$.

6 Conclusion

We solved the Mean-Curvature Flow problem for the Level Set method using several different initial conditions. Numerically proved the stability and consistency of both Regularized and Central difference schemes in combination with RK-mersn scheme and shown that OpenMP is suitable for this equation.

References

[1] *OpenMP Application Program Interface.* http://www.openmp.org/drupal/mp-documents/spec25.pdf.

[2] M. Beneš. *Analysis of Equations in the Phase Field Model.* Masaryk University, Brno, (1998). 17-35, 1998, proceedings on the conference Equadiff 97.

[3] M. Beneš. *Mathematical and computational aspect of solidification of pure substances.* Acta Math. Univ. Comenianae, (2000). Act I, Scene 3, Lines 70–72, are apropos.

[4] G. Dzuik. *Mean curvature flow and related topics.* Springer, Berlin-Heidelberg-New York, (2003).

[5] G. Dzuik. *Numerical approximations of mean curvature flow of graphs and level sets.* Springer, Berlin-Heidelberg-New York, (2003).

[6] J. Sethian. *Level Set Methods and Fast Marching Methods: Evolving interfaces in computational geometry, fluid Mechanics, computer vision, and materials science.* Cambridge University Press, Cambridge, UK, (1999).

Position Sensitive Detector Data Acquisition Path*

Martin Dráb

4th year of PGS, email: drab@kepler.fjfi.cvut.cz
Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU
advisor: Ladislav Kalvoda, Department of Solid State Engineering, Faculty of Nuclear Sciences and Physical Engineering, CTU

Abstract. The specific initial and main aim of the project INDECS was to drive the neutron diffractometer KSN-2 in its upgraded form maintaining the position sensitive detectors (PSD). For the purpose of collecting data from these detectors a special complex structure called PSD Acquisition Path (PSDAP) was created and it will be integrated into the concept of project INDECS as one of its External Execution Modules (EEM).

Abstrakt. Původním hlavním cílem projektu INDECS bylo řídit neutronový difraktometr KSN-2, který ve své vylepšené podobě obsahuje pozičně citlivé detektory (PSD). Za účelem sbírání dat z těchto detektorů byla vytvořena speciální komplexní struktura nazvaná PSD Acquisition Path (PSDAP), která bude integrována do konceptu projektu INDECS v podobě External Execution Modulu (EEM).

1 Introduction

Upgrade of the KSN-2 neutron diffractometer from a simple (one channel) counting detector type to a type with a set of position sensitive detectors (PSD) required a completely different and much more complicated way of collecting data from the detectors. It was also one of the reasons for launching project INDECS, to create a software that would do just that, among other things related to driving the diffractometer.

In this article we would briefly describe the design of the part of the project INDECS that is meant to collect data from the PSDs of KSN-2 diffractometer and do its basic evaluation to the form of a neutron-counting histogram, which is supposed to be the raw output of the neutron diffractometer for the physicists, who then use this form of data to do further processing and thereby extracting other studied information about the measured samples. This part is called the PSD Acquisition Path or the PSDAP.

At some points we are going to be rather specific on the implementation and hardware that is currently used for the KSN-2 data collection and processing, but the global concept of the PSDAP is designed in a way that its individual parts can be replaced with adequate parts for different hardware setup, should this change in the future, or to adapt it for a different diffractometer setup.

*This work has been supported by grants MSM6840770021 and JINR 22-03007.

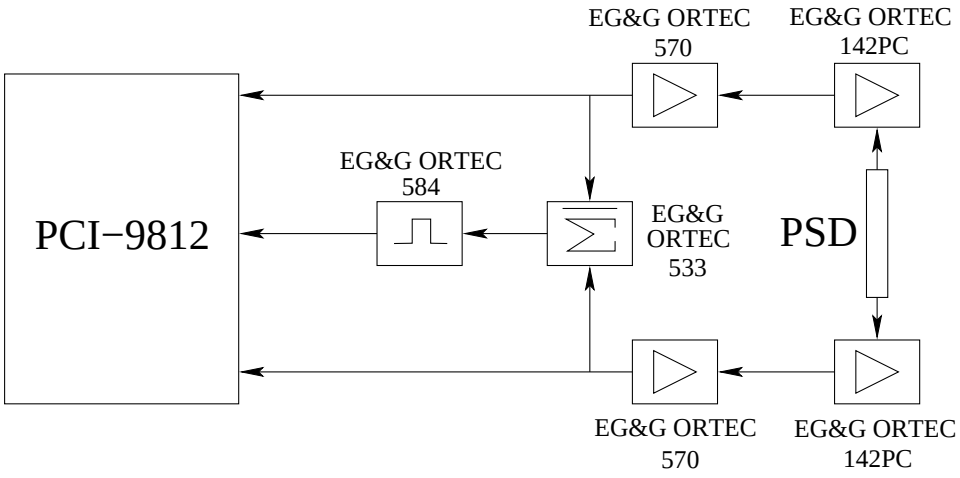


Figure 1: Electronics of the position sensitive detector used on the KSN-2 neutron diffractometer.

1.1 KSN-2 detector hardware

The hardware surrounding the actual PSD on the KSN-2 diffractometer is shown in a block scheme on figure 1. For a little more specific description, see for instance [2] (chapter 2), and for the description of the principles of the PSD, see the [1] or any other common literature about the PSDs.

For our purposes it is sufficient to know, that the PSD acts like a big resistor and that the signal pulse generated by the event of an incoming neutron is split in two and travels through the resistor to either end of the detector, where it is preamplified by a preamplifier (EG&G Ortec 142PC in our case) and then amplified and shaped by the shaping amplifier, which in our case is the EG&G Ortec 570), where not only is the pulse amplified to a level that we can further sample, but it is also given a proper and quite nicely looking (compared to the actual signal coming out of the preamplifier) semi-gaussian shape, an example of which you can see on figure 2. This signal on either side of the PSD is then sampled by an A/D converter card, which in the case of KSN-2 is the ADLink PCI-9812, and the sampling is triggered by the predefined trigger level to distinguish the higher peaks of a neutron event from the much lower peaks produced by different sources of radiation. And this is the point where the role of the PSDAP begins.

2 PSD Acquisition Path

The PSDAP has several processing steps to do. First the raw sampled signal has to be obtained. Then it can possibly be split into multiple neutron events, which may be detected by one signal trigger event. After that a position of the event on the detector has to be determined. And finally the event's position has to be written to the adequate bin in the resulting histogram. As these are the steps that are done sequentially one by one with each signal pulse, we call this processing tool a "path", because each signal pulse has to walk this processing path step by setp from the source into the hostogram.

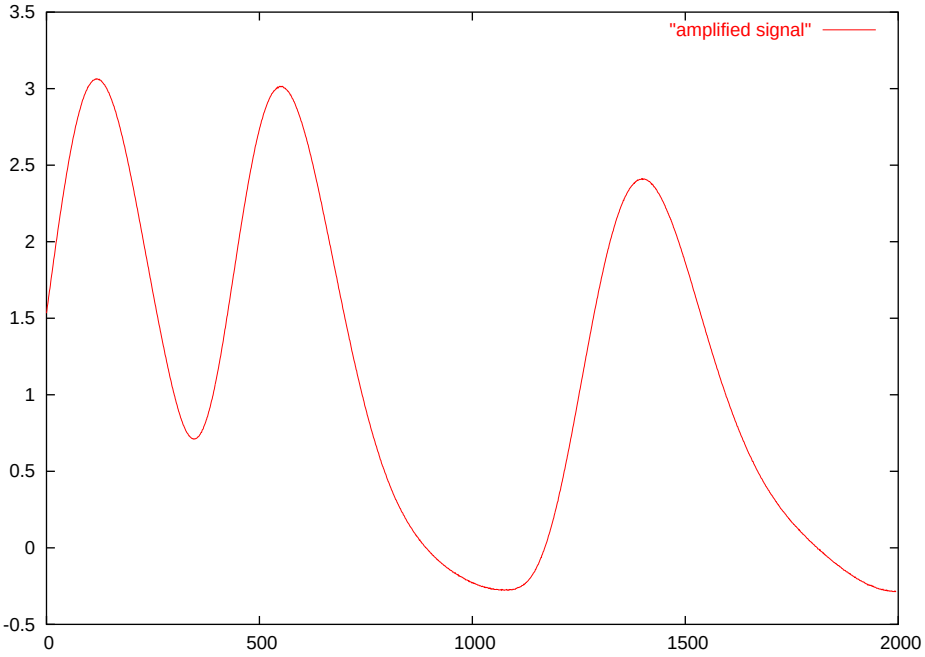


Figure 2: Semi-gaussian-shaped signal that comes out of the shaping amplifier upon an event detection.

There are two ways to incorporate this into the structure of project INDECS' resources. One way would be to create each step along the path as a separate External Execution Module (EEM, see [3]) and let the data be pushed between them by virtual instructions passing through the Execution Engine (EE, see [3]) VPU each time. Though this might be much more flexible, it is also a bit more and unnecessarily complicated and more importantly, rather slow, as the events are coming at rates about thousands of events per second and quite large amount of data must be pushed through the path (up to 64 KB per event for the PCI-9812). Since we need to miss as little events as possible we also need to process the sampled data as fast as possible to be able and ready to process next.

For this reason the other way of implementing was chosen. The entire PSDAP is implemented as just one complex EEM, which however is consisting of submodules, each doing one step along the path. The path has to be configured by assembling appropriate submodules of the PSDAP together before it can be used. For some modules there are more possibilities to choose from different modules doing slightly different work. Let's describe them a little closely in the following subchapters.

The top-level implementation of the PSDAP is substantiated by a library called `libpsdap` and all the modules are implemented on the MMSR level of project INDECS.

2.1 Data Source

The Data Source is the first module at the entry of the path. Its purpose is to acquire the raw sampled signal from the signal source, decode them and send them further along

the path. Currently there exist two types of the Data Source module.

2.1.1 PCI9812

The PCI9812 Data Source module is the Data Source module that acquires the raw signal data from two channels of the ADLink PCI-9812 A/D converter card, decodes them into two separate signal sequences and sends them together with a timestamp and data size in a third channel to the next step of the PSDAP. Because (as said above) we want to miss as little neutron events as possible (and also because the computer so far dedicated to data acquisition for the first channel of the KSN-2 is just a PII 400 MHz), we are currently acquiring data under the RTLinux OS, which is an implementation of the hard real-time OS running under common Linux itself (see [2]).

A special real-time driver was written for the PCI-9812. This driver is launched as a real-time thread scheduled by the hardware interrupt of the PCI-9812, so that it gets a very high priority and can run anytime necessary regardless of the other state of the OS. The data are transferred using fixed number of preallocated DMA buffers, one per event, and pushed through the RT-FIFO mechanism to the non-real-time application process, which is the PSDAP.

There are several parameters that can be preset, among which the most important ones are the trigger level which is used for triggering the event. The event length, which determines how much samples are samplet per one event and, though, what would be the length of the sampled signal. And the sampling rate at which the signal would be sampled. Depending on the setting of the shaping constants of the shaping amplifier, the optimal frequency to catch the event should not be much less than 10 MHz, but to have some reserve for the signal analysis a full 20 MHz sampling rate of the PCI-9812 is recommended.

One downside of the RTLinux implementation of the driver is that its free variant is so far only implemented on the old 2.4.x Linux kernel architecture, which is now obsoleted by the progressive 2.6.x version. However, with a reasonably new PC computer, a litle different variation of the driver (using partly dynamic buffer queues) can possibly be created and run even on standard non-real-time Linux 2.6.x kernels. This work has, however, not been started, yet.

2.1.2 SDCF

The SDCF Data Source module is an analogy of the PCI9812 module, but it does not acquire the data from the PCI-9812 card, but is decoded from a SDCF file, into which the raw signal can be stored by the below described Signal Storage module.

This decoding module actually consists of three MMSR transcoders. First transcoder reads the data from the given SDCF file. Second is a general demultiplexer transcoder for the SDCF streams. This part of the decoder is implemented by the `libmdsdcf` library, which, of course, uses the general Multiplexer/Demultiplexer (`libmd`), the SDCF, the MMSR and the Stream Cache library. This general demultiplexer extracts the global header packets of the SDCF stream and first two data substreams, all of which are separately decoded by its individual decoder transcoders each. These stream decoder transcoders are also implemented within the `libmdsdcf`. When the first two data

streams are decoded, their data along with appropriate timestamps are forwarded to the last transcoder of this decoding module, which is the Synchronization Barrier Decoder transcoder (also called "synbar"), which basically takes the data and timestamps from the two separate independent channels and puts the data from the same time together. So it synchronizes the two channels to be sent along with the timestamp and datalength information in a separate third channel further along the PSDAP.

2.2 Signal Storage

Signal Storage module is a module that is used to store the signal data somewhere, possibly into a file of some kind, otherwise it is transparent, so it sends out the same data that it receives. This also means, that there can possibly be more of these modules chained at the specific point of the the PSDAP, though it is strongly discouraged, as storage itself may be quite a delaying work at these datarates and if the computer and its relevant peripherials aren't fast enough, the whole processing can possibly be delayed so much, that it may miss some neutron events that would normally be detected and thereby the effectivity of the whole system can go down.

The signal storage modules can be placed both after the Data Source module and after the Multi Event Separator module (see below).

2.2.1 SDCF

The SDCF Signal Storage module is used to store the sampled event (meaning a signal on two channels and a timestamp) into a SDCF file. It is using the SDCF library for that and it is a reversed process to that of the SDCF Data Source module. Each data channel is stored in a separate SDCF data substream and before each event a global header with a synchronization timestamp is forced. This is (so far) the most effective way to store the acquired signal, in the means of redundancy, however no signal compression on the SDCF file has been implemented, yet.

2.2.2 RAW

The RAW Signal Storage module is an analogy of the SDCF Signal Storage module, but instead of storing the event signals into the SDCF file, it stores them into a raw text file, one event per file. The format is simple, just two columns of signal data and commented header containing information like timestamp and sampling frequency. This is not the most effective way of storing the event data, but it comes handy when you want to do an eye inspection of the data or manual processing of the signal by other programs like GNU Plot or MatLab, which can easily read data from raw text files.

2.2.3 Send/Receive

This is a little different kind of module. It is a communication module, that can generate a special kind of data transfer virtual instructions and send them to the predefined target of the EE that the EEM containing this module is attached to. Instead of passing the event signal data through and further along the PSDAP, the data are actually diverted from

this PSDAP and sent via these data transfer virtual instructions for further processing to another PSDAP, possibly in another computer.

When this kind of virtual instructions arrive to the PSDAP EEM, it can be received by the same module at the same position of that PSDAP, decoded and sent further along the PSDAP for further processing there. The main purpose of this module is to be able to divert the next processing steps of the acquiring PSDAP to another computer, if the acquiring computer is not fast enough to do all the processing.

2.3 Multi Event Separator

The Multi Event Separator is another step in the PSDAP path. Its purpose is to separate possible multiple events sampled at one shot on one trigger. So if during the original signal acquisition we sample a signal of some length, where more than one neutron event occurs. This signal can be split into multiple events and sent further along the PSDAP as separate events.

There can be various methods of event separation, but given that the shaping preamplifier gives us nice semi-gaussian shapes of the event pulse and that it has certain dead time to prevent total overlapping and noise from other unwanted radiation, we can very well use the easiest method of separation by the same trigger level as by which the initial acquisition was fired. Of course, another methods of separation can possibly be investigated and appropriate modules written in the future.

2.4 Peak Analyzer

The Peak Analyzer module is the part that does the main processing along the PSDAP. It takes the sampled signal from the two ends of the PSD on the input. And calculates the position of the event along the PSD where it occured.

The point is to find the corresponding peaks (generated by the event) on either of the channels, the peaks can not be further from one another than is the time needed to travel from one end of the PSD to another, and by comparing the heights of the peak from the two channels determine the position. Because from each position on the PSD the signal travels a specific distance accross the PSD to either end of the detector, and since it is a big resistor, the further the signal travels there the bigger the resistivity it has to pass through and though the lower the ampliture it has. So when the event occurs in the middle of the detector, the peaks are the same height, when it is close to one end of the detector, the corresponding peak is high, while the other is much lower and vice versa.

This, however, as much of another physically measured variables, is introduced with certain distortion, so each particular detector should be callibrated by covering it with shielding and opening just on several channels (positions) of the PSD. This can construct a compenstation curve, which is then used as a transformation function for calculating the exact position. Peak Analyzer module can construct this compensation curve when run under a special mode. The curve can be sent or received by special virtual instructions.

2.5 Event Storage

The Event Storage module has similar functionality to the signal storage module, but so far only a transparent module that can send the events via special virtual instructions exists. The event in this case is represented only by a single number determining the position (also called channel) on the position sensitive detector, where the neutron event occured. No other variant of this module has been created, yet. But if there would be any use for it, another variants can be crated in the future.

2.6 Histogrammer

This is the final point of the PSDAP. It is a module which maintains a histogram with a preset number of bins (that generally represents the number of channels of the PSD given by its resolution). And the events, that come in the form of the position, are sorted into appropriate bins and counted there. Resulting histogram can be sent upon request to the specified target of the attached EE and then used for further processing outside of the PSDAP. The histogram can also be reset by a virtual instruction.

2.7 Final Notes

The PSDAP EEM can send and receive virtual commands, some of which have been mentioned above. Another of these commands are a start and stop commands, which determine when to start and when to stop acquiring data. You can set various parameters of the PSDAP by sending it virtual commands, including its configuration and compensation curve. You can make is start and/or stop by an external event coming from the INDECS system, namely it can be a timer for a measurement over a specific period of time, or it can be a threshold on the monitor detector counter, so that the measurement stops after a certain number of neutrons entering the measured sample, and so on. And finally you can let the PSDAP send you some status information about the processing.

3 Conclusion

Most of the parts of the PSDAP are finished already, some of them are close to be finished. The PSDAP still has to undergo some real testing, so far we are testing it only with the data that we have collected separately. A full integration of the PSDAP into the INDECS system and its thorough testing has to be done. Also EEMs and drivers for another devices of the KSN-2 neutron diffractometer still need to be written, so that the KSN-2 can be fully driven by the INDECS system. Motor handling and temperature control are some of them.

References

[1] M. Dráb. *Project: INDECS*. (Rešeršní práce) FJFI, ČVUT, Praha, (2001).

[2] M. Dráb. *Project: INDECS*. (Undergraduate Research Work) Faculty of Nuclear Sciences and Physical Engineering, Czech Technical University, Prague, (2002).

[3] M. Dráb. *Project INDECS: The Design of a Superior System of Data Collection and Analysis*. (Diploma Thesis) Faculty of Nuclear Sciences and Physical Engineering, Czech Technical University, Prague, (2003).

Dynamic Effect in Capillary Pressure-Saturation Relationship

Radek Fučík*

2nd year of PGS, email: radek.fucik@email.cz
Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU
advisor: Jiří Mikyška, Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU

Abstract. Classical models for capillary pressure - saturation relationship have been shown to hold only in the state of thermodynamical equilibrium and thus a modified dynamic capillary pressure-saturation relationship has been proposed by Hassanizadeh et al. that includes the classical capillary pressure function in the state of thermodynamic equilibrium and a product of relaxation time τ and a partial time-derivative of saturation. However, this approach importantly affects the nature of the governing two-phase flow equations for immiscible and incompressible flow in porous media. New numerical and theoretical studies are required in order to understand this phenomena. In the presented work, a onedimensional problem and an implicit numerical scheme is presented to model various effects of the order in magnitude of the dynamic effect coefficient τ on the saturation and pressure profiles in homogeneous porous medium.

Abstrakt. Klasické modely pro závislost kapilárního tlaku na saturaci platí pouze ve stavu termodynamické rovnováhy. Hassanizadeh et al. navrhují nový model pro kapilární tlak, který zahrnuje jak klasickou funkci kapilárního tlaku na saturaci v termodynamické rovnováze, tak nově i časovou derivaci saturace násobenou relaxačním časem τ. Tento přístup ovšem mění typ doposud používaných rovnic pro simulaci dvoufázového nestlačitelného a nemísivého filtračního proudění a je proto zapotřebí zjistit, jakým způsobem se změní stávající modely při implementaci dynamického kapilárního tlaku. V této práci je uvažována jednorozměrná úloha, která je řešena konečnými diferencemi. Navrhnuté numerické schéma je použito k porovnání profilů saturace a tlaku pro řádově různé hodnoty relaxačního parametru τ.

1 Introduction

In many countries, more than half of the population depend on groundwater as their supply for drinking water. The groundwater sources are often endangered by leaks from disposal dumps, accidental spills of substances used in industry or leaking storage tanks. Mathematical modelling is one of the important tools that helps to predict the spreading of the contaminant in the water saturated zones. In addition, the mathematical models can facilitate extraction of valuable substances such as oil or gas.

This manuscript focuses on the dynamic phenomena in the capillary pressure - saturation relationship that has been examined in various papers in the past decades. The main objective is to propose a numerical scheme that implements the dynamic capillary pressure - saturation relationship for heterogeneous porous media. In this report, only

*www.RadekFucik.com

preliminary results of numerical experiments in the homogeneous porous medium are given in order to allow for future generalization of the numerical code also for heterogeneous porous media.

2 Background

Fundamental constitutive quantities used in modelling flow in porous media are described in the following subsections. Thorough definitions, descriptions, and examples can be found in [8], [17], [1], [16], [2], [6], or [7].

2.1 Wettability

As two immiscible phases are present in the porous media, a meniscus of fluid-fluid interface is formed as a result of the presence of the solid phase (sand grains). The interaction between adhesive and cohesive forces within the fluids leads to the specific angle ϑ between the solid surface and the fluid-fluid interface. The wettability of fluid is then determined as:

$\vartheta = 0$	$\vartheta \in (0, \frac{\pi}{2})$	$\vartheta > \frac{\pi}{2}$
completely wetting,	partially wetting,	non-wetting.

2.2 Saturation

The fluid distribution in immiscible multiphase flow in porous media is described by the saturation S_α $[-]$ that indicates the volumetric portion of void space within pores occupied by the fluid phase α. Therefore, S_α is always between 0 and 1, and the sum of saturations S_α of all fluids present in the porous media is 1, i.e., $\sum_\alpha S_\alpha = 1$.

Since not all volume of the fluid phase can be displaced in multiphase flow from a porous medium due to hysteretic effects, the α-phase residual saturation quantity $S_{r\alpha}$ $[-]$ is introduced. It expresses the minimal saturation of the phase α that will retain in the porous medium due to adhesion effects with respect to the solid matrix. Therefore, the effective saturation S^e_α $[-]$ that describes only volumetric portions of displaceable fluid phases is introduced as

$$S^e_\alpha = \frac{S_\alpha - S_{r\alpha}}{1 - \sum_\beta S_{r\beta}}. \tag{1}$$

2.3 Capillary pressure

Following the standard definitions in literature, the capillary pressure p_c $[ML^{-1}]$ on the pore scale is defined as the difference between the non-wetting phase pressure p_n $[ML^{-1}]$ and the wetting phase pressure p_w $[ML^{-1}]$, i.e.,

$$p_c = p_n - p_w. \tag{2}$$

This definition is then averaged over a representative elementary volume (REV), see [1], [16], and thus holds for both pore and macroscopic scale relationship for modelling capillarity phenomena.

The capillary pressure function has been commonly considered as a function of wetting phase saturations and so it has been widely used in model equations in literature, see for instance [21], [11], [9], or [10]. The following Brooks and Corey (3) explicit capillary pressure - effective wetting phase saturation parametrization has been used in two-phase flow models and is given by [1]

$$p_c^{eq}(S_w^e) = p_d(S_w^e)^{-\frac{1}{\lambda}}, \tag{3}$$

where p_d $[ML^{-1}]$ is the entry pressure and λ $[-]$ describes pore distribution of the grains in porous material. The Brooks and Corey relationship (3) is suitable for modelling heterogeneous porous media because the difference in the entry pressure coefficients p_d in different porous materials preserves the barrier effect that has been observed in experiments, for details see [21], [16], [1], [8]. As the main objective of the ongoing research is to study dynamic effects of capillarity in heterogeneous porous media, other capillary pressure - saturation models like that by van Genuchten [25] which does not involve barrier effect will not be considered in this manuscript.

2.4 Dynamic capillary pressure

The classical capillary pressure - saturation relationships such as (3) has been used in almost all mathematical studies on porous media flow modelling in the past decades. Recently, theoretical studies [15], [14], [5], [12], [13], [3], [23], as well as the empirical approach in [24] have produced new aspects in the two-phase flow theories. The most important result is that the classical capillary pressure - saturation relationships hold only in the state of thermodynamic equilibrium. Therefore, the classical approach cannot be used in the modelling of capillarity when the fluid content is in motion and, consequently, a new capillary pressure - saturation relationship is proposed in the following form:

$$p_n - p_w = p_c^{eq}(S_w^e) - \tau(S_w)\frac{\partial S_w}{\partial t} =: p_c(S_w, \partial_t S_w), \tag{4}$$

where p_c^{eq} is the capillary pressure - saturation relationship in equilibrium and τ $[ML^{-1}T^{-1}]$ is the dynamic effect coefficient. In (4), the partial derivative of S_w after t is shortly denoted as $\partial_t S_w$ in the second argument of the capillary pressure function p_c.

Various researchers have developed formulae that are similar to (4). The characteristic dynamic effect quantity τ is regarded as a measure for the distance of the system from equilibrium [17].

Early in 1978, Stauffer [24], (or see [17], [18], [23]), proposed a linear dependence in (4) and proposed the following definition of τ:

$$\tau_S = \frac{\alpha_S \mu_w \Phi}{K\lambda}\left(\frac{p_d}{\rho_w g}\right)^2, \tag{5}$$

where $\alpha_S = 0.1$ $[-]$ denotes the scaling parameter, μ_w $[ML^{-1}T^{-1}]$ is the wetting phase dynamic viscosity, Φ $[-]$ is the porosity of the material, K $[L^2]$ is the intrinsic permeability, ρ_w $[ML^{-3}]$ is the wetting phase density and g $[LT^{-2}]$ is the gravitational acceleration

[1] A superscript ^{eq} is used in the definition (3) with respect to latter and it indicates the capillary pressure - saturation relationship model in equilibrium.

constant. Both λ and p_d are the Brooks and Corey parameters from relationship (3). Thus, the coefficient τ_S can be calculated for a given porous medium and wetting fluid.

The Stauffer model for the dynamic effect coefficient τ is obtained by correlating experimental data. The values of τ_S vary between $\tau_S = 2.7{\cdot}10^4\ Pas$ and $\tau_S = 7.7{\cdot}10^4\ Pas$, see [17, page 27], but the value of τ_S for sand parameters used in this manuscript is $\tau_S = 1.88 \cdot 10^5\ Pa\,s$. However, other researchers suggest that the magnitude of τ should be in the order of $10^2 - 10^3\ Pa\,s$, [4], or, on the other hand, it should be also in the order of $10^4 - 10^8\ Pa\,s$ as estimated in [14]. Moreover, some authors assume a general nonlinear dependence

$$\tau = \tau(S_w), \tag{6}$$

where the explicit dependence remains an open problem and thus only constant values of τ are studied in next sections.

3 Mathematical model

3.1 Governing equations

A mathematical model describing the two-phase flow in a onedimensional domain is presented in this section in order to demonstrate how the two-phase flow in porous medium is affected by the introduction of the dynamic capillary pressure relationship (4) instead of the classical relationships in thermodynamic equilibrium.

The governing two-phase flow equations in onedimensional domain $[0, L]$ are given by the $p_w - S_n$ formulation [1]

$$\Phi \frac{\partial S_n}{\partial t} = -\frac{\partial}{\partial x} u_n, \tag{7}$$

$$u_n = -\frac{K}{\mu_n} k_{rn} \left(\frac{\partial}{\partial x}(p_w + p_c) - \rho_n\, g \right), \tag{8}$$

$$-\Phi \frac{\partial S_n}{\partial t} = -\frac{\partial}{\partial x} u_w, \tag{9}$$

$$u_w = -\frac{K}{\mu_w} k_{rw} \left(\frac{\partial p_w}{\partial x} - \rho_w\, g \right) \tag{10}$$

where $k_{r\alpha} = k_{r\alpha}(S_n)$ $[-]$ is the α-phase relative permeability function and u_α $[LT^{-1}]$ is the α-phase Darcy velocity, for details see [8], [16], [1], or [2]. Initial and boundary conditions for equations (7-10) are given separatedly for each experimental problem.

3.2 Discrete problem

A standard finite difference discretization technique is used in order to determine approximate discrete solution $S_{n,i}^k$, $p_{w,i}^k$ of the problem (7-10), generally defined as $f_i^k = f(k\Delta t, i\Delta x)$, where $i = 0, 1, \ldots, m$, $m\Delta x = L$, and $k = 0, 1, \ldots$.

Since the nonlinear problem (7-10) involves the dynamic capillary pressure function defined in (4) that includes time derivative of S_n, an implicit numerical scheme is proposed

in the following form:

$$\Phi \frac{S_{n,i}^{k+1} - S_{n,i}^{k}}{\Delta t} = -\frac{u_{n,i,i+1}^{k,k+1} - u_{n,i-1,i}^{k,k+1}}{\Delta x}, \tag{11}$$

$$-\Phi \frac{S_{n,i}^{k+1} - S_{n,i}^{k}}{\Delta t} = -\frac{u_{w,i,i+1}^{k+1} - u_{w,i-1,i}^{k+1}}{\Delta x}, \tag{12}$$

$$\tag{13}$$

where the upwinded discrete Darcy velocities are given as

$$u_{n,i,i+1}^{k,k+1} = -\frac{K}{\mu_n} k_{rn}(S_{n,*}^{k+1}) \Big(\underbrace{\frac{p_{w,i+1}^{k+1} - p_{w,i}^{k+1}}{\Delta x} + \frac{p_{c,i+1}^{k,k+1} - p_{c,i}^{k,k+1}}{\Delta x} - \rho_n \, g}_{\mathrm{dir}_n} \Big) \tag{14}$$

$$S_{n,*}^{k+1} = \begin{cases} S_{n,i+1}^{k+1} & \text{if } \mathrm{dir}_n > 0. \\ \\ S_{n,i}^{k+1} & \text{if } \mathrm{dir}_n < 0. \end{cases} \quad \text{upwinded saturation } S_{n,*}$$

$$p_{c,i}^{k,k+1} = p_c \left(1 - S_{n,i}^{k+1}, -\frac{S_{n,i}^{k+1} - S_{n,i}^{k}}{\Delta t} \right).$$

and

$$u_{w,i,i+1}^{k+1} = -\frac{K}{\mu_w} k_{rw}(S_{n,\#}^{k+1}) \Big(\underbrace{\frac{p_{w,i+1}^{k+1} - p_{w,i}^{k+1}}{\Delta x} - \rho_w \, g}_{\mathrm{dir}_w} \Big) \tag{15}$$

$$S_{n,\#}^{k+1} = \begin{cases} S_{n,i+1}^{k+1} & \text{if } \mathrm{dir}_w > 0. \\ \\ S_{n,i}^{k+1} & \text{if } \mathrm{dir}_w < 0. \end{cases} \quad \text{upwinded saturation } S_{n,\#}$$

The numerical scheme is solved using the Newton iteration method for a system of nonlinear equations. The Jacobi matrix used in the Newton iteration method is sparse and can be reordered to a penta-diagonal matrix. It is therefore inverted using the *LU* decomposition algorithm for multi-diagonal matrices.

4 Numerical experiments

4.1 Comparison to semianalytical solution

Firstly, the numerical scheme (11-12) was compared to the semianalytical solution for onedimensional advection-diffusion problem, see [19], [26], [20], [10], [22], [9], [11], [8], or [7].

The semianalytical solution can be computed for onedimensional problem (7-10) without gravity (i.e., $g = 0$), including only classical capillary pressure - saturation relationship

(3), and with a special initial and boundary conditions in the form

$$S_n(0,x) = 0 \qquad \text{for all } x > 0, \tag{16}$$
$$p_w(0,x) = p_{atm} \qquad \text{for all } x > 0, \tag{17}$$
$$u_w(t,0) = 0 \qquad \text{for all } t > 0, \tag{18}$$
$$p_w(t,0) = p_{atm} \qquad \text{for all } t > 0, \tag{19}$$
$$S_n(t,L) = 0 \qquad \text{for all } t > 0, \tag{20}$$
$$u_w(t,L) = At^{-\frac{1}{2}} \qquad \text{for all } t > 0, \tag{21}$$

where constant A $[LT^{-\frac{1}{2}}]$ is the parameter of the semianalytical solution, see [10], and characterizes the input flux magnitude with value $A = 9.321 \cdot 10^{-3}\ ms^{-\frac{1}{2}}$ for this numerical experiment. Since the pressure is differentiated in the problem equations (7-10), the solution does not depend on the prescribed value of p_{atm} and so $p_{atm} = 0$ is used for all numerical experiments .

Although the semianalytical solution gives a good quantitative comparison with the solution obtained by the numerical scheme, the qualitative study (i.e., computation of the experimental order of convergence) is impossible due to singularity of the boundary flux (21) at $t = 0$. The results obtained for the sand and fluid properties in Table 1 are shown in Figure 1 at time $t = 10,000\ s$.

	Symbol	Units	Value
Porosity	Φ	$[-]$	0.42
Intrinsic Permeability	K	$[m^2]$	$2.73 \cdot 10^{-11}$
Residual Water Sat.	S_{wr}	$[-]$	0
Brooks-Corey	p_d	$[Pa]$	3433.5
parameters	λ	$[-]$	2
Water viscosity	μ_w	$[kg\ m^{-1}s^{-1}]$	10^{-3}
Air viscosity	μ_n	$[kg\ m^{-1}s^{-1}]$	$1.83 \cdot 10^{-4}$

Table 1: Parameter setup for the sand and fluid properties.

4.2 Numerical experiments with dynamic effect

Assuming the previously presented reliability of the quantitative comparison of the numerical scheme solution to the semianalytical solution, a series of tests were proceeded in order to determine influence of various values of the dynamic effect coefficient τ on the solution.

A onedimensional vertically positioned column filled with homogeneous sand is initially fully water saturated. At $t = 0$ a pressure head in the lower end of the column ($x = L$) begins to decrease and the air enters the column from the top ($x = 0$). The pressure drop is characterised by constant wetting phase flux $u_w(t,L) = A$. Together,

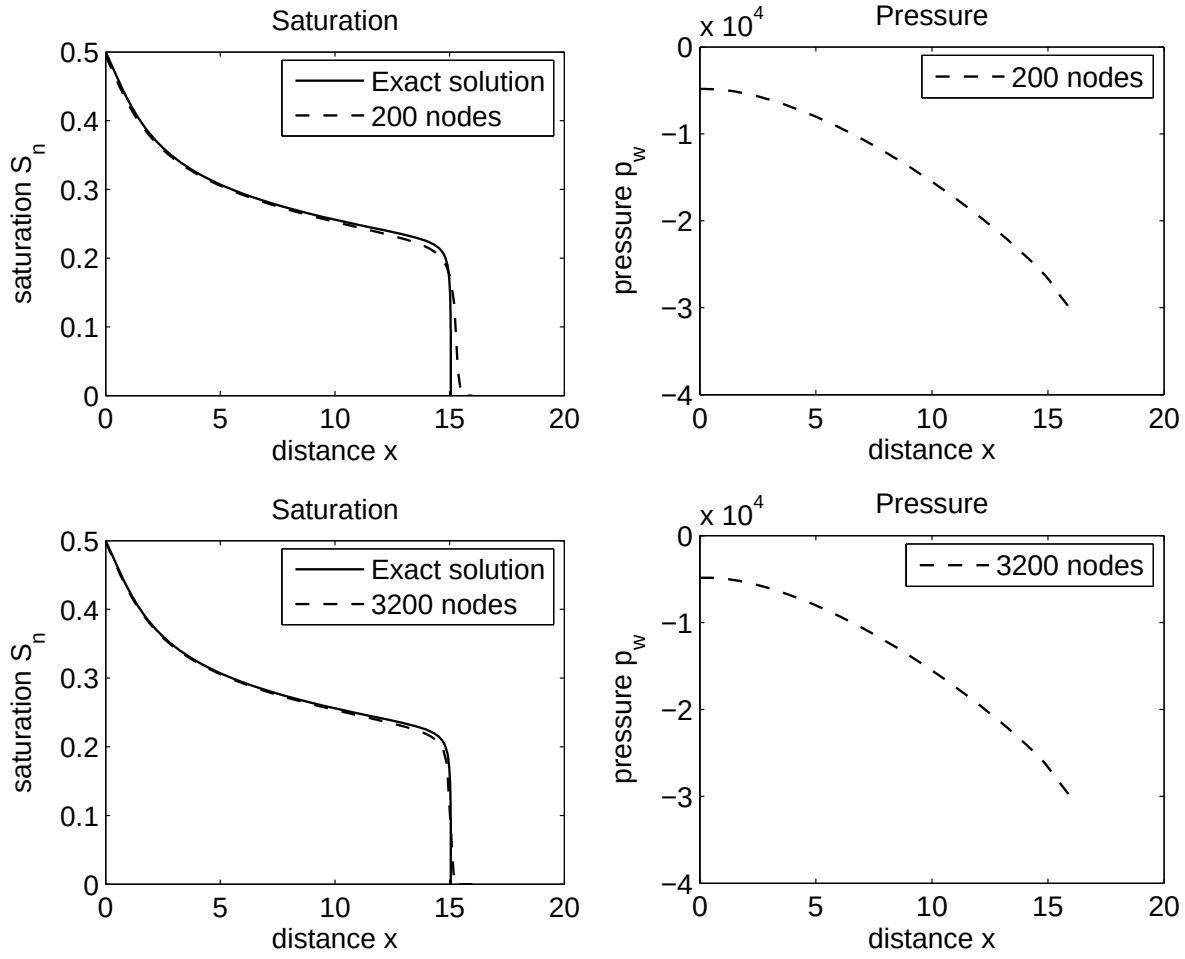


Figure 1: Semianalytical solution compared to the numerical solution obtained by the scheme (11-12).

the initial and boundary conditions are given as follows:

$$S_n(0,x) = 0 \qquad \text{for all } x > 0, \tag{22}$$

$$p_w(0,x) = p_{atm} - p_c^{eq}(1) + \rho_w g x \qquad \text{for all } x > 0, \tag{23}$$

$$S_n(t,0) = 0 \qquad \text{for all } t > 0, \tag{24}$$

$$p_w(t,0) = p_{atm} - p_c(1 - S_n(t,0), -\partial_t S_n(t,0)) \qquad \text{for all } t > 0, \tag{25}$$

$$S_n(t,L) = 0 \qquad \text{for all } t > 0, \tag{26}$$

$$u_w(t,L) = A \qquad \text{for all } t > 0, \tag{27}$$

Numerical simulations were done using the fluid and material properties shown in Table 1 with $A = 5 \cdot 10^{-5}\ ms^{-1}$. The final time of the simulations is $t = 10,000s$.

Reference solutions are obtained using only classical capillary pressure - saturation relationship in the equilibrium and are shown in Figure 2. Next, Figure 3 shows solution obtained for the Stauffer coefficient τ_S. In order to examine the situations for higher orders of magnitude of τ, two simulations for $\tau = 10^7\ kg\,m^{-1}s^{-1}$ and $\tau = 10^8\ kg\,m^{-1}s^{-1}$

were carried out and are depicted in Figure 4 and 5, respectively.

The results indicate that the Stauffer model for the parameter τ_S does not significantly changes the pressure/saturation profiles with respect to the reference state. However, the increase in orders of magnitude of the dynamic effect coefficient τ is more important as it is obvious in Figures 4 and 5

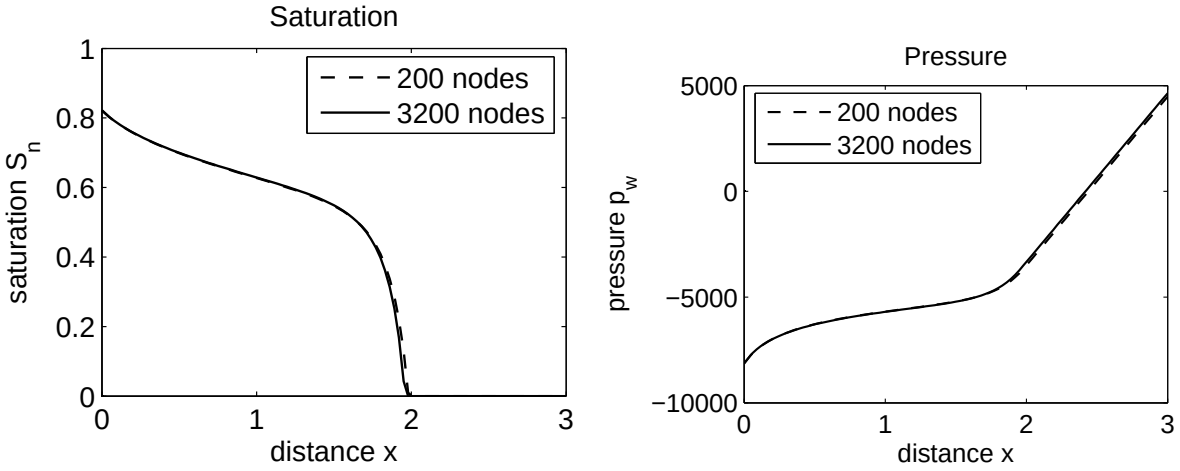


Figure 2: Numerical solutions with no dynamic effect (reference solution), i.e., $\tau = 0\ Pa\,s$.

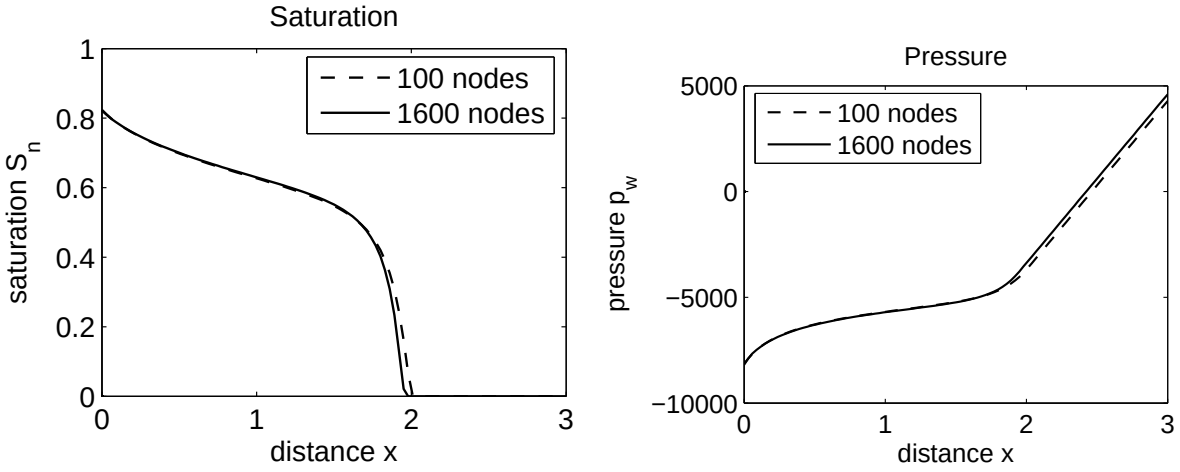


Figure 3: Numerical solutions using Stauffer dynamic effect coefficient τ_S. For the given problem (Table 1) its value is $\tau_S = 1.88 \cdot 10^5\ Pa\,s$.

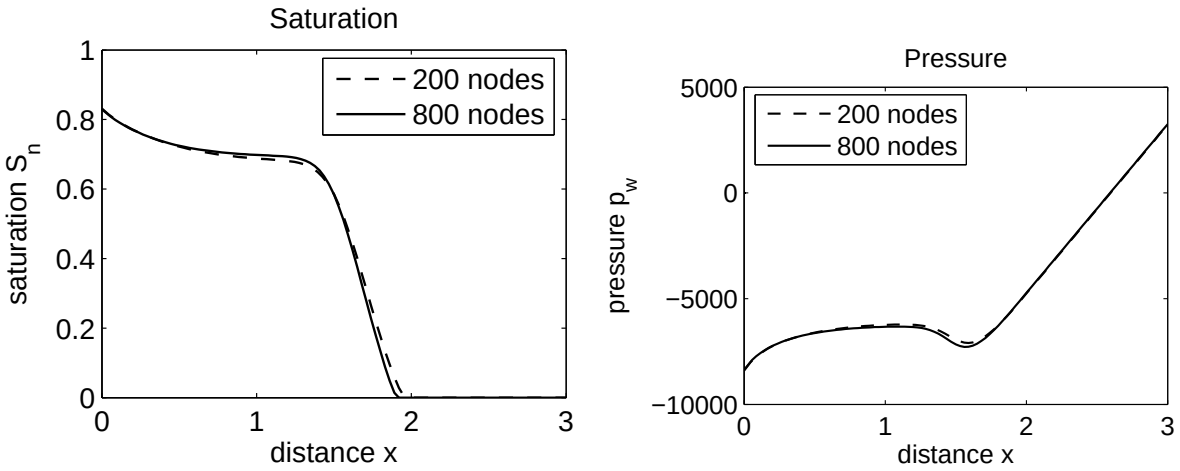


Figure 4: Numerical solutions obtained for $\tau = 10^7\ Pa\,s.$).

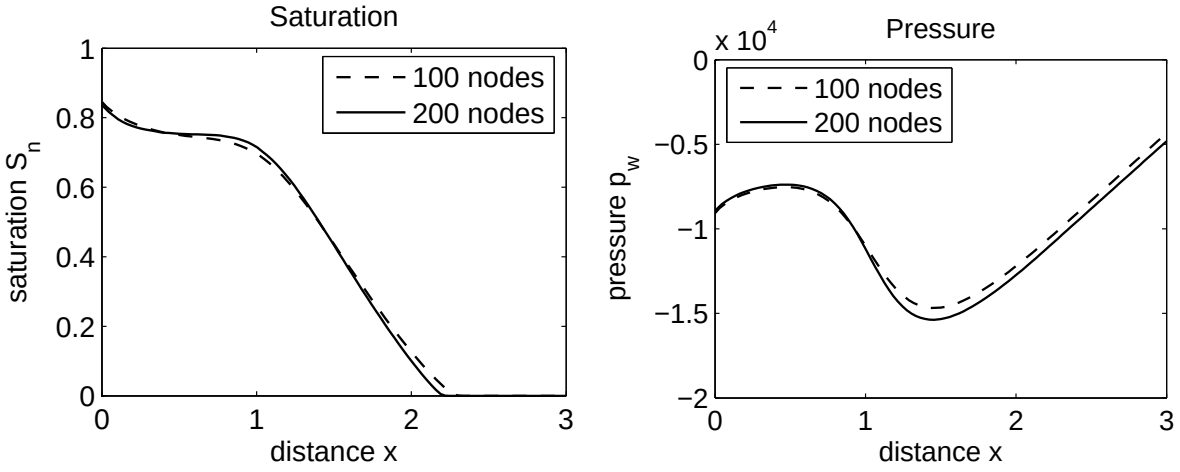


Figure 5: Numerical solutions obtained for $\tau = 10^8\ Pa\,s.$).

5 Conclusion and future work

This manuscript presents recently obtained numerical simulations using the nonclassical dynamic capillary pressure in simulating two-phase incompressible flow in porous medium. The presented results indicate the numerical scheme is applicable for the given problem and allows for experimental implementation of two-phase flow in heterogeneous porous medium.

Consideration of material heterogeneity involves various possibilities how the interfacial conditions between different porous materials can be treated. The investigation of all possible implementations is going to be the main goal of the author's future work since the main objective of the project is to develop robust and stable implementation of interfacial conditions for two and three dimensional codes including the dynamic capillary pressure - saturation relationships also across the material interface.

Acknowledgement

This work has been partly supported by the project "Applied Mathematics in Technical and Physical Sciences" MSM 6840770010, by the project "Environmental modelling" KONTAKT ME878, and by the project "Jindřich-Nečas Center for Mathematical Modelling" LC06052, all of the Ministry of Education of the Czech Republic, and by the National Science Foundation through the award 0222286 (CMG RESEARCH: "Numerical and Experimental Validation of Stochastic Upscaling for Subsurface Contamination Problems Involving Multiphase Volatile Chlorinated Solvents").

The author would gratefully thank Toshihiro Sakaki [2] for providing experimental data for the numerical simulations.

References

[1] P. Bastian. *Numerical Computation of Multiphase Flows in Porous Media.* Christian-Albrechts-Universitat Kiel, (1999).

[2] J. Bear and A. Verruijt. *Modeling Groundwater Flow and Pollution.* D. Reidel, Holland, Dordrecht, (1990).

[3] A. Y. Beliaev and S. M. Hassanizadeh. *A theoretical model of hysteresis and dynamic effects in the capillary relation for two-phase flow in porous media.* Transport in Porous Media **43** (2001), 487–510.

[4] H. K. Dahle, M. A. Celia, and S. M. Hassanizadeh. *Bundle-of-tubes model for calculating dynamic effects in the capillary-pressure-saturation relationship.* Transport in Porous Media **58(5-22)** (2005), DOI 10.1007/s1242–004–5466–4.

[5] D. B. Das, S. M. Hassanizadeh, B. E. Rotter, and B. Ataie-Ashtiani. *A numerical study of micro-heterogeneity effects on upscaled properties of two-phase flow in porous media.* Transport in Porous Media **56** (2004), 329–350.

[2] CESEP, Colorado School of Mines, Golden CO, cesep.mines.edu

[6] R. Fučík. *Numerical solution of multiphase porous media flow, Research work.* FNSPE of Czech Technical University Prague, Prague, (2004).

[7] R. Fučík. *Numerical analysis of multiphase porous media flow, Research project.* FNSPE of Czech Technical University Prague, Prague, (2005).

[8] R. Fučík. *Numerical Analysis of Multiphase Porous Media Flow in Groundwater Contamination Problems , Graduate Thesis.* FNSPE of Czech Technical University Prague, Prague, (2006).

[9] R. Fučík, M. Beneš, J. Mikyška, and T. H. Illangasekare. Generalization of the benchmark solution for the two-phase porous-media flow. In 'Finite Elements Models, MODFLOW, and More : Solving Groundwater problems', 181–184, (2004).

[10] R. Fučík, J. Mikyška, M. Beneš, and T. H. Illangasekare. *An improved semi-analytical solution for verification of numerical models of two-phase flow in porous media.* Vadoze Zone Journal **6** (2007), 93–104.

[11] R. Fučík, J. Mikyška, and T. H. Illangasekare. *Evaluation of saturation-dependent flux on two-phase flow using generalized semi-analytic solution.* Proceedings on the Czech Japanese Seminar in Applied Mathematics FNSPE CVUT Prague (2004), 25–37.

[12] W. G. Gray and S. M. Hassanizadeh. *Paradoxes and realities in unsaturated flow theory.* Water Resources Research **27(8)** (1991), 1847–1854.

[13] W. G. Gray and S. M. Hassanizadeh. *Unsaturated flow theory including interfacial phenomena.* Water Resources Research **27(8)** (1991), 1855–1863.

[14] S. M. Hassanizadeh, M. A. Celia, and H. K. Dahle. *Dynamic effect in the capillary pressure-saturation relationship and its impacts on unsaturated flow.* Vadoze Zone Journal **1** (2002), 38–57.

[15] S. M. Hassanizadeh and W. G. Gray. *Thermodynamic basis of capillary pressure in porous media.* Water Resources Research **29(10)** (1993), 3389–3405.

[16] R. Helmig. *Multiphase Flow and Transport Processes in the Subsurface : A Contribution to the Modeling of Hydrosystems.* Springer Verlag, Berlin, (1997).

[17] S. Manthey. *Two-phase flow processes with dynamic effects in porous media - parameter estimation and simulation.* Institut fur Wasserbau der Universitat Stuttgart, Stutgart, (2006).

[18] S. Manthey, S. M. Hassanizadeh, and R. Helmig. *Macro-scale dynamic effects in homogeneous and heterogeneous porous media.* Transport in Porous Media **58** (2005), 121–145.

[19] D. B. McWhorter and D. K. Sunada. *Exact integral solutions for two-phase flow.* Water Resources Research **26** (1990), 399–413.

[20] D. B. McWhorter and D. K. Sunada. *Reply.* Water Resources Research **28** (1992), 1479.

[21] J. Mikyška. *Numerical Model for Simulation of Behaviour of Non-Aqueous Phase Liquids in Heterogeneous Porous Media Containing Sharp Texture Transitions, Ph.D. Thesis.* FNSPE of Czech Technical University, Prague, (2005).

[22] J. Mikyška and T. H. Illangasekare. Analytical and numerical solution for one-dimensional two-phase flow in homogeneous porous medium. In 'submitted to of the 2nd International Conference on Porous Media and its Applications in Science and Engineering, June 17-21 2007', Kauai, Hawaii, USA, (2007).

[23] O. Oung, S. M. Hassanizadeh, and A. Bezuijen. *Two-phase flow experiments in a geocentrifuge and the significance of dynamic capillary pressure effect.* Journal of Porous Media **8(3)** (2006), 247–257.

[24] F. Stauffer. Time dependence of the relations between capillary pressure, water content and conductivity during drainage of porous media. In 'On scale effects in porous media, IAHR, Thessaloniki, Greece', (1978).

[25] M. T. van Genuchten. *A closed-form equation for predicting the hydraulic conductivity of unsaturated soils.* Soil Science Society of America Journal **44** (1980), 892–898.

[26] G. S. B. Z.-X. Chen and P. A. Witherspoon. *Comment on 'exact integral solutions for two-phase flow'.* Water Resources Research **28** (1992), 1477–1478.

Syntéza periodicko-stochastických textur

Martin Hatka

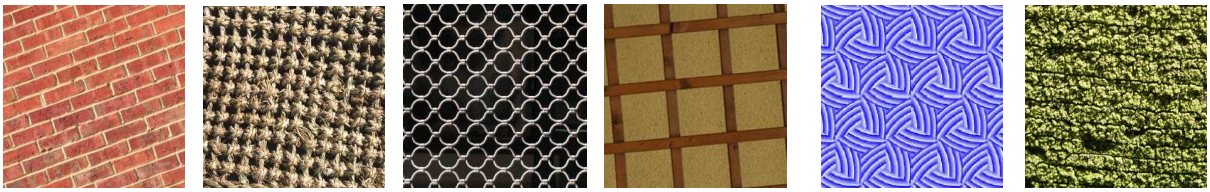
2. ročník PGS, email: hatka@utia.cas.cz
Katedra matematiky, Fakulta jaderná a fyzikálně inženýrská, ČVUT
školitel: Michal Haindl, Ústav teorie informace a automatizace, AV ČR

Abstract. This paper describes two methods for seamless enlargement of difficult colour textures containing both regular periodic and stochastic components. Such textures cannot be modelled using neither simple tiling nor using purely stochastic models. The first novel method automatically recognizes and separates periodic and random texture components. Each of this components is subsequently modelled using the coresponding optimal method. Both independently enlarged texture components are combined in the resulting synthetic near regular texture. The second method detects two main direction of periodicity in regular periodic component and generates several double-toroidal tiles of the same general shape, which can seamlesly enlarge given near-periodic texture without visible regularity. While the presented texture synthesis methods allow only moderate compression, they are extremely fast due to complete separation of the analytical step of the algorithm from the texture synthesis part. The methods are universal and easily viable in a graphical hardware for purpose of real-time rendering of any type of near regular static textures.

Abstrakt. Tento článek popisuje dvě metody pro syntézu složitých barevných textur obsahujících periodickou i stochastickou složku. Takové textury nelze modelovat buď jen jednoduchým dlaždicováním nebo čistě stochastickými modely. První nová metoda automaticky rozpoznává a odděluje periodickou a stochastickou složku textury. Každá z těchto komponent je následně modelována pomocí odpovídajících optimálních metod. Obě nezávisle syntetizované komponenty jsou zkombinovány, čímž vznikne výsledná syntetická periodicko-stochastická textura. Druhá metoda detekuje dva hlavní směry periodicity pravidelné periodické komponenty a generuje několik toroidních dlaždic stejného obecného tvaru. Pomocí těchto dlaždic lze generovat periodicko-stochastickou texturu bez viditelných rušivých pravidelností. Ačkoliv prezentované metody pro syntézu textur umožňují poze částečnou kompresi, jsou extrémně rychlé, a to díky completně oddělené fázi analýzy od fáze syntézy algoritmu. Metody jsou univerzální a snadno implementovatelné v grafickém hardwaru, zejména za účelem renderingu libovolného typu periodicko-stochastických statických textur v reálném čase.

1 Úvod

Mnoho aplikací v počítačové grafice, v počítačovém vidění a při zpracování obrazu vyžaduje textury libovolných rozměrů. Pro tyto aplikace je syntéza textur velmi důležitá, protože je alternativním a ve většině případů i jediným možným způsobem jejich vytváření. Cílem syntézy textur je vytvoření nové textury z daného texturního vzorku, přičemž nová textura by měla realisticky odpovídat textuře původní. Vzhled obou textur by měl být takový, jako by byly vytvořeny na základě stejného generujícího procesu. Problémem je ale navržení algoritmu, který bude efektivní a zároveň bude dosahovat kvalitních výsledků.



Obrázek 1: Několik příkladů realistických periodicko-stochastických textur, obrázky jsou o velikosti 256×256 obrazových bodů.

Obrázek 2: Vlevo ukázka periodické textury, vpravo stochastické.

V závislosti na další aplikaci lze klást důraz jak na určiré kvalitativní vlastnosti textury, tak i na vlastnosti algoritmů pro jejich generování. Algoritmy pro modelování textur mají přirozeně část analýzy, ve které se odhadují parametry potřebné pro další syntézu, a část syntézy, ve které se generuje nová textura. Tyto dvě části nejsou často odděleny. Aby byly metody využitelné v praktických aplikacích, bývá častým požadavkem jejich rychlost, zejména pak rychlost syntézy.

Tento článek je zaměřen na modelování periodicko-stochastických textur (obr. 1, obr. 2). Periodicko-stochastická textura obsahuje globální pravidelné struktury, které představují zásadní problém pro vzorkovací metody, a nepravidelné stochastické struktury, které nemohou být věrohodně reprodukovány pouhým dlaždicováním. Pokud se omezíme pouze na periodicko-stochastické textury, pak lze navrhnout specializovanou metodu pro jejich syntézu. V tomto článku budou prezentovány dvě příbuzné metody, jež lze pro syntézu periodicko-stochastických textur využít.

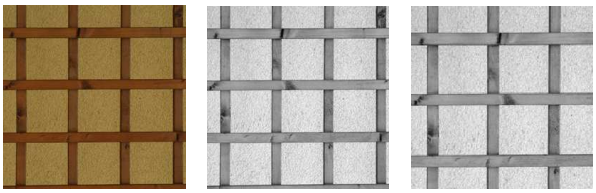
První z nich původně vychází ze záměru věnovat se editaci textur. Další motivací je možnost výrazné komprese texturních dat, zejména pak u stochastické složky. Základním prvkem metody je oddělení periodické struktury od zbývající, zpravidla stochastické části. Periodickou strukturu lze pak rozšiřovat vhodným vzorkovacím algoritmem a stochastickou část lze vcelku úspěšně modelovat pomocí některého adaptivního modelu. Vezmeme-li v úvahu více takových textur, můžeme jejich periodické a stochastické části libovolně zaměnovat a generovat textury zcela nové.

Druhá metoda vznikla částečně z nedostatků původní metody. Umožňuje generovat periodicko-stochastické textury, jejichž periodické struktury jsou libovolně natočeny a dosažené výsledky jsou na tomto natočení nezávislé. S využitím faktu, že se ve vzorku vyskytuje jistá periodicita, dokážeme extrahovat několik minimálních dlaždic, které jsou si velmi podobné, a jejich hrany optimalizovat tak, že do sebe všechny vzájemně zapadají. Jejich kombinací dokážeme generovat výstupní texturu libovolných rozměrů, přičemž nedochází k umělému vnášení periodicity do stochastických struktur textury.

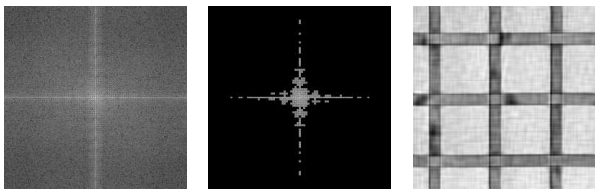
2 Modelování periodicko-stochastických textur

2.1 Separace periodické a neperiodické části

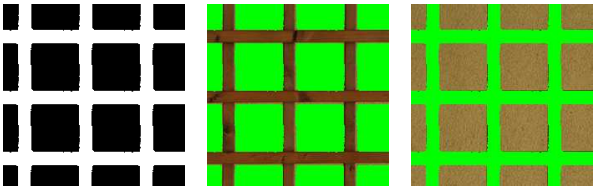
Periodická a neperiodická část textury je detekována ve zjednodušené monospektrální textuře získané pomocí Karhunen-Loevovy transformace vstupní barevné textury. Nově vzniklá spektrální pásma jsou vzájemně nekorelovaná, pro další zpracování se uvažuje



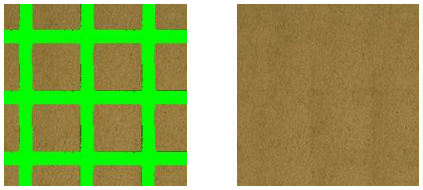
Obrázek 3: Vstupní textura (vlevo) je pomocí Karhunen-Loevovy transformace převedena na monochromatickou (uprostřed) a najde se největší periodický výřez (vpravo), který se navíc převzorkuje.



Obrázek 4: V amplitudovém spektru (vlevo) se ponechají jen výrazné frekvence (uprostřed). Dále se provede zpětná Fourierova transformace (vpravo).



Obrázek 5: Po filtraci a binarizaci s přihlédnutím k periodicitě (vlevo) je oddělena periodická složka (uprostřed) od stochastické (vpravo).



Obrázek 6: Pro potřeby učení adaptivních metod je třeba stochastickou část (vlevo) zrekonstuovat, aby byla bez neznámých pixelů (vpravo).

spektrální pásmo nesoucí maximum informace a odpovídající největšímu vlastnímu číslu (obr. 3).

Abychom získali dlaždici odpovídající největší periodické části textury, musí být známy periody podél horizontální a vertikální osy. Existuje několik metod pro nalezení period založených buď na různých statistických přístupech nebo na Fourierově transformaci (viz. diskuse v [1]). Pro tento účel byla použita suma čtvercových diferencí, tak jako v [1].

Po nalezení horizontální a vertikální periody se ze vstupní textury vyřízne největší možná periodická část. Výřez je pak převzorkován na rozměry odpovídající mocninám dvou (obr. 3), které jsou vyžadovány pro filtraci založené na rychlé Fourierově transformaci. Filtr ponechá pouze koeficienty, které jsou lokálními maximy v amplitudovém spektru, a koeficienty v jejich okolí. Navíc je na lokální maxima kladena podmínka, že musí být větší, než předem stanovený práh. Jako okolí lokálních maxim uvažujeme hierarchické okolí prvního nebo druhého řádu (obr. 4).

Po filtraci se provede zpětná Fourierova transformace, převzorkuje se zpět na původní rozměr a tím se dostane odfiltrovaný výřez. Tento výřez je pak binarizován pomocí určitého prahu. Pro nalezení korespondence k periodické a neperiodické části originální textury se binární obrázek testuje na posuny o velikosti horizontální a vertikální periody a rozhoduje se, zda určitý obrazový bod patří nebo nepatří do periodické struktury (obr. 5).

2.2 Syntéza

Po oddělení lze libovolnými způsoby provádět syntézu periodické a neperiodické části, a to nezávisle na sobě, různými metodami. Pro rozšiřování periodické části je vhodné využít metodu založenou na dlaždicování. V našem případě jsme využili dříve publikovanou metodu *Toroidní váleček* [4], kterou jsme navíc mohli zjednodušit díky předem známým velikostem period a tedy i předem známým rozměrům minimální toroidní dlaždice. Pro rozšiřování neperiodické části lze využít adaptivních metod založených na matematických modelech nebo lze využít i některou z vhodných vzorkovacích metod, v závislosti na charakteru neperiodické části.

Syntéza periodické části je jednoduchá, uvažujeme pouze její známé body. Ovšem problém nastává při syntéze stochastické části, pokud chceme použít některou z adaptivních metod. Adaptivní metody vyžadují pro fázi učení dostatečně velký vzorek vstupní textury, obvykle alespoň 256×256 bodů, ve kterém jsou všechny body známé. Po odfiltrování periodické struktury však nemáme zajištěno, že budeme mít k dispozici dostatečně velký vzorek pro učení.

Budeme-li předpokládat, že stochastická složka je dostatečně homogenní, tedy v místě periodické části by se nevyskytovaly výrazné nehomogenity, pak lze pro doplnění neznámých oblastí využít vhodnou vzorkovací metodu (obr. 6). Pro naše účely jsme vyvinuli metodu, jejíž základní myšlenka vychází z algoritmu *Image Quilting* [2]. Z důvodu zachování složitejších elementů v neperiodické části využíváme syntézy pomocí bloků textury. Pro optimalizaci hran bloků při vkládání využíváme řezů s minimální chybou. Pro syntézu využíváme pouze bloky, které neobsahují ani jeden neznámý pixel. Tento způsob se osvědčil, je dostatečný pro většinu textur a často lze úspěšně zrekonstuovat celou původní podkladovou texturu.

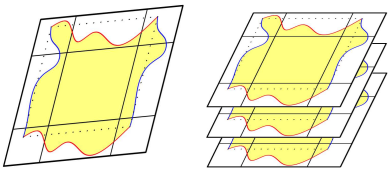
Po nezávislé syntéze obou částí se již jen vloží periodická část přes syntetickou texturu stochastické části. Tímto způsobem lze také kombinovat periodické a stochastické části libovolných textur.

2.3 Zhodnocení metody

Metoda funguje spolehlivě pro většinu periodických struktur obsažených ve vstupních texturách. Jednou z možných komplikací je požadavek, aby periodická sturktura byla periodická ve smyslu celého vzorku textury, a to z důvodu využívání Fourierovské filtrace. Pokud lze vstupní texturu oříznout tak, že maximální výřez je periodický ve vodorovném i svislém směru, nevzniká zde žádný další principielní problém. Ovšem pokud bude periodická část v textuře natočena o obecný úhel, nemáme již zajištěno, že lze nalézt maximální periodický výřez v obou směrech a nelze potom Fourierovskou filtrací periodickou část odfiltrovat.

Výsledek filtrace také dosti závisí na charakteru vstupní textury, zejména pokud je ve stochastické části přítomna nějaká významná frekvence. Pak se nemusí podařit požadované odfiltrování periodické části.

Obrázek 7: Vstupní textura (vlevo) a výseče korelačního pole pro hledání period ve dvou různých směrech. Vpravo pak vyříznutá minimální dlaždice.



Obrázek 8: Minimální dlaždice a více dlaždic se stejnými hranami.

3 Vzorkování periodicko-stochastických textur

V případě předchozí metody narážíme na problém, když je periodická struktura natočena. Protože předchozí metoda selhává zpravidla už při detekci periodicit v horizontálním nebo vertikálním směru, je třeba se poohlédnout po jiné, obecnější metodě, jak detekovat významné periodicity.

3.1 Detekce významných periodicit

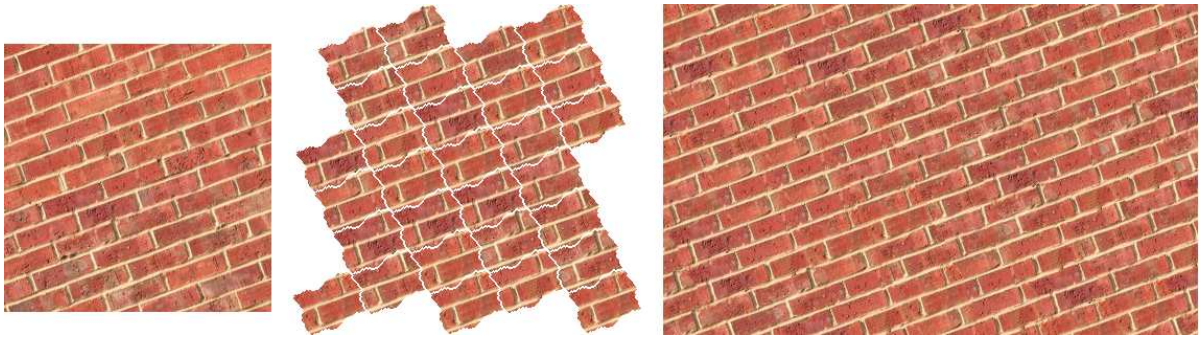
Je zřejmé, že pro hrybý odhad velikosti periody a směru periodicity lze snadno využít amplitudového spektra Fourierovy transformace. My však potřebujeme zjistit periodu ve dvou různých směrech přesně. Nalezneme je pomocí lokálních maxim v poli korelačních koeficientů pro různé posuvy. Abychom korelační pole nemuseli počítat celé, využijeme odhadu z Fourierovy transformace a budeme korelační koeficienty počítat jen na výsečích (obr. 7), které určíme na základě dvou maximálních koeficientů v amplitudovém spektru. Tyto dva maximální koeficienty musí určovat dva různé nerovnoběžné směry. Přesné periody a jejich směry najdeme na základě korelačního koeficientu počítaného pro různé posuvy vstupní textury.

3.2 Minimální toroidní dlaždice

Přesné periody nám současně určují velikost a směr hran minimální toroidní dlaždice, kterou budeme chtít získat ze vstupní textury. Pro hledání optimální minimální dlaždice a jejich optimálních hran zavádíme obecné horizontální a vertikální oblasti překryvu (obr. 8). Uvažujeme, že protejší okraje dlaždice se částečně překrývají a v těchto oblastech překryvu hledáme optimální řez s minimální chybou. Vhodný je algoritmus A^*, případně Dijkstrův algoritmus.

3.3 Několik toroidních dlaždic

Jelikož budeme chtít získat syntetickou texturu, která nebude vykazovat pravidelnost, můžeme hned vyříznout několik minimálních dlaždic a jejich hrany optimalizovat tak, aby do sebe protější hrany zapadaly. Jejích vzájemnou libovolnou kombinací pak můžeme generovat texturu požadovaných rozměrů. Hledání dalších vhodných dlaždic provádíme pomocí korelace. Vyříznutou minimální dlaždici včetně oblastí překryvu posouváme po



Obrázek 9: Princip syntézy textury. Vlevo vstupní textura, uprostřed princip skládání minimálních toroidních dlaždic generujících novou výstupní texturu a vpravo nová syntetická textura.

vstupní textuře a na základě korelačního koeficientu hledáme další vhodné dlaždice. Jelikož několik takto nalezených dlaždic je velmi podobných, hledáme stejný optimální řez pro všechny dlaždice najednou (obr. 8), čímž zajistíme jejich vzájemnou návaznost.

3.4 Syntéza

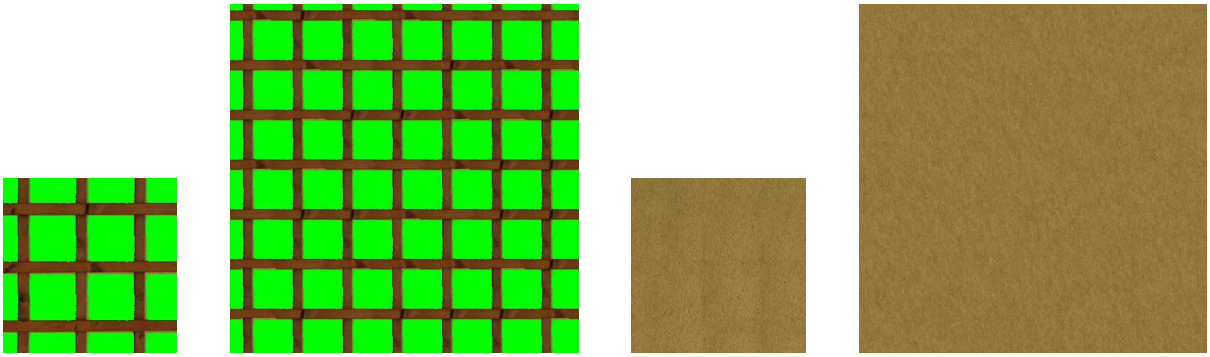
Syntéza textury je velmi jednoduchá a nevyžaduje žádné komplikované výpočty. Proces syntézy je pouhé vyplnění konečné oblasti několika vzájemně navazujícími toroidními dlaždicemi (obr. 9). Dlaždice jsou umísťovány postupně po krocích daných dvěma hlavními periodami. Po vyplnění konečné oblasti tímto způsobem získáme novou texturu libovolných rozměrů.

3.5 Zhodnocení metody

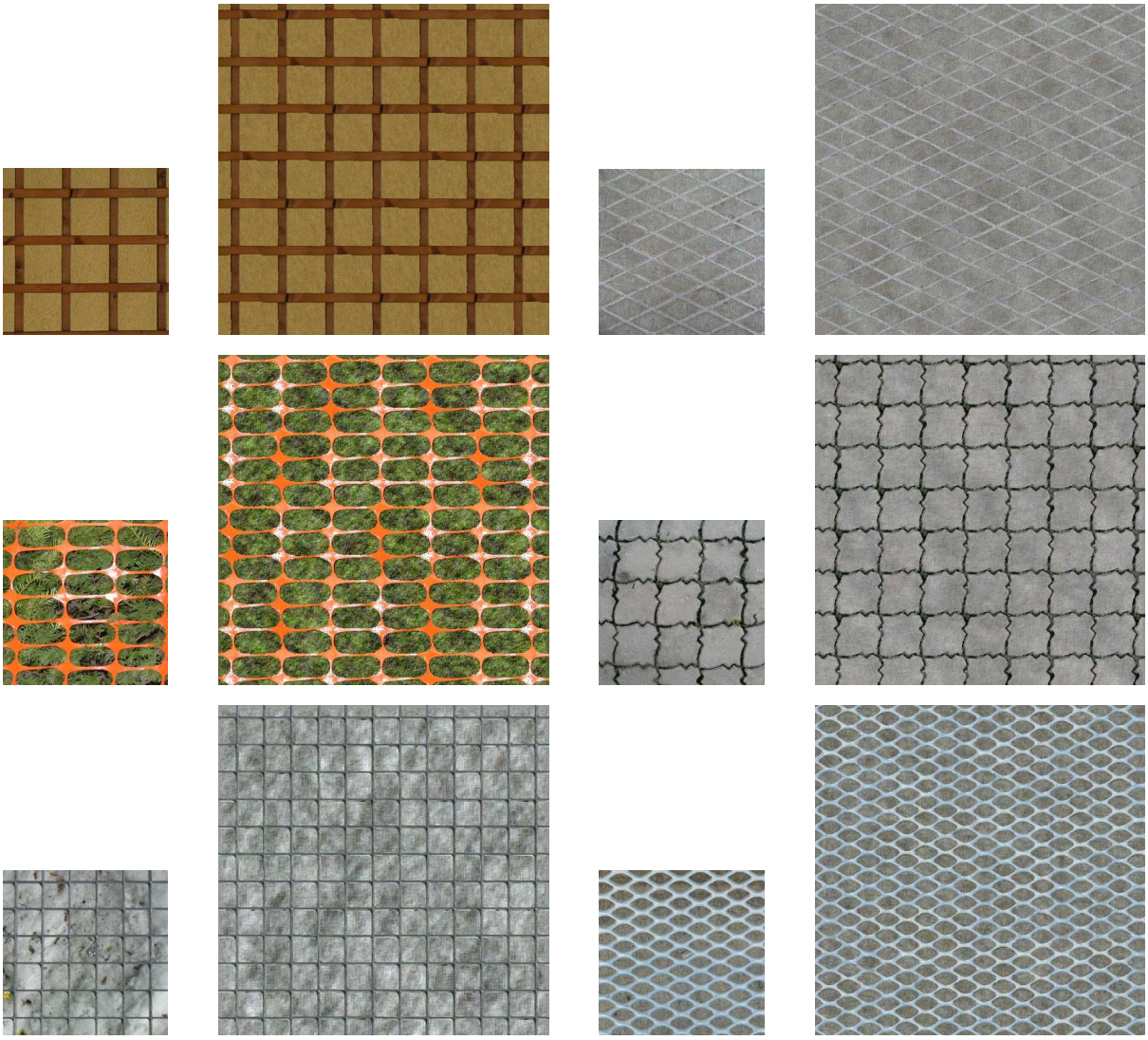
Tato metoda velmi úspěšně generuje syntetické textury z libovolného vzorku periodicko-stochastické textury. Značnou výhodou je nezávislost na natočení periodické části. V případě, že budeme vyžadovat syntetickou texturu otočenou jinak, než je vzorek, stačí otočit vzorek a pak teprve metodu použít. Není třeba otáčet až výstupní texturu, která je vždy mnohem větších rozměrů a její otáčení by tedy bylo časově náročnější.

Častým požadavkem je rychlá fáze syntézy, nejlépe v reálném čase. Analytická část slouží k odhadu nebo hledání parametrů, které se využijí při syntéze. Analýza je obvykle časově velmi náročná a nebývají na ni kladeny žádné zásadní časové nároky. Část syntézy by však měla být co nejrychlejší a v ideálním případě by měla být oddělena od analýzy, aby nebyla zatížena náročnými výpočty. Právě tyto požadavky metoda splňuje a umožňuje proto syntézu v reálném čase.

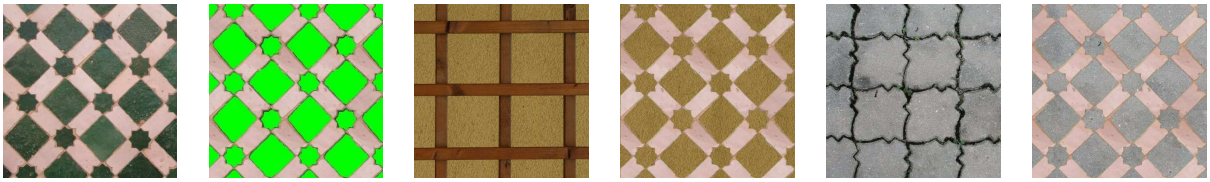
Jako nevýhoda se může jevit obecný tvar toroidních dlaždic, který může na první pohled komplikovat manipulaci s nimi. Z důvodu zachování vzájemné návaznosti více dlaždic však nelze hrany dále optimalizovat a je nutné zachovat jejich obecný charakter.



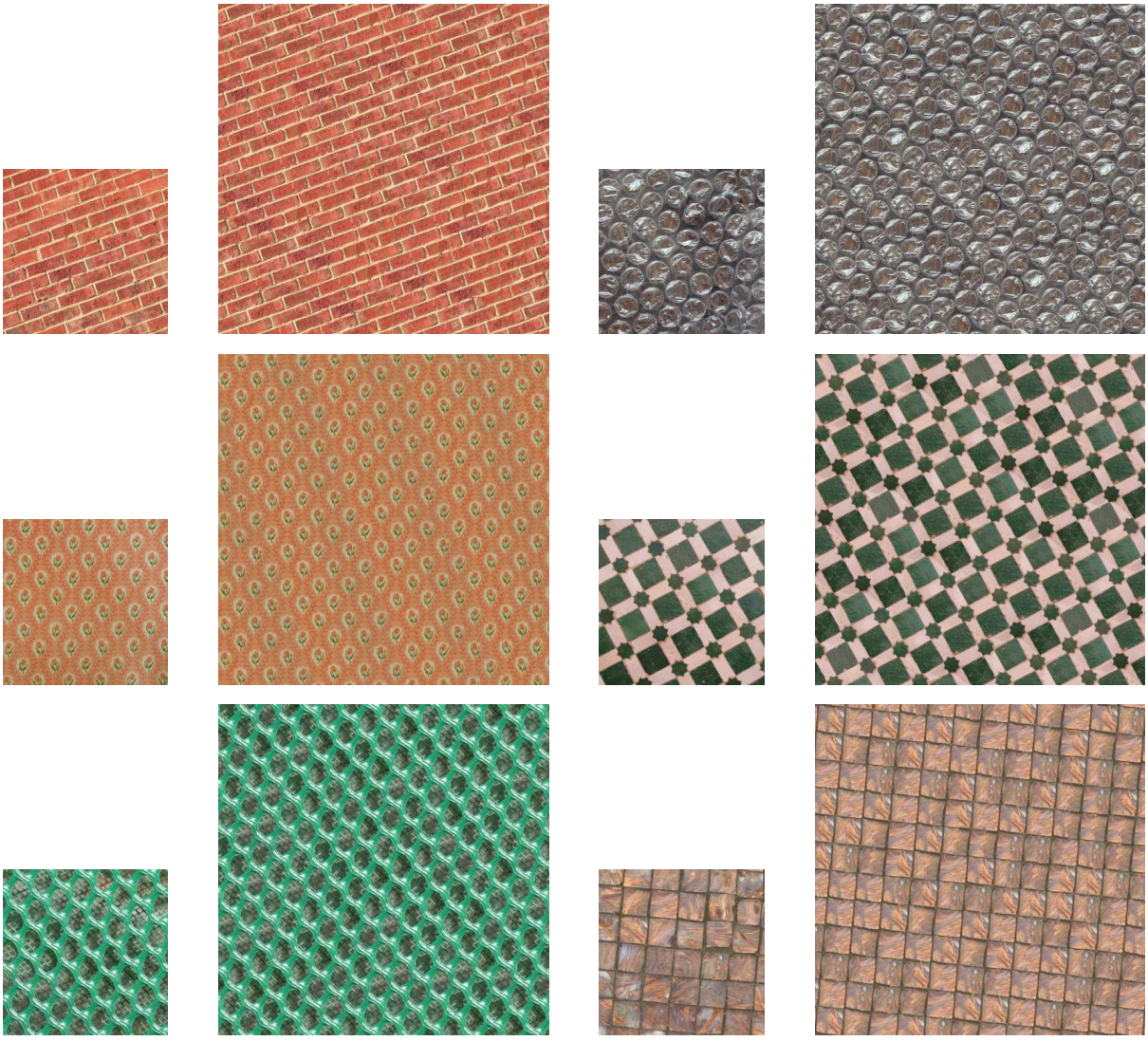
Obrázek 10: Syntéza jednotlivých složek periodicko-stochastické textury. Vlevo vzorkování periodické složky a vpravo modelování stochastické složky.



Obrázek 11: Výsledky syntézy periodicko-stochastických textur. Periodická část je rozšiřována dlaždicováním, stochastická část je modelována.



Obrázek 12: Ukázka editace textur. Pravidelná složka textury je kombinována se stochastickou složkou textury jiné.



Obrázek 13: Výsledky vzorkování periodicko-stochastických textur, pro syntézu nové textury bylo použito 4 až 8 dlaždic.

4 Výsledky

Obě metody dosahují velmi dobrých výsledků a byly testovány na širokém spektru periodicko-stochastických textur. Většina výsledných syntetických textur je vizuálně nerozlišitelná od vstupních vzorků, navíc lze provést syntézu v reálném čase a nezávisle na analýze, zatímco analýza je časově mnohem náročnější. Pro úplnost uveďme, že analýza vzorku o velikosti 256×256 nebo 512×512 bodů vyžaduje čas v řádu několika minut. Časy byly měřeny na PC Intel Core 2 Duo 2.4GHz s 2GB RAM.

Na obrázku 10 je ukázka odděleného zpracování periodické a stochastické složky textury. Periodická složka je vzorkována, v našem případě algoritmem *Toroidní váleček* [4], stochastická složka je modelována pomocí kauzálního autoregresního modelu [5]. Obrázek 11 prezentuje výsledky syntézy, periodická složka byla opět vzorkována a stochastická modelována. Na obrázku 12 je znázorněna alternativní možnost využití metody k editaci textur, kombinování periodické a stochastické složky různých textur.

Výsledky syntézy periodicko-stochastických textur vzorkovací metodou jsou na obrázku 13, pro syntézu bylo použito 4 až 8 toroidních dlaždic, v závislosti na charakteru vstupního vzorku.

5 Závěr

Syntéza textur je alternativním a většinou jediným možným způsobem generování textur k přímému použití v počítačové vizualizaci. Má široké uplatnění v počítačové grafice, zejména pak v oblasti virtuální reality, protože v různých aplikacích je třeba modelovat objekty reálného světa. Aby tyto objekty vypadaly co nejvěrohodněji, je nutné je pokrýt vhodným povrchem. A zde se přímo nabízí syntéza textur. Místo velké textury, která by pokryla celý objekt, stačí mít malý vzorek, ze kterého se vhodnou metodou získá textura patřičných rozměrů. Takto získaná textura se pak mapuje na povrch objektu.

Vhodnou metodou se rozumí taková metoda, která odpovídá požadavkům aplikace. Hlavními faktory ovlivňující výběr metody jsou výsledná kvalita, rychlost a možnost komprese. Často je třeba zvolit určitý kompromis mezi kvalitou a rychlostí. Navíc je tato práce zaměřena na periodicko-stochastické textury. Využijeme-li vlastností periodicko-stochastických textur, pak lze pro toto spektrum textur vyvinout efektivní specializované metody.

Obě metody jsou schopny modelovat syntetické periodicko-stochastické textury z daného texturního vzorku, přičemž vstupní vzorek a syntetická textura jsou ve většině případů vizuálně nerozlišitelné.

První z metod umožňuje průměrnou kompresi při manipulaci s periodickou složkou, zatímco stochastickou složku lze díky adaptivním metodám modelovat na základě několika málo parametrů, čímž lze dosáhnout značné komprese.

Druhá z popsaných metod je použitelná nejen pro syntézu periodicko-stochastikých textur a dosahuje velmi dobrých výsledků. Z charakteru metody je zřejmé, že nebude fungovat na vzorcích textur, které jsou poškozeny nerovnoměrnou intenzitou jasu ve vzorku, nejsou dostatečně reprezentativní a nebo které jsou poškozeny geometrickou transformací. Dodejme, že v těchto případech selhávají metody považované za špičkové.

Literatura

[1] K. Djado, R. Egli, and F. Deschênes. Extraction of a representative tile from a near-periodic texture. In 'GRAPHITE '05: Proceedings of the 3rd international conference on Computer graphics and interactive techniques in Australasia and South East Asia', 331–337, New York, NY, USA, (2005). ACM Press.

[2] A. A. Efros and W. T. Freeman. Image quilting for texture synthesis and transfer. In 'SIGGRAPH '01: Proceedings of the 28th annual conference on Computer graphics and interactive techniques', 341–346, New York, NY, USA, (2001). ACM Press.

[3] M. Haindl and M. Hatka. BTF roller. In 'Texture 2005: Proceedings of 4th Internatinal Workshop on Texture Analysis and Synthesis', M. Chantler and O. Drbohlav, (eds.), 89–94, Edinburgh, (October 2005). Heriot-Watt University.

[4] M. Haindl and M. Hatka. A roller - fast sampling-based texture synthesis algorithm. In 'The 13th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision 2005', V. Skala, (ed.), 80–83, Plzen, (February 2005). University of Western Bohemia.

[5] M. Haindl and V. Havlíček. A multiscale colour texture model. In 'Proceedings of the 16th International Conference on Pattern Recognition', R. Kasturi, D. Laurendeau, and C. Suen, (eds.), 255–258, Los Alamitos, (August 2002). IEEE Computer Society.

Improvement of Model-based Prediction via Assimilation of Measured Data*

Radek Hofman

1st year of PGS, email: hofman@utia.cas.cz
Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU
advisor: Petr Pecha, Institute of Information Theory and Automation, AS CR

Abstract. The paper briefly presents the development of software system HARP which is being developed in IITA. The HARP system is designed as a decision support tool (DSS) for purposes of fast assessment of radiological consequences of accidental releases of radionuclides into living environment. Particular attention is devoted to assimilation subsystem of the HARP system and its exploitation in analysis and prediction of further evolution of a scenario dealing with long-term deposition of ^{137}Cs over terrain. The present work demonstrates utilization of Kalman filter in assimilation process of the scenario and also a model used for modeling of long-term exposure due to groundshine in Japan DSS OSCAAR (Off-Site Consequence Analysis code for Atmospheric Releases in reactor accidents) is introduced.

Abstrakt. Příspěvek stručně popisuje současný stav vývoje softwarového systému HARP vyvíjeného v UTIA. Jedná se o pomocný nástroj pro hodnocení a analýzu následků úniků radionuklidů z jaderných elektráren do okolního prostředí. Zvláštní pozornost je věnována asimilačnímu subsytému systému HARP a jeho využití pro analýzu a predikci vývoje aktivity z dlouhodobé depozice radioaktivního ^{137}Cs na terénu. V příspěvku je popsán Kalmanův filtr a jeho využití v asimilačním procesu konkrétního scénáře využívajícího model pro dávky záření z dlouhodobé depozice radionuklidů na zemském povrchu adaptovaného z japonského balíku modelů OSCAAR (Off-Site Consequence Analysis code for Atmospheric Releases in reactor accidents).

1 Introduction

In the current state of development is software system HARP capable to model atmospheric dispersion of radioactive pollutants and subsequent dose distributions and health effects in the exposed population. Main objective is to improve reliability of the model predictions via advanced statistical techniques of assimilation of model results with observations from terrain. The aim is to develop modeling, simulation and educational tool with unified user-friendly graphical interface for utilization in radiation protection.

2 HARP system

The HARP system consists of three main parts: Atmospheric dispersion model (ADM), Dose model (DOS) and Food-chain model (FCM), more in [3]. In recent development

*This work is part of the grant project GAČR No. 102/07/1596, which is funded by Grant Agency of the Czech Republic.

of HARP were numerical algorithms modified from deterministic to their probabilistic versions, see [2]. Probabilistic version enables for sensitivity analysis, uncertainty analysis and also for statistical estimate of error covariance structure of generated data. The block diagram of system architecture is in the Fig. (2).

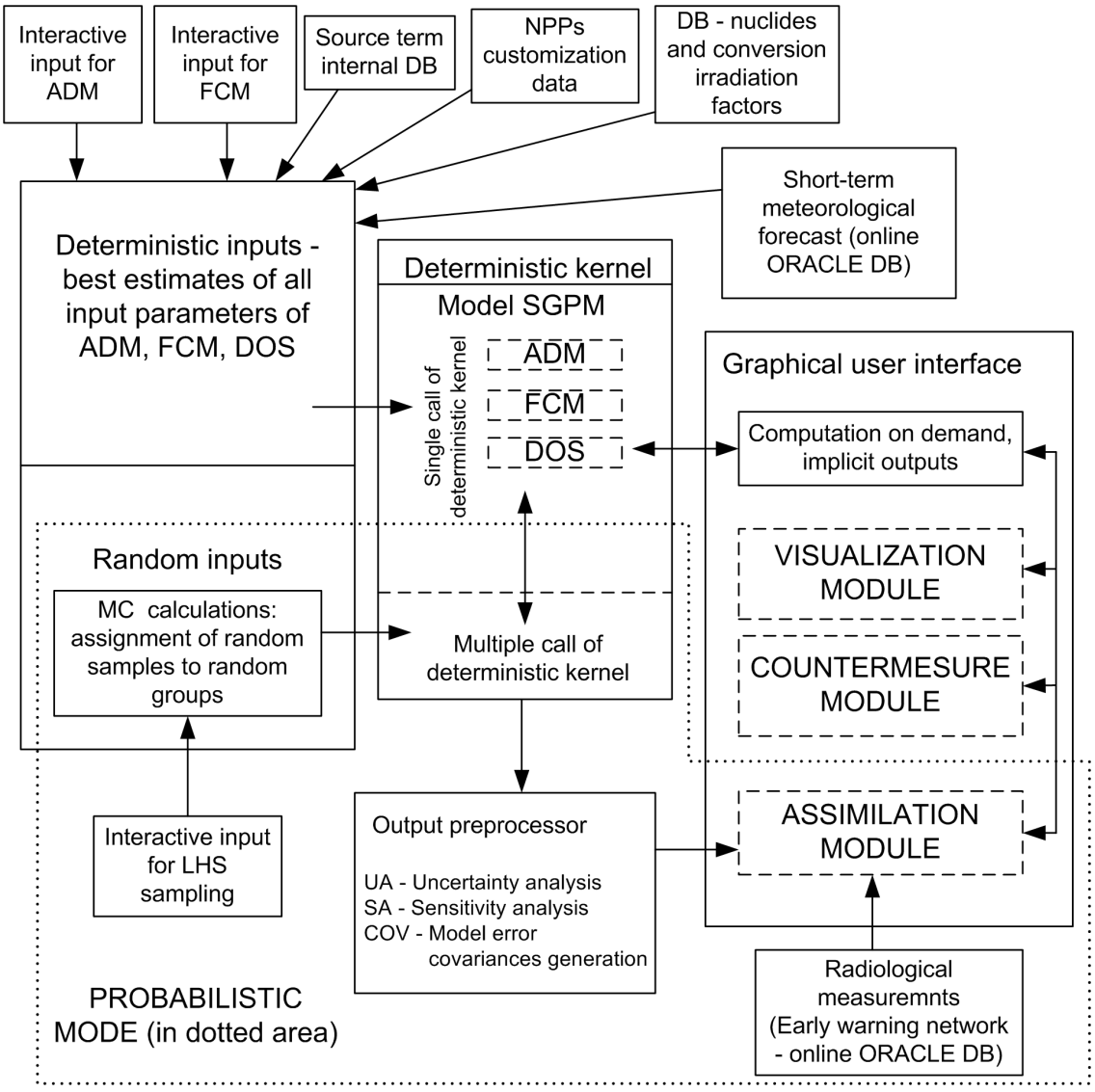


Figure 1: The architecture of the HARP system. Deterministic numerical kernel is interconnected to visualization and assimilation submodule via graphical user interface.

2.1 Assimilation submodule

Assimilation submodule offers comfortable graphical user interface for interactive insertion of data and its maintenance and evaluation. Numerical and assimilation subsystem have direct binding to visualization submodule (see Fig. (2)), so both modeled data and measurements can be easily visualized on relevant scalable map background. Evaluation of results is also supported by data tables and comparative graphs. Also access

to ORACLE database of meteorological forecasts and measurement stations included in Radiation Monitoring Network of the Czech Republic is established there. In the current state of the art are implemented following assimilation algorithms: Interpolation procedures, successive corrections method, optimal interpolation and Kalman filter.

2.2 Atmospheric dispersion model

ADM in HARP is based on segmented Gaussian plume model (SGPM). The segmentation allows us to use different set of input parameters for each of the segments. In the current state of development the model has more than hundred input parameters and the most significant of them $P_1 - P_{12}$ (resulting from sensitivity analysis) are listed in Table (2.2). In the table are distributions of random parameters multiplicatively applied to nominal values of input parameters in order to obtain probability distributions of those. Parameter distributions are expert-based estimates supplemented by measurements. The influence of the rest of model input parameters on variation of model output is assumed to be unimportant and these parameters are on input set to their best estimate values.

We divide the parameters into to groups: Local and global. The global parameters don't vary through the segments and remains same for all of them. Values of local parameters vary in time and/or with spatial location. SGPM enables to take into account

Variable	**G/L**	**Min**	**Mean**	**Max**	**Distribution**	σ
P_1 - intensity of release	G		1.0		normal	0.20
P_2 - horiz. dispersion	G		1.0		normal	0.13
P_3 - horiz. fluct. of wind dir.	L	-5	0	5	disc. uniform	
P_4 - dry deposition of elem.	G	0.41	1.0	1.6	uniform	
P_5 - dry deposition of aero.	G	0.41	1.0	1.6	uniform	
P_6 - elution of elem. iodine	G	0.2	1.0	5.0	log-uniform	
P_7 - elution of aero.	G	0.2	1.0	5.0	log-uniform	
P_8 - advection speed of plume	L	-1.0	0.0	1.0	uniform	
P_9 - wind profile	G	0.5	1.0	1.5	uniform	
P_{10} - vertical dispersion	G		1.0		normal	0.13
P_{11} - mixing layer height	G	0.6	1.175	1.75	uniform	
P_{12} - heat flux	G	0.0	0.5	1.0	uniform	

Table 1: The most significant parameter of ADM and distribution of multiplicative factors used for their generating from nominal values.

realistic weather forecasts hourly provided by the Czech Hydro-Meteorological Institute. ADM also accounts for many factors affecting the plume depletion (dry/wet deposition, influence of terrain type etc.).

3 Scenario for long-term deposition of ^{137}Cs over terrain

The plume moving over the terrain leaves behind a radioactive trace due to dry and wet activity deposition. Movement of a plume over observed area lasts usually few hours.

When the plume leaves observed area, the trace represents an initial condition for prediction of further evolution of radiological situation on terrain. Our analyzed scenario starts just after the plume leaves the observed terrain. An emphasis is laid on prediction of long term evolution of radiological situation in time horizon of years up to tens of years. This knowledge is important for planning of long-term countermeasures relating to food-chain model, which is also a part of the HARP system.

The only nuclide assumed in this scenario is ^{137}Cs. As half-time to decay of ^{137}Cs is approximately 30 years, it is one of most significant and dominant source of radioactivity in long term scenarios.

Initial conditions (background field) for assimilation is given by the ADM when the radioactive plume is gone. As s model of radiation situation evolution is used the relation according to Eq. (3) from the OSCAAR model. The crucial task is to estimate error covariance structure of the model and the background field. As a first attempt, we are trying to estimate error covariance structure by Monte-Carlo approach as a sample covariance of a drawn sample of size $N \approx 10^3$. The sample was generated according to given probability distributions, see Table (3.1).

Alternative way of estimation of error covariance structure could be a spatial filter widely used in meteorology because of high dimensionality of problems solved there. Spatial filters are based on assumption: The bigger distance between two points the smaller correlation between modeled/measured values in these points. This assumption is rather unrealistic and physically inexact and also denies one of major advantages of assimilation methods - embodying of physical information. Spatial filters for determination of correlation between points i and j proposed by Bergthorson and Doos are as follows:

$$\mathbf{P}_{ij}^{f} = exp\left[-\frac{r_{ij}^2}{L}\right] \tag{1}$$

$$\mathbf{P}_{ij}^{f} = (1 + \frac{r_{ij}}{L})exp\left[-\frac{r_{ij}}{L}\right] \tag{2}$$

where r_{ij} is the distance between the points and L is a chosen constant called radius of influence.

3.1 OSCAAR model

Abbreviation OSCAAR stands for Off-Site Consequence Analysis code for Atmospheric Releases in reactor accidents and it has been developed within the research activities on probabilistic safety assessment at the Japan Atomic Research Institute [7]. OSCAAR consists of a series of interlinked modules and that are used to calculate the atmospheric dispersion and deposition of selected radionuclides. In this work are adopted formulae and principles used in OSCAAR for prediction of dose rate due to long-term groundshine. It can be expressed by the Eq. (3).

$$D_s(t) = SD_k \cdot R(t) \cdot E(t) \cdot DF_g \cdot L \cdot \sum_i f_i \cdot (OF_i^{out} + OF_i^{in} \cdot SF) \tag{3}$$

The interpretation of terms in Eq. (3) is in the Table (3.1). The following exponential

$D_g(t)$	dose rate on day t after deposition of a radionuclide ($Sv\ s^{-1}$)
SD_k	total deposition of the radionuclide at place k ($Bq\ m^{-2}$)
$R(t)$	factor to account for radioactive dacay occuring between the deposition and t
$E(t)$	factor to account for the environmental dacay of groundshine ($Sv\ s^{-1}\ perBq\ m^{-2}$)
DF_g	dose-rate conversion factor for groundshine
L	geometric factor (%)
f_i	fraction of i-th occupation group (%)
OF_i^{out}	outdoor occupancy factor for i-th occupation group (%)
OF_i^{in}	indoor occupancy factor for i-th occupation group (%)
SF	shielding factor for wooden of brick house (%)

Table 2: Interpretation of term in Eq. (3)

functions represent the two factors of $R(t)$ and $E(t)$ as a functions of time. The experiments had shown that the mitigation of groundshine due to environmental decay follows relation given by superposition of two exponentials (Eq. (5), (6)).

$$R(t) = exp(-ln2 \cdot \frac{t}{T_y}) \tag{4}$$

$$E(t) = d_f \cdot exp(-ln2 \cdot \frac{t}{T_{sf}}) + d_s \cdot exp(-ln2 \cdot \frac{t}{T_{ss}}) \tag{5}$$

where

$$d_f + d_s = 1 \tag{6}$$

Ground deposition model formulae are semi-empirical, it means that some of equation parameters are determined empiricaly upon measurements and the parameter values depend on the local conditions in the place of model application (soil type etc.). The dose conversion factor was calculated by the method of Kocher (1980) in which the exposed individual was assumed to stand on a smooth, infinite plane surface with uniform concentration of source of radioactivity. Data used in the groundshine dose calculations

Variable	**Mean**	**Min**	**Max**	**Distribution**	**Units**
d_s	0.52	0.40	0.71	Uniform	-
T_{sf}	1.1	0.41	1.4	Uniform	y
T_{ss}	28	24.3	29.4	Uniform	y
L	0.45	0.2	0.7	Uniform	-
SF(wood)	0.52	0.26	0.78	Uniform	-
SF(brick)	0.2	0.1	0.3	Uniform	-
DF_g	5.86×10^{-16}	-	-		$Sv\ s^{-1}/Bq\ m^{-2}$

Table 3: Parameter values used for ground exposure calculations in OSCAAR model.

are given in Table (3.1). The parameter distributions were determined for ^{137}Cs from Chernobyl disaster. The appropriate data for other elements are not available but it is

assumed that the long-term influence of most of them is not significant. For elements with high half-time to decay are assumed the same equations of groundshine mitigation as for ^{137}Cs. As in the HARP, the approach used in OSCAAR adopted probabilistic methodology and it allows us to determine error covariance structure of the model. It is a necessary condition for application of advanced assimilation techniques to the model (Kalman filter, 4DVAR).

4 Data assimilation

The goal of data assimilation is to provide an analysis which relies on measurements and so called background field from a model forecast. Other inputs to data assimilation process can be physical constraints on the problem or any additional prior knowledge not included in the model. Merging of these contending sources of information had shown to be very promising in many branches of contemporary Earth sciences like meteorology and oceanography.

In data assimilation we try to adjust model according to measured values what represents research effort to move from isolated model prediction forward reality. An automatic procedure for bringing observations into the model is called objective analysis. The major progress of the objective analysis was achieved in the field of meteorological forecasting techniques that represents efficient tool in struggle with tendency to chaotic destruction of physical knowledge, see [5]. Advanced assimilation methods are capable to take into account measurements and model errors in form of error covariance matrices.

In the Fig. (4) we can see the schematic of two stage assimilation process. In the first stage called data update step are modeled values adjusted according to all measurements available in certain time step. This part of data assimilation process is often called intermittent assimilation. This step allows us to get new and hopefully better initial conditions for time update step which performs the prediction of evolution of an analyzed quantity. Advanced assimilation algorithms also enables for prediction of evolution of model errors. Without data update step we could get a prediction substantially diverging from the true evolution.

4.1 Kalman filter

The Kalman filter ([1]) has long been regarded as the optimal solution to many tracking and data prediction tasks. The purpose of filtering is to extract the required information from a signal, ignoring everything else. Kalman described his filter using state space technique which enables filter to be used as either a smoother, a filter or a predictor. In this paper is presented exploitation of Kalman filtering method in a special assimilation scenario.

As was already stated in previous paragraph, initial condition for the task of prediction of radiological situation evolution is given as a result of ADM when the plume is gone. Reliability of this initial value $\mathbf{x}_f$ (often called background field) can be improved by assimilation process. If there are available some measurements $\mathbf{y}_o^t$ at time t which we assume to be more reliable then the model, we can adjust the model according to their values with respect to physical information contained in the model. Error covariance

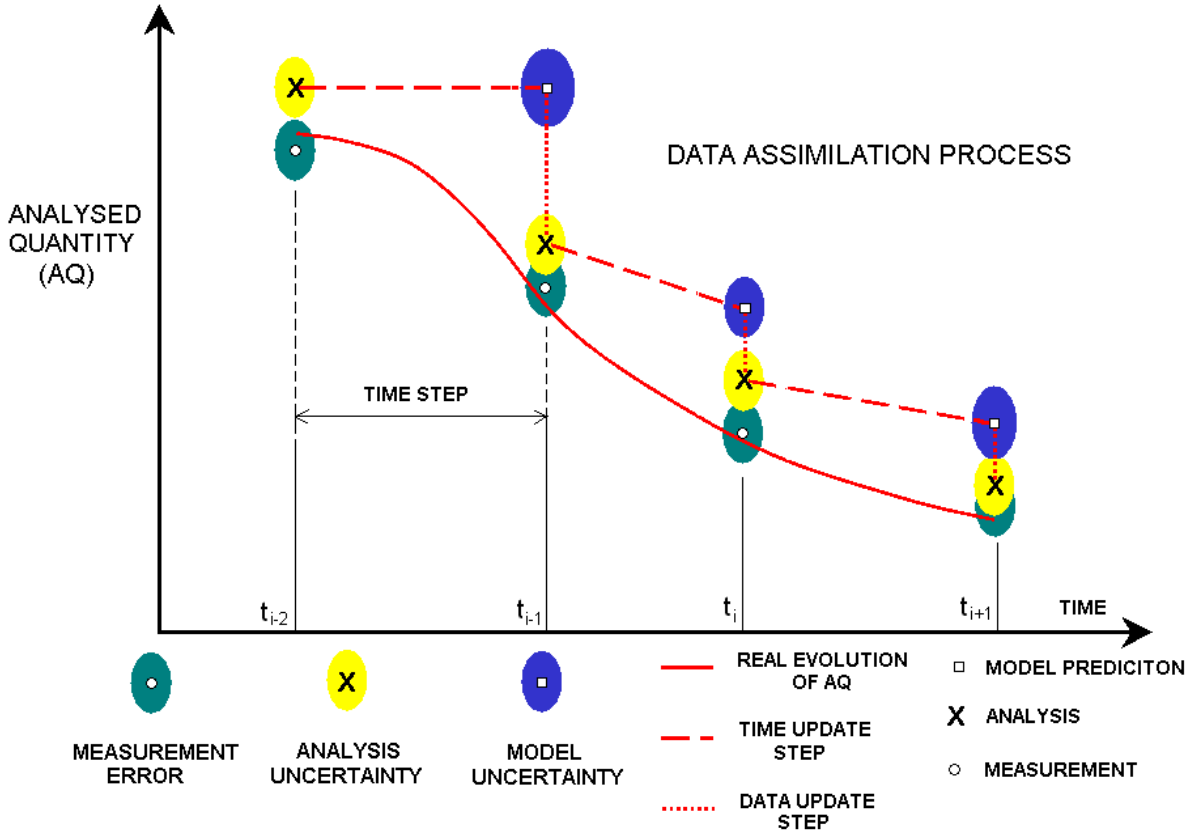


Figure 2: The schematic of data assimilation process.

structure is expressed in form of error covariance matrices of model $\mathbf{P}_f^t$ and measurements $\mathbf{R}^t$. The result of this process in time t is a new better initial condition $\mathbf{x}_a^t$ called analysis (Eq. (7), Eq. (8)) and information on its error covariance structure $\mathbf{P}_a^t$ (Eq. (9)). $\mathbf{H}$ is a linear operator for transformation of points from space of model into the space of measurements. This process is called data update step of Kalman Filter.

$$\mathbf{x}_a^t = \mathbf{x}_f^t + \mathbf{K}^t(\mathbf{y}_o^t - \mathbf{H}\mathbf{x}_f^t) \tag{7}$$

$$\mathbf{K}^t = \mathbf{P}_f^t\mathbf{H}^T(\mathbf{R}^t + \mathbf{H}\mathbf{P}_f^t\mathbf{H}^T)^{-1} \tag{8}$$

$$\mathbf{P}_a^t = (\mathbf{I} - \mathbf{K}^t\mathbf{H})\mathbf{P}_f^t \tag{9}$$

The second step is called time update and in this step is performed time evolution of an analyzed quantity via linear model $\mathbf{M}$ (Eq. (10)) and also evolution of its error covariance structure (Eq. (11)). Output of second step of Kalman filter is prediction $\mathbf{x}_f^{t+1}$ and information on error of this prediction $\mathbf{P}_f^{t+1}$.

$$\mathbf{x}_f^{t+1} = \mathbf{M}\mathbf{x}_a^t \tag{10}$$

$$\mathbf{P}_f^{t+1} = \mathbf{M}\mathbf{P}_a^t\mathbf{M}^T + \mathbf{Q} \tag{11}$$

This two steps can be iteratively repeated as long as new measurements are available.

5 Results and Conclusion

In assumed scenario the exposure was assumed with no shielding, so shielding coefficients were set to 1. Because of lack of real measurements testing of an assimilation process was performed with simulated measurements sampled from the same numerical model using perturbed input parameters. Early results which will be presented in oral part of presentation show that this task can be successfully solved via two step data assimilation process, but there are still some problems especially with estimation of error covariance structure and its propagation forward in time.

The achieved results had shown so far that the differentiation of ADM input parameters to local and global introduced in paragraph 2.2 substantially influences error covariation structure of the model. Choice of parameters to vary in order to estimate error covariance structure is important part of assimilation process. Some results were already published in [6]. The results from spatial filter (Eq. (1), (2)) could by used for weighing of statistical estimate of error covariance structure and to mitigate the influence of global vs. local property of certain input parameter.

In the next development of assimilation methodology and HARP system is intended to implement some other advanced assimilation methods based on Bayesian approach.

References

[1] K. E. *Atmospheric modeling, data assimilation and predictability.* Cambridge Univ. Press, Cambridge, (2003).

[2] P. P. and P. E. *Modelling of random activity concentration fields in air for purposes of probabilistic estimation of radiological burden.* In '10th Int. Conf. on Harmonization within Atmospheric Dispersion Modelling, Ioannina, Greece', (2005).

[3] P. P., H. R., and P. E. *Training simulator for analysis of environmental consequences of accidental radioactivity releases.* In '6th EUROSIM Congress on Modelling and Simulation, Ljubljana, Slovenia', (2007).

[4] P. Pecha and R. Hofman. *Integration of data assimilation subsystem into environmental model of harmful substances propagation.* In 'Harmo11 - 11th Internal Conf. Cambridge', (2007).

[5] D. R. *Atmospheric data analysis.* Cambridge Univ. Press, Cambridge, (1991).

[6] H. R. *Assimilation scenario for long-term deposition of* ^{137}cs. In '8th International PhD Workshop on Systems and Control a Young Generation Viewpoint, Balatonfured, Hungary', (2007).

[7] H. T. and M. T. *OSCAAR Model -Description and Evaluation of Model Performance*, (2006).

Informational Cathegorical Data Clustering*

Jan Hora

3rd year of PGS, email: hora@utia.cas.cz
Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU
advisor: Jiří Grim, Institute of Information Theory and Automation, AS CR

Abstract. The EM algorithm has been used repeatedly to identify latent classes in categorical data by estimating finite distribution mixtures of product components. Unfortunately, the underlying mixtures are not uniquely identifiable and, moreover, the estimated mixture parameters are starting-point dependent. For this reason we use the latent class model only to define a set of "elementary" classes by estimating a mixture of a large number components. As such a mixture we use also an optimally smoothed kernel estimate. We propose a hierarchical "bottom up" cluster analysis based on unifying the elementary latent classes sequentially. The clustering procedure is controlled by minimum information loss criterion.

Abstrakt. Shlukování kategoriálních dat je často řešeno hledáním tzv. latentních tříd pomocí EM algoritmu. Tento přístup ovšem závisí na počátečním řešení a naráží na problém neidentifikovatelosti směsi. Popisovaná metoda vyhledává shluky nikoliv jako jednotlivé komponenty směsi jako v případě latentních tříd, ale jako podsměsi vzniklé sloučením několika jednoduchých tříd z odhadnuté distribuční směsi s vyšším počtem komponent. Extrémní variantou takové směsi může být jádrový odhad, jehož optimální vyhlazení je v práci popsáno. V práci je dále představena metoda hierarchického shlukování s kritériem nejmenší informační ztráty.

1 Introduction

The concept of cluster analysis is closely related to the similarity of objects or distance of data vectors defined by a metric. The cluster analysis of categorical (nominal, qualitative) data is difficult because the standard arithmetical operations are undefined and also there is no generally acceptable definition of distance for multivariate categorical data. For these reasons the available methods of cluster analysis cannot be applied directly to categorical data.

At present the standard approach to cluster analysis of categorical data is to introduce some similarity measure or distance function in a heuristical manner. It appears that the only statistically justified method to analyze multivariate categorical data is the latent class model of Lazarsfeld [5]. Motivated by sociological research he proposed the fitting of multivariate Bernoulli mixtures to binary data with the aim to identify possible latent classes of respondents. Serious drawback of the Lazarsfeld's idea has been the tedious and somewhat arbitrary methods used for fitting the models. The numerical problems have been removed by the computationally efficient EM algorithm [1]. In the last years the original idea of Lazarsfeld has been widely applied and frequently modified by different authors (cf. e.g. [3] and [9] for extensive references).

*This research was supported by the grant GACR 102/07/1594 of the Czech Grant Agency and by the projects of the Grant Agency of MŠMT 2C06019 ZIMOLEZ and 1M0572 DAR.

In this paper we propose a hierarchical approach to cluster analysis of categorical data in the context of data mining. Applying the latent class model to large multivariate databases we assume a large number of classes ($M \approx 10^1 \div 10^2$) with the aim to approximate the unknown probability distribution. The EM algorithm yields different parameter estimates but the approximation accuracy of the estimated mixture is comparable. The initial parameters of the estimated mixture can be chosen randomly without affecting the quality of estimates essentially. Unlike the latent class analysis we use the estimated mixture components only to identify "elementary" latent classes with the posterior component weights playing the role of membership functions. The underlying decision problem can be characterized by the statistical decision information. We assume that the statistical properties of data can be described by the estimated mixture even if the "elementary" components are not defined uniquely. We assume that potential clusters can be identified by the optimal decomposition of the estimated mixture into sub-mixtures. We propose a hierarchical clustering procedure based on sequential unifying of the elementary latent classes. The procedure is controlled by the minimum information loss criterion.

Another way to describe the given data is the kernel estimate. In this paper we present an optimally smoothed kernel estimate which can be used as a distribution mixture for above mentioned clustering.

The paper is organized as follows. We first describe the idea of latent class analysis and the related problem of estimating discrete product mixtures by means of EM algorithm (Sec. 2). Section 3 introduces the statistical information criterion, Sec. 4 describes the method of hierarchical cluster analysis and the section 5 describes the possibility of kernel estimation. The application of the method is illustrated by numerical example in Sec. 6. Finally we discuss the main results in the Conclusion.

2 Latent Class Model

Let us suppose that some objects are described by a vector of discrete variables taking values from finite sets:

$$\boldsymbol{x} = (x_1, \ldots, x_N), \quad x_n \in \mathcal{X}_n, \ |\mathcal{X}_n| < \infty, \quad \boldsymbol{x} \in \mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_N. \tag{1}$$

We assume that the variables are categorical (i.e. non-numerical, nominal, qualitative) without any type of ordering. Considering the problem of cluster analysis we are given a set of data vectors

$$\mathcal{S} = \{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(K)}\}, \quad \boldsymbol{x}^{(k)} \in \mathcal{X} \tag{2}$$

and the goal of cluster analysis is to partition the set $\mathcal{S}$ into "natural" well separated subsets of similar objects

$$\Re = \{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_M\}, \quad \mathcal{S} = \cup_{j=1}^{J} \mathcal{S}_j, \qquad \mathcal{S}_i \cap \mathcal{S}_j = \emptyset, \quad \text{for } i \neq j. \tag{3}$$

In this sense the concept of cluster analysis is closely related to some similarity or dissimilarity measures. Unfortunately, in case of categorical variables the arithmetical operations are undefined and therefore we cannot compute means and variances nor there is any generally acceptable way to define distance for the categorical data vectors $\boldsymbol{x} \in \mathcal{X}$. Binary

data, as a special case, may appear to be naturally ordered, however, the values 0 and 1 are often assigned quite arbitrarily. For these reasons the available algorithms of cluster analysis are not directly applicable to categorical data.

The standard way to avoid this difficulty is to introduce a similarity measure or distance function for categorical data in a heuristical manner. It may appear quite easy to define a distance table for a single categorical variable, especially in case of some well interpretable values. However, in a multidimensional space the problem of distance definition becomes difficult because of uneasy foreseen consequences of interference of different distance tables.

As it appears the only statistically justified approach to clustering categorical data can be traced back to the latent structure analysis of Lazarsfeld [5] who proposed to identify latent classes in binary data by estimating multivariate Bernoulli mixtures. The method is easily generalized to categorical variables and it is often applied in different modifications as "latent class analysis" [9]. The latent class model is defined as a finite mixture of a given number of product components

$$P(\boldsymbol{x}) = \sum_{m\in\mathcal{M}} w_m F(\boldsymbol{x}|m), \quad \boldsymbol{x}\in\mathcal{X}, \qquad \mathcal{M} = \{1,\dots,M\}. \tag{4}$$

Here w_m are non-negative probabilistic weights

$$\sum_{m\in\mathcal{M}} w_m = 1, \qquad 0 < w_m < 1, \quad m\in\mathcal{M}, \tag{5}$$

$F(\boldsymbol{x}|m)$ are the mixture components defined as products of univariate conditional (component specific) discrete distributions $f_n(x_n|m)$

$$F(\boldsymbol{x}|m) = \prod_{n\in\mathcal{N}} f_n(x_n|m), \quad \mathcal{N} = \{1,\dots,N\} \tag{6}$$

and $\mathcal{M},\mathcal{N}$ are the index sets of components and variables respectively.

The latent class model (4) naturally defines a statistical decision problem. Having estimated the mixture parameters we can compute the conditional probabilities

$$q(m|\boldsymbol{x}) = \frac{w_m F(\boldsymbol{x}|m)}{\sum_{j\in\mathcal{M}} w_j F(\boldsymbol{x}|j)}, \quad \boldsymbol{x}\in\mathcal{X}, \quad m\in\mathcal{M} \tag{7}$$

which can be viewed as membership functions of the estimated latent classes. They are particularly useful if there is some interpretation of the mixture components, e.g. if the components can be shown to correspond to some real "latent classes" [5], "hidden causes" [6] or "clusters" having a specific meaning.

A unique classification of data vectors $\boldsymbol{x}\in\mathcal{X}$ can be obtained by means of Bayes decision function (with the ties arbitrarily decided)

$$d(\boldsymbol{x}) = \arg\max_{j\in\mathcal{M}}\{q(j|\boldsymbol{x})\}, \quad \boldsymbol{x}\in\mathcal{X}. \tag{8}$$

By using the Bayes decision function $d(\boldsymbol{x})$ we obtain the elementary "latent class" partition $\Re$ of the set $\mathcal{S}$ by classifying the points $\boldsymbol{x}\in\mathcal{S}$:

$$\Re = \{\mathcal{S}_1,\mathcal{S}_2,\dots,\mathcal{S}_M\}, \quad \mathcal{S}_m = \{\boldsymbol{x}\in\mathcal{S} : d(\boldsymbol{x}) = m\}, \quad m\in\mathcal{M}. \tag{9}$$

In other words the partition $\Re$ is defined by the maximum posterior weights $q(m|\boldsymbol{x})$ and represents the result of latent class analysis in the original form as proposed by Lazarsfeld (cf. [5], [3]), [9]). The latent class model (4) seems to be one of the most widely applicable tools of cluster analysis of categorical data. The original idea of Lazarsfeld has been used by many authors to identify individual classes of bacteria (cf. e.g. [3]) and more recently Vermunt et al. [9] describe different modifications of the latent class analysis as applied in diverse fields.

The standard way of estimating mixtures is to use EM algorithm (cf. [1], [4]). In particular to compute maximum-likelihood estimates of mixture parameters we maximize the log-likelihood function

$$L = \frac{1}{|\mathcal{S}|} \sum_{\boldsymbol{x} \in \mathcal{S}} \log P(\boldsymbol{x}) = \frac{1}{|\mathcal{S}|} \sum_{\boldsymbol{x} \in \mathcal{S}} \log \left[\sum_{m \in \mathcal{M}} w_m F(\boldsymbol{x}|m) \right] \tag{10}$$

A serious disadvantage of the latent class analysis relates to the fact that the resulting clusters may be non-unique. It is obvious that, if the estimated mixture is not defined uniquely, then the corresponding interpretation of data in terms of latent classes may become questionable. Unfortunately, there are at least three sources of uncertainty which may influence the resulting mixture parameters. First, there is no exact method to choose the proper number of mixture components (cf. [4]). Another source of multiple solutions is the existence of local maxima of the log-likelihood function (10). For this reason we can expect different locally optimal solutions depending on the chosen initial parameters. However, even if we succeed to manage the computational aspects of mixture estimation, there is still the well known theoretical problem that the latent class model is not identifiable (cf. [3]). In particular it is easily verified that any non-degenerate mixture (4) can be expressed equivalently in infinitely many different ways [2].

3 Minimum Information Loss Criterion

In view of Sec. 2 the latent class model (4) is the only information source about the structural properties of the data set $\mathcal{S}$. For this reason we identify the clusters by means of the optimal decomposition of mixture (4) into sub-mixtures.

Recall that having estimated the mixture parameters we can define the elementary latent classes by classifying the data vectors $\boldsymbol{x} \in \mathcal{S}$ according to the maximum posterior weight $q(m|\boldsymbol{x})$ (cf. (8), (9)). The underlying decision problem can be characterized by the statistical decision information. By using the Shannon formula we can write

$$I(\mathcal{X}, \mathcal{M}) = H(\mathcal{M}) - H(\mathcal{M}|\mathcal{X}), \qquad H(\mathcal{M}) = \sum_{m \in \mathcal{M}} -w_m \log w_m, \tag{11}$$

$$H(\mathcal{M}|\mathcal{X}) = \sum_{\boldsymbol{x} \in \mathcal{X}} P(\boldsymbol{x}) H_x(\mathcal{M}) = \sum_{\boldsymbol{x} \in \mathcal{X}} P(\boldsymbol{x}) \sum_{m \in \mathcal{M}} -q(m|\boldsymbol{x}) \log q(m|\boldsymbol{x}). \tag{12}$$

Here $H(\mathcal{M})$ is the uncertainty connected with estimating the outcome $m \in \mathcal{M}$ of a random experiment with the probabilities $\{w_1, \ldots, w_M\}$ without any other knowledge. Given a vector $\boldsymbol{x} \in \mathcal{X}$ we can improve the estimation accuracy by computing the more

specific conditional probabilities $q(m|\boldsymbol{x})$. The statistical decision information $I(\mathcal{X},\mathcal{M})$ contained in the latent class model is defined as the difference between the a priori entropy $H(\mathcal{M})$ and the mean conditional entropy $H(\mathcal{M}|\mathcal{X})$ which corresponds to the knowledge of $\boldsymbol{x} \in \mathcal{X}$.

It can be seen that a partition $\mathcal{U}$ of the index set $\mathcal{M}$

$$\mathcal{U} = \{\mathcal{M}_1, \mathcal{M}_2, ..., \mathcal{M}_C\}, \qquad \bigcup_{c=1}^{C} \mathcal{M}_c = \mathcal{M}, \quad i \neq j \Rightarrow \mathcal{M}_i \cap \mathcal{M}_j = \emptyset. \tag{13}$$

actually defines a decomposition of the estimated mixture into sub-mixtures:

$$P(\boldsymbol{x}) = \sum_{c=1}^{C} \sum_{m \in \mathcal{M}_c} w_m F(\boldsymbol{x}|m) = \sum_{c=1}^{C} P(\boldsymbol{x}|\mathcal{M}_c)\ p(c), \qquad \boldsymbol{x} \in \mathcal{X}, \tag{14}$$

$$P(\boldsymbol{x}|\mathcal{M}_c) = \sum_{m \in \mathcal{M}_c} \frac{w_m}{p(c)} F(\boldsymbol{x}|m), \qquad p(c) = \sum_{m \in \mathcal{M}_c} w_m, \qquad c = 1, \ldots, C. \tag{15}$$

The sub-mixtures $P(\boldsymbol{x}|\mathcal{M}_c)$ can be then used to define the partition of the space $\mathcal{X}$ into corresponding clusters. We can write

$$p(c|\boldsymbol{x}) = \frac{p(c)P(\boldsymbol{x}|\mathcal{M}_c)}{P(\boldsymbol{x})} = \frac{\sum_{m \in \mathcal{M}_c} w_m F(\boldsymbol{x}|m)}{P(\boldsymbol{x})} = \sum_{m \in \mathcal{M}_c} q(m|\boldsymbol{x}), \tag{16}$$

$$d(\boldsymbol{x}|\mathcal{U}) = \arg\max_{c} \{p(c|\boldsymbol{x})\}, \quad \boldsymbol{x} \in \mathcal{X} \tag{17}$$

and by using the decision function $d(\boldsymbol{x}|\mathcal{U})$ we obtain the partition

$$\Phi = \{\mathcal{X}^{(1)}, \mathcal{X}^{(2)}, \ldots, \mathcal{X}^{(C)}\}, \quad \mathcal{X} = \cup_{j=1}^{C} \mathcal{X}^{(j)}, \qquad \mathcal{X}^{(i)} \cap \mathcal{X}^{(j)} = \emptyset, \quad \text{for } i \neq j. \tag{18}$$

for which we get

$$P(\mathcal{X}^{(c)}) = \sum_{\boldsymbol{x} \in \mathcal{X}^{(c)}} P(\boldsymbol{x}) = \sum_{c=1}^{C} P(\mathcal{X}^{(c)}|\mathcal{M}_c)\ p(c); \qquad P(\mathcal{X}^{(c)}|\mathcal{M}_c) = \sum_{\boldsymbol{x} \in \mathcal{X}^{(c)}} P(\boldsymbol{x}|\mathcal{M}_c) \tag{19}$$

Here the clusters $\mathcal{X}^{(c)} \in \Phi$ correspond to the respective sub-mixtures $P(\boldsymbol{x}|\mathcal{M}_c)$. By using the Shannon formula we can express the statistical decision information contained in the decomposed mixture. In analogy with (11), (12) we can write

$$I(\Phi,\mathcal{U}) = H(\Phi) - H(\Phi|\mathcal{U}), \qquad H(\Phi) = \sum_{\mathcal{X}^{(c)} \in \Phi} -P(\mathcal{X}^{(c)}) \log P(\mathcal{X}^{(c)}), \tag{20}$$

$$H(\Phi|\mathcal{U}) = \sum_{c \in \mathcal{U}} p(c) H_{\mathcal{M}_c}(\Phi), \qquad H_{\mathcal{M}_c}(\Phi) = \sum_{\mathcal{X}^{(i)} \in \Phi} -P(\mathcal{X}^{(i)}|\mathcal{M}_c) \log P(\mathcal{X}^{(i)}|\mathcal{M}_c). \tag{21}$$

Intuitively it is clear that by fusing sub-mixtures (or components) we loose some decision information. Indeed, we can easily verify that the decision information decreases if we join any two subsets $\mathcal{M}_i, \mathcal{M}_j \in \mathcal{U}$ of a given partition $\mathcal{U}$ (for more see [10]).

4 Minimum Information Loss Cluster Analysis

In view of the above equations any cluster analysis based on mixture decomposition is connected with some information loss from the point of view of the underlying decision problem. Naturally we would be interested in maximizing the statistical information about Φ contained in $\mathcal{U}$ and therefore the elementary information loss seems to be a suitable criterion to control the process of sequential fusion of the components and submixtures in the original latent class model. However, our experiments have shown that better results are gained when relative information is used.

In order to evaluate the information loss connected with the partitions Φ and $\mathcal{U}$ we have to estimate the information $I(\Phi,\mathcal{U})$ by means of the observation sample $\mathcal{S}$. By using the estimates

$$\hat{q}(m|\mathcal{X}^{(j)}) = \frac{1}{|\mathcal{S}_j|}\sum_{x\in\mathcal{S}_j} q(m|\boldsymbol{x}), \quad \hat{P}(\mathcal{X}^{(j)}) = \frac{|\mathcal{S}_j|}{|\mathcal{S}|}, \quad j \in \mathcal{M} \tag{22}$$

we can write

$$\hat{p}(c|\mathcal{X}^{(c)}) = \sum_{m\in\mathcal{M}_c} \hat{q}(m|\mathcal{X}^{(j)}) = \sum_{m\in\mathcal{M}_c} \frac{1}{S_c}\sum_{\boldsymbol{x}\in S_c} q(m|\boldsymbol{x}) \tag{23}$$

and therefore

$$\hat{H}(\mathcal{U}|\Phi) = \sum_{\mathcal{X}^{(c)}\in\Phi} P(\mathcal{X}^{(c)})\hat{H}_{\mathcal{X}^{(c)}}(\Phi) = -\sum_{\mathcal{X}^{(c)}\in\Phi} P(\mathcal{X}^{(c)})\sum_{c\in\mathcal{U}} \hat{p}(c|\mathcal{X}^{(c)})\log\hat{p}(c|\mathcal{X}^{(c)}) \tag{24}$$

and finally we obtain the criterion Q_{ij}

$$Q_{ij} = \frac{\hat{I}(\Phi,\mathcal{U})}{\hat{H}(\mathcal{U})} - \frac{\hat{I}(\Phi',\mathcal{U}')}{\hat{H}(\mathcal{U}')} = -\frac{\hat{H}(\mathcal{U}|\Phi)}{\hat{H}(\mathcal{U})} + \frac{\hat{H}(\mathcal{U}'|\Phi')}{\hat{H}(\mathcal{U}')} \tag{25}$$

The described criterion Q_{ij} is an estimate of the relative information loss arising after union of the two subsets $\mathcal{M}_i,\mathcal{M}_j$ in the partition $\mathcal{U}$ and the corresponding clusters $\mathcal{X}^{(i)}$, $\mathcal{X}^{(j)}$ in the partition Φ. In the following we use the estimated relative information loss Q_{ij} as a criterion for the optimal choice of the pair of subsets to be unified. In other words, in each step of the procedure we unify the two sets $\mathcal{M}_i,\mathcal{M}_j \in \mathcal{U}$ and $\mathcal{X}^{(i)},\mathcal{X}^{(j)} \in \Phi$ for which the resulting information loss Q_{ij} is minimized.

In the considered decision-making framework a natural goal of cluster analysis is to preserve maximum decision information with a minimum number of clusters. Let us remark that the most general result of the above algorithm is the sequence of information loss values $\{Q_{ij}^{(k)}\}_{k=1}^{M}$ produced by the hierarchical clustering procedure. The form of the sequence suggests different possibilities of final clustering and simultaneously it can be seen how justified are the resulting clusters. For a given mixture (4) the sequence $\{Q_{ij}^{(k)}\}_{k=1}^{M}$ is defined uniquely and the form of the sequence should be similar for comparably good estimates of the underlying latent class model.

The proposed method of cluster analysis of categorical data can be summarized as follows:

Algorithm:

1. Estimation of the latent class model (4) for the categorical data set $\mathcal{S}$ by means of EM algorithm for a sufficiently large M.
2. Definition of the basic mixture $\mathcal{U} = \{\{1\}, \{2\}, \ldots, \{M\}\}$ and the basic latent class partition $\Phi = \{\mathcal{X}^{(1)}, \mathcal{X}^{(2)}, \ldots, \mathcal{X}^{(M)}\}$.
3. Sequential unifying the most similar subsets $\mathcal{M}_i, \mathcal{M}_j \in \mathcal{U}$ and $\mathcal{X}^{(i)}, \mathcal{X}^{(j)} \in \Phi$ for which the resulting relative information loss Q_{ij} (cf. (25)) is minimal.
4. Choice of the optimal partition $\mathcal{U}^*$ according to the point of $Q_{ij}^{(k)}$.
5. Definition of the resulting clusters in $\mathcal{S}$ by means of the decision function $d(\boldsymbol{x}|\mathcal{U}^*)$.

5 Discrete Kernel Estimate

In previous sections we worked with distribution mixtures with component count $M \ll |S|$ which parameters were estimated by EM algorithm. Another way to approximate the propability distribution $P(\boldsymbol{x})$ is a non-parametric kernel estimate. For a given data sample S the discrete Parzen estimate can be defined as

$$\hat{P}(\boldsymbol{x}) = \frac{1}{|S|} \sum_{\boldsymbol{y} \in S} G(\boldsymbol{x}|\boldsymbol{y}) \tag{26}$$

where the kernel function $G(\boldsymbol{x}|\boldsymbol{y})$ can be expressed as a product of binomial functions g_n

$$G(\boldsymbol{x}|\boldsymbol{y}) = \prod_{n \in \mathcal{N}} g_n(x_n|y_n), \qquad g_n(x_n|y_n) = \begin{cases} \alpha_n & \text{for } x_n = y_n, \\ \beta_n & \text{for } x_n \neq y_n, \end{cases} \qquad \begin{matrix} \alpha_n \in \left\langle \frac{1}{2}, 1 \right\rangle \\ \beta_n = \frac{1-\alpha_n}{K_n - 1} \end{matrix} \tag{27}$$

where the parameters α_n and β_n are the smoothing parameters.

The kernel estimate (26) is formally a distribution mixture with uniform component weights equal to $\frac{1}{|S|}$ and, therefore, it could be used for the cluster analysis in the way described in the previous sections, without loosing the decision information during the mixture estimation.

5.1 Optimally Smoothed Kernel Estimate

Choosing the values of the smoothing parameters $\alpha_n, n \in \mathcal{N}$ seriously affects the quality of the estimate. We use a method based on log-likelihood cross-validation (Duin [7]) which is based on maximizing the log-likelihood function.

In order to avoid the trivial results ($\alpha_n = 1$) we use the following modification of the log-likelihood function:

$$L(S, \boldsymbol{\alpha}) = \sum_{\boldsymbol{x} \in S} \log \frac{1}{|S| - 1} \sum_{\boldsymbol{y} \in S, \boldsymbol{y} \neq \boldsymbol{x}} \prod_{n \in \mathcal{N}} g_n(x_n|y_n, \alpha_n) \tag{28}$$

As there is no general analytic solution we use the iterative method based on EM algorithm for maximizing the criterion (28). The EM algorithm modification is following

E-step:

$$q^{(t+1)}(\boldsymbol{y}|\boldsymbol{x}) = \frac{G(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\alpha}^{(t)})}{\sum_{\boldsymbol{z}\in S, \boldsymbol{z}!=\boldsymbol{x}} G(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{\alpha}^{(t)})} \tag{29}$$

M-step: (implicit form)

$$\boldsymbol{\alpha}^{(t+1)} = \arg\max_{\alpha} \Big\{ \sum_{\boldsymbol{x}\in S} \sum_{\boldsymbol{y}\in S, \boldsymbol{x}\neq\boldsymbol{y}} q^{(t+1)}(\boldsymbol{y}|\boldsymbol{x}) \log \frac{G(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\alpha})}{G(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\alpha}^{(t+1))}} \Big\} \tag{30}$$

which can be explicitly written as

$$\alpha_n^{(t+1)} = \frac{\sum_{\boldsymbol{x}\in S} \sum_{\boldsymbol{y}\in S, x_n=y_n, \boldsymbol{x}\neq\boldsymbol{y}} q^{(t+1)}(\boldsymbol{y}|\boldsymbol{x})}{\sum_{\boldsymbol{x}\in S} \sum_{\boldsymbol{y}\in S, \boldsymbol{x}\neq\boldsymbol{y}} q^{(t+1)}(\boldsymbol{y}|\boldsymbol{x})} \qquad n \in \mathcal{N} \tag{31}$$

6 Numerical examples Handwritten Non-stylized Numerals

In the example the proposed minimum information loss cluster analysis has been applied to classification of handwritten non-stylized numerals on a binary raster. We have used 400 000 numerals from the NIST database uniformly representing the classes 0,1,...,9. Each of the numerals in the data base has been normalized to a square 16×16 binary raster, i.e. it has been represented by a 256-dimensional binary vector.

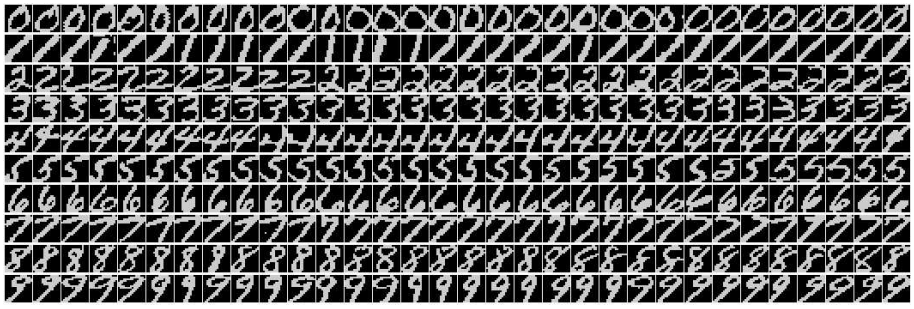


Figure 1: Examples of numerals from the NIST database normalized to the 16x16 binary raster.

Normally the NIST numerals are used as a benchmark problem for a supervised pattern recognition. The supervised classifier is trained for each class separately with the resulting relatively low classification error. Obviously, the non-supervised solution of the problem cannot be expected to achieve comparable accuracy, however, from the point of view of cluster analysis, we have again the possibility of a visual inspection of results.

Fig. 1 illustrates the properties of the NIST database. In the rows there are examples of numerals from the database. Again we have estimated the latent class model in the form of a 256-dimensional Bernoulli mixture. We have chosen a model of $M = 60$ components.

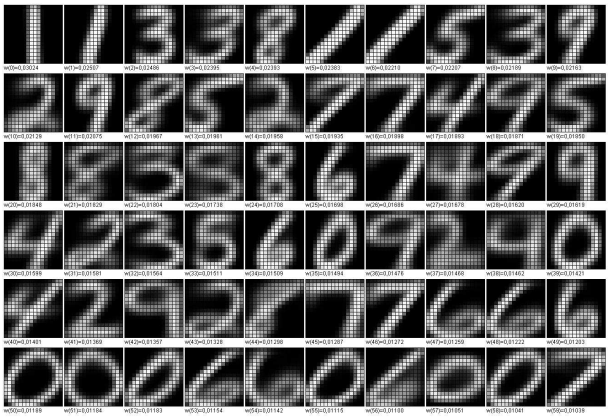
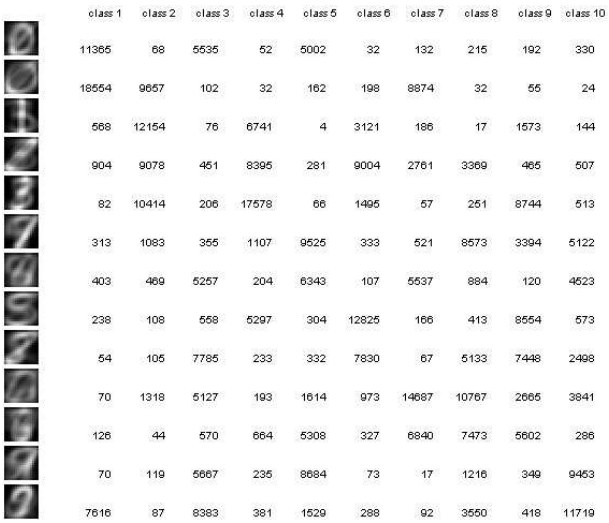


Figure 2: The component parameters (in square arrangement) of the mixture of 60 components estimated from the NIST database.



	class 1	class 2	class 3	class 4	class 5	class 6	class 7	class 8	class 9	class 10
	11365	68	5535	52	5002	32	132	215	192	330
	18554	9657	102	32	162	198	8874	32	55	24
	568	12154	76	6741	4	3121	186	17	1573	144
	904	9078	451	8395	281	9004	2761	3369	465	507
	82	10414	206	17578	66	1495	57	251	8744	513
	313	1083	355	1107	9525	333	521	8573	3394	5122
	403	469	5257	204	6343	107	5537	884	120	4523
	238	108	558	5297	304	12825	166	413	8554	573
	54	105	7785	233	332	7830	67	5133	7448	2498
	70	1318	5127	193	1614	973	14687	10767	2665	3841
	126	44	570	664	5308	327	6840	7473	5602	286
	70	119	5667	235	8684	73	17	1216	349	9453
	7616	87	8383	381	1529	288	92	3550	418	11719

Figure 3: Resulting cluster means for the final 13 clusters. The matrix illustrates the coincidence of the resulting clusters with the original classes.

The EM algorithm has been initialized randomly with the uniform component weights and stopped after 30 iterations and the estimated parameters θ_{mn} (in the raster arrangement) are shown in Fig. 2. The estimated elementary latent classes as characterized by the components in Fig. 2 have been unified sequentially by using the algorithm of Sec. 5. The hierarchical procedure based on pairwise unifying the most similar sub-mixtures has been stopped at the level of 13 clusters which precedes a local increase of the information loss Q_{ij}.

Fig. 3 describes the properties of the resulting clusters. The number of clusters is higher than 10 because for some numerals there are different variants which are too dissimilar in the high-dimensional description. The distribution of data vectors in the resulting clusters with respect to the true classes can be seen in the corresponding rows.

7 Conclusion

The latent class models have been used repeatedly as a tool of cluster analysis of multivariate categorical data since the standard approaches are usually not directly applicable. Unfortunately, the underlying discrete distribution mixtures with product components are not uniquely identifiable. In order to avoid the problem of identifiability the latent class model is applied only to identify elementary latent classes. We assume that the potential clusters can be constructed by unifying the elementary classes even if they are not defined uniquely. A hierarchical procedure is proposed to define the optimal decomposition of the underlying mixture or optimally smoothed kernel estimate. The hierarchical cluster analysis is controlled by minimum information loss criterion.

References

[1] Dempster, A.P., Laird, N.M. and Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. B*, **39** (1977) l-38

[2] Grim J.: EM cluster analysis for categorical data. In: *Structural, Syntactic and Statistical Pattern Recognition.* (Yeung D. Y., Kwok J. T., Fred A. eds.), (Lecture Notes in Computer Science. 4109). Springer, Berlin 2006, pp. 640-648.

[3] Gyllenberg M., Koski T., Reilink E., Verlaan M.: Non-uniqueness in probabilistic numerical identification of bacteria. *Journal of Applied Prob.*, **31** (1994) 542-548

[4] McLachlan G.J. and Peel D.: *Finite Mixture Models*, John Wiley & Sons, New York, Toronto, (2000)

[5] Lazarsfeld P.F., Henry N.W.: *Latent structure analysis.* Houghton Miflin, Boston (1968)

[6] Pearl J.: *Probabilistic reasoning in intelligence systems: networks of plausible inference.* Morgan-Kaufman, San Mateo, CA (1988)

[7] Duin, P.W.: On the choice of smoothing parameters for Parzen estimates of propability density functions. *IEEE Trans. on Computers*, (1994) C-25, No. 11, pp. 1175-1179

[8] Grim J., Novovičová J., Pudil P., Somol P., Ferri F. J.: Initializing normal mixtures of densities. *Proceedings of the 14th International Conference on Pattern Recognition* - (Jain, A.; Venkatesh, S.; Lovell, B.) IEEE, Los Alamitos 1998, pp. 886-890.

[9] Vermunt J.K., Magidson J.: Latent Class Cluster Analysis. In: *Advances in Latent Class Analysis*, (Eds. Hagenaars J.A., McCutcheon A.L., Cambridge University Press (2002)

[10] Grim J., Hora J.: Minimum Information Loss Cluster Analysis for Cathegorical Data. In: *Machine Learning and Data Mining in Pattern Recognition.* (Perner P. ed.), (Lecture Notes in Artificial Intelligence 4571). Springer, Berlin 2007, pp. 233 - 247.

Image Database Designed for Fast and Robust Image Search

Ondřej Horáček*

4th year of PGS, email: horacek@utia.cas.cz
Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU
advisor: Jan Flusser, Institute of Information Theory and Automation, AS CR

Abstract. Image retrieval deals with a problem of finding similar pictures in image database. Our task is to find originals of modified images, typically stolen and republished on the web. Our problem is specific in terms of the database size (millions of photos), demanded speed of the search (seconds), and unknown image modifications (loss of quality, radiometric degradation, crop, etc.). Proposed method works in the following tree steps: 1. Image preprocess – normalization for robustness to the modifications. 2. Retrieval of candidates from the database index – stochastic decision in each vertex of the index tree is used to find the most relevant candidates. 3. Verification of the candidates – modified phase correlation is used. The method was implemented in practice with very good results. Based on wide experiments, it was shown that the success rate of the search depends on the level of image modification.

Abstrakt. Image retrieval se zabývá vyhledáním snímků v obrazové databázi na základě určité podobnosti. Naším úkolem je vyhledat v databázi originály snímků dodatečně upravených, konkrétně neoprávněně publikovaných na webu. Tato úloha je specifická velikostí databáze (miliony snímků), požadovanou rychlostí odezvy (sekundy) a předem neznámým poškozením snímku (ztráta kvality, radiometrické poškození, ořez atd.). Předkládaná metoda pracuje v následujících třech krocích: 1. Předzpracování snímku – normalizací je zajištěna odolnost vůči změnám. 2. Vyhledání kandidátů v indexu databáze – díky stochastickému rozhodování v indexovém stromu databáze jsou nalezeny nejpravděpodobnější kandidáti. 3. Ověření kandidátů – používáme modifikovanou fázovou korelaci. Metoda již byla implementována a dosahuje velmi dobrých výsledků. Na základě různorodých experimentů je ukázáno, že úspěšnost vyhledávání závisí nejvíce na míře modifikací snímku.

1 Introduction

Large image databases are often run on a commercial basis – browsing through and viewing images is free of charge while downloading and re-using them on your webpages and articles is a subject of a fee. However, some users republish the downloaded images without paying the fee, which is a violation of copyright law. The copyright owner thus wants to regularly scan suspicious domains or websites to check if there are any unauthorized copies of the database images.

This paper describes an original method which we developed for an international advertising and press company. This company runs a database of more than 10 millions

*The author thanks his colleagues Jakub Bican and Jan Kamenický for their cooperation and consultations.

photographs updated everyday. They estimate hundreds thousands images being used without permission on the web. Detection of illegal copies is complicated by two principal difficulties – the unauthorized images are usually modified before they are post on the web and the response of the system must be extremely fast because of an enormous number of database images. Although this problem formulation looks like an image retrieval task, this is not the case. In traditional image retrieval, we want to find in the database all *similar* images to the query image, where similarity is evaluated by colors, textures, content, etc. Here we want to identify only the *equivalent* images to the query (we call this task *image identification*. This is why we cannot apply most of standard image retrieval techniques. By the term "equivalent images" we understand any pair of images which differ from one another by the following transformations.

- Quality reduction. Either compression to different image format or resize changes the image representation, although the image seem very similar to human eye (in Fig. 1b).
- Radiometric and color distortions. We consider changes of image brightness and contrast (in Fig. 1c), changes of color tone or conversion of the image to gray-scale (in Fig. 1d).
- Crop of the image. Image part can be cropped from the background, still we consider that the major part of the image is preserved. Also, a frame can be added to the image or aspect ratio can be changed (in Fig. 1e).
- Local changes. A logo can be added to the image, or a thin label can go throw the image (in Fig. 1f).
- Combinations. Reasonable combination of distortions mentioned above is also considered. However, their increasing amount and significance will surely impact the algorithm results (in Fig. 1g).

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

Figure 1: Possible modifications of the query image. The first image is the original stored in the database.

We decided to include the above transforms into our "equivalence relation" because, according to the earlier statistics performed by the company, they are frequently present in unauthorized copies.

It is not possible to use directly any of the existing methods. Probably the closest published method is by Obdržálek et al. in [4]. It is suitable for recognition of man-made objects with partially planar surface, such as goods in a supermarket. It can handle very general situations including partial occlusion and affine transform of the objects but this is useless for our purpose. Our method is different is several aspects. We can not use several maximum extremal regions per image, because on our database images they may not exist. Our index tree is thousand times bigger than that one considered in [4], so we need to use different, space-efficient tree structure. Finally, we include another independent image comparison to eliminate false positives matches.

We present an original image identification method, which is based on a hierarchical structure of the database, representation of the images by proper invariant features, and a fast tree-searching algorithm. Our method has got very good identification rate in a reasonable response time. In this article, we present main idea of the method as well as selected details.

2 Algorithm outline

A kind of binary decision tree is used for the database indexing. Some image features are needed to characterize the image in the index. We require the features to be robust to considered degradations, stable, but mainly to be extensible and discriminate enough for any database size. We use image intensities in various pixel positions, surely after dealing with the image distortions. This choice is simple in principle, but we found it effective in presented method. Our image identification works in these three steps:

1. Normalization. Robustness to the modification is ensured by normalization of the images during a process. In other words: Each of the considered modification corresponds with a change of an evaluative image quality. We apply the modification once again in an amount, that was established to set the qualities to the same value for all the images. This process annuls the impact of considered modifications.

2. Stochastic index. The database images are organized in binary decision tree. Thus, we obtain several candidates for match with a query image very quickly. Decisions in the tree are based simply on image intensity at certain position. The position and threshold are set for each vertex during build of tree. For individual query image, stability of decisions in the tree is evaluated. We alternate the unstable decisions during the image identification. So, we get many candidates per query.

3. Candidate verification. Edge information of the image serves for the the final comparison of a candidate with the query image. More concretely, phase correlation restricted to low-pass fourier transform is used.

3 Normalization

Both the query image and the database images are preprocessed previously. We apply some normalizations to make the images invariant to considered modifications. First, we

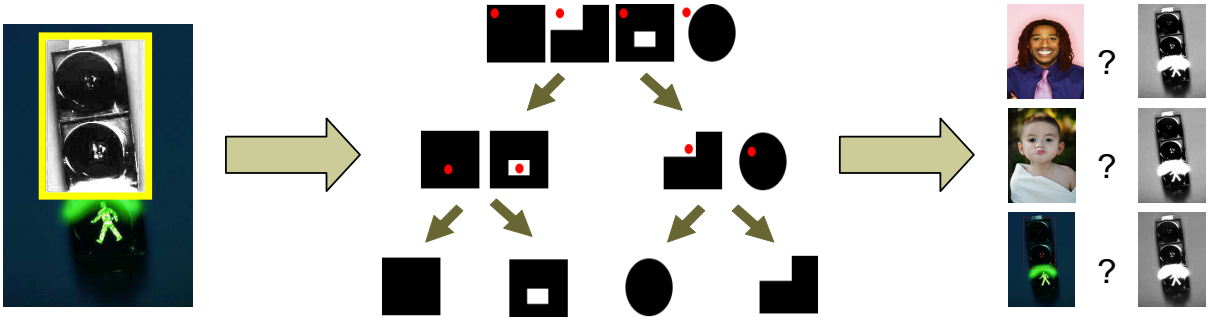


Figure 2: Schema of the image identification. First, normalization of the query image is done. Then, the most probable candidates are found in the index tree. And finally, the candidates are verified by modified phase correlation.

convert the images to gray-scale to handle color distortions. Then, the lowest and the highest 10 percent image histogram fraction are found and their centers of gravity are set to fixed values (by addition / multiplication of the image intensity values). This ensures image invariance to brightness and contrast changes.

The image crop is tougher modification to handle. The first idea could be to use features invariant to crop, such as corners (found for example by Harris corner detector [1]) or intersection points. We do not use them because it is not possible to stably match (identify) those points for all the database images after crop. We consider that the crop preserves major part of the image. For most of the images, the part is separable from the background by color. So, we segment the major part of the image and bound it by a crop invariant frame.

We introduce a special color-based segmentation for separation of the image major part. The algorithm finds a frame in the image which fulfill following requirements: it is big enough, ir does not lay on the image border, it is color specific and it as stable as possible. It was developed heuristically with respect to experiment results. Our algorithm works in principle as follows: In principle, it divides the image into blocks, computes the block color character, and finds neighborhood blocks with the same character for each possible starting block. The segmented region is broadly bounded by box, which we call a frame. Stability and "quality" of a the frame is evaluated. For the database images, just area bounded by the best frame is used for image identification in the index tree. It is reasonable to consider that this frame will be found in the modified image as one of the best, too. Therefore we use the best five variants of the frame for the image search.

4 Stochastic index

The task for the index is to retrieve image from the database quickly. Response time to a query must be principally shorter than proportional to the database size, which goes to millions of images. Input for the index is a query image that have been normalized and blurred the same way as the database images were. They should be very similar, but still, we need to evaluate stability of each decision in the index tree to be robust to

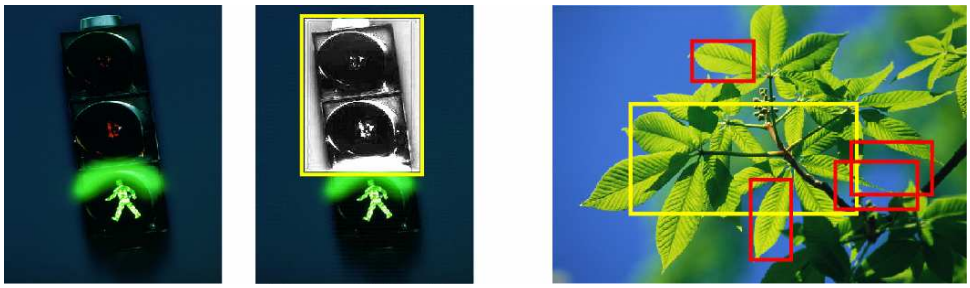


Figure 3: Image preprocess. The original is converted to gray-scale, crop-invariant frame is found and its brightness and contrast are normalized. On the right, multiple invariant frames used for the image identification are shown.

minor image changes.

The database images are organized in a binary decision tree, commonly used to handle huge amount of data (a survey is done in [3]). Decision in the tree are based on intensities of image pixels. The pixel (threshold) position is taken relatively to the valid image area (frame). Once we have two different branches of the tree, each of them can contain images differing mainly (and therefore the most stably) in a different image part. Such a threshold position is established during the tree build. So, each node of the tree contains threshold value and relative position of the threshold pixel.

Now we describe the search for image in the index tree. In a node, pixel of the query image is compared with the threshold value and its sub-branch (next node) is chosen. We establish stability of the comparison as a likelihood, that the image belong to the same (left/right) branch even after following image distortions. We assume that the threshold pixel can change its intensity (with uniform distribution in certain intensity interval) and it can change its position (the miss-place has a two-dimensional gaussian distribution) (in Fig. 4). This stochastic model is similar to the one presented by Obdržálek [4], but their usage of the model is different.

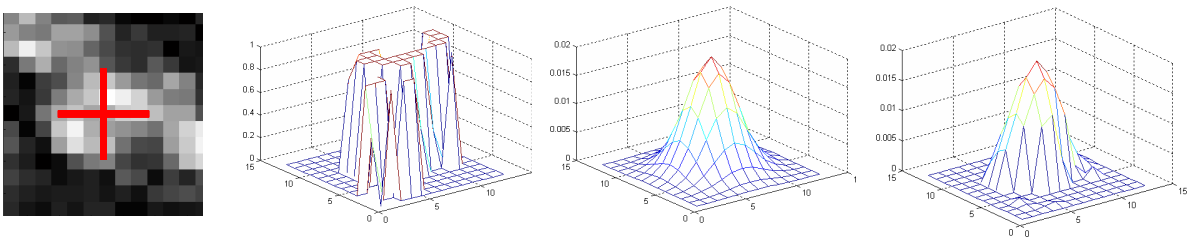


Figure 4: Stability of decision in the tree. The image and the threshold position, likelihood that a pixel is above the threshold, likelihood of threshold position change, and their multiplication.

This way we go throw the index tree till we reach a leaf containing an image – candidate for the match. Now, we select node with the least stable decision. Now we continue throw the alternate branch and get an other match candidate. This way we found 20 images

from the index with the highes probability to match the query image. Note, that some candidates are found even for totally dissimilar query image.

It was mentioned above that the index tree needs to be prepared previously. Because of the huge number of images, the tree is build gradually. Our algorithm works in principle as follows:

1. Read normalized version of a database image from a disk one by one.
2. Find a leaf in the partially builded tree for the image and register the image to the leaf.
3. If the leaf contains enough images, convert the leaf to inner tree node and redistribute its images as follows:
   (a) Choose randomly several positions of a threshold.
   (b) For each threshold candidate, get intensity values at the processed position (for all the images belonging to the leaf).
   (c) For the threshold candidate, compute the threshold value as a median of the intensity values.
   (d) For the threshold candidate, compute its stability as a sum of decision stabilities for all the images belonging to the leaf. It means, evaluate the likelihood described in the paragraph about search (and in Fig. 4).
   (e) Choose the most stable threshold of the threshold candidates and save its position and the processed node (leaf).
   (f) Create left and right sub-node of the processed node. Based on the new threshold, redistribute the images of the processed node to them.
4. After all the database images have been processed by previous steps and are registered to some leaf, divide the leafs till they contain only one image. (Do it the same way as in steps 3a to 3f).

5 Rest of the method

In the last step, the candidate images are compared one by one with the query image. We use modified phase correlation (originally introduced by Kuglin [2] in 1975). Phase correlation is robust to overlap, shift and radiometric degradation. We restrict the correlation only to low-pass of the fourier spectrum to make the comparison more stable for image quality changes (in Fig. 5).

For better performance in practical experiments, many improvements have been made on the basic method scope described above. The algorithm use several parameters and options, that affect both the identification rate and algorithm speed. There is another tradeoff with the normalization: the more quantities are normalized, the more information is lost and it is harder to identify the original image; but it makes it possible to identify images with harder combination of distortions. We overcome this situation by usage of more than one versions of preprocess (with different normalization and different parameter

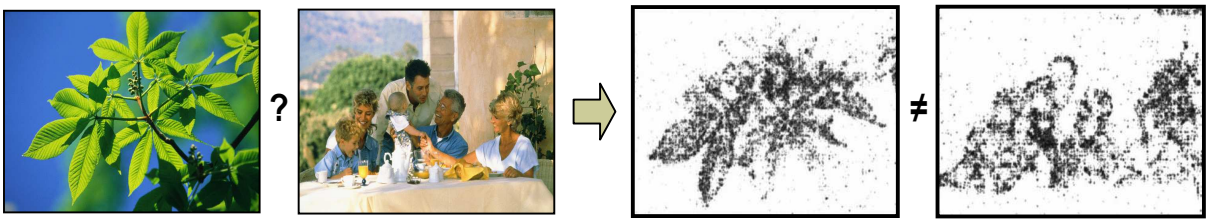


Figure 5: Candidate verification. Low-pass of phase correlation bases the image comparison on their major edges.

set). And, of course, more than one index tree. We use two independent index trees in the prototype implementation: one is build from images with normalized histogram, the second is build and searches for the image parts bounded by the invariant frames.

The method is being implemented as a configurable framework. A structured configuration file controls application of zero to several normalization algorithms per image, build and search in several index trees as well as it contains all parameters for all the algorithms. (Note, that some simple preprocess algorithms, such as mirroring or blurring, are not described in this article). We have also implemented implemented optional index ability to work with color images. In that case, a threshold value is replaced by a plain in RGB color space.

6 Experiment results

Prototype of the proposed method was implemented in Matlab. Tests have been done on 100 000 image database. First, a thousand of query images was generated for each considered modification (the images are still "equivalent", see samples on Fig. 1). Strength of the modification is varying around the level expected from the practice. The identification ratio is very good, better than we expected. Original images are found successfully in 99.5 % of cases. The rest 0.5 % are cases, when the database contained a very-similar image (e.g. with some retouching), that has been identified before the original one. For typical automatically republished images, identification rate is more than 90 %, which is very good for practical use of the method. The identification rate surely decrease for harder modifications, but even for combinations of radiometric degradations, crop and logo is still about 20%. See table 1 for more details.

Response speed of the Matlab prototype is up to 20 seconds per image. Image retrieval from the index tree takes about 0.2 second, rest belong to the one by one candidate verification by the fourier transform. Build of the index tree takes less than a second per image. But, the database build can take several days with some non-trivial normalization (such as invariant frame – about 3 seconds per image). Overall, the method speed as well as the identification rate depends on appropriate set of parameters.

Degradation	True positives	False positives
Original	99.5 %	0.4 %
Logo added	98.2 %	0.2 %
Scale	94.6 %	0.4 %
Brightness and contrast	71.2 %	0.7 %
Crop	45.0 %	0.2 %
Scale + logo	93.4 %	0.2 %
Scale + logo + frame	35.8 %	0.4 %
Radiometric deg. + crop + logo	18.2 %	0.5 %
Not in the database	0.0 %	0.3 %

Table 1: Identification rate on 100000–image database. The identification rate depends on kind of image modification.

7 Conclusion

In this article, we presented our method for modified image identification. The task is specific by character of the modifications, the database size, and required response speed. The method is novel in normalization during the image preprocessing (invariant frame, normalization) and in stochastic backtracking throw the image index. It was shown in experiments on huge database that the method performs very well. It is ready to catch majority of illegally republished database images on scanned web sites.

References

[1] C. Harris and M. J. Stephens. A combined corner and edge detector. In Alvey Conference, pages 147–152, 1988.

[2] C. D. Kuglin and D. C. Hines. The phase correlation image alignment method. Assorted Conferences and Workshops, pages 163–165, September 1975.

[3] Sreerama K. Murthy. Automatic construction of decision trees from data: A multi-disciplinary survey. Data Min. Knowl. Discov, 2(4):345–389, 1998.

[4] Obdržálek, Štěpán and Matas, Jiří. Sub-linear indexing for large scale objec recognition. In WF Clocksin, AW Fitzgibbon, and PHS Torr, editors, BMVC 2005: Proceedings of the 16th British Machine Vision Conference, volume 1, pages 1–10, London, UK, September 2005. BMVA.

Classical Particle and Time-periodic Aharonov-Bohm Flux

Tomáš Kalvoda

2nd year of PGS, email: tom.kalvoda@gmail.com
Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU
advisor: Pavel Šťovíček, Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU

Abstract. The behaviour of a classical charged particle confined in a plane, under influence of homogeneous magnetic field and time-periodic Aharonov-Bohm flux is studied. At first the canonical transformation to action-angle coordinates is constructed. Then the resonant effect between the strength of the magnetic field and the frequency of the Aharonov-Bohm flux is studied by the means of averaging method. The result is demonstrated on particular example and the numerical solution of the original problem is compared with the analytical result obtained with the help of averaged system.

Abstrakt. Tento příspěvek ze zabývá chováním klasické nabité částice pohybující se v rovině, na niž působí homogenní magnetické pole na tuto rovinu kolmé a periodicky časově závislý Aharonov-Bohmův tok. Nejprve je sestrojena kanonická transformace do proměnných akce-úhel. Poté je studován rezonantní efekt mezi sílou magnetického pole a frekvencí Aharonova-Bohmova toku metodou středování. Výsledek je demonstrován na konkrétním případě porovnáním numerického řešení původního problému s analytickým výsledkem získaným z vystředovaného systému.

1 Introduction

We are interested in qualitative behaviour of a classical charged particle in a plane influenced by a homogeneous magnetic field perpendicular to the plane, and time-periodic Aharonov-Bohm flux. The system is studied from the viewpoint of nonrelativistic classical mechanics.

Let the Cartesian coordinates in the plane be denoted by $q = (q_1, q_2) \in \mathbb{R}^2$. Suppose that there is a particle with mass m and charge e confined to this plane. The vector potential A consists of two parts. The homogeneous magnetic field of strength $b > 0$ perpendicular to the q-plane is generated by the potential

$$A_h(q) = \frac{-b}{2} q^\perp,$$

where $q^\perp = (-q_2, q_1)$. The second part corresponds to the Aharonov-Bohm flux $\Phi(t)$ located in the origin of the coordinate system and is given by

$$A_{ab}(q,t) = \frac{\Phi(t)}{2\pi|q|} q^\perp.$$

This term contains a singularity in the origin, the phase space $(\mathbb{R}^2 \setminus \{0\}) \times \mathbb{R}^2$ is not simply connected. It is assumed that the flux Φ is real valued, continuously differentiable, and periodic function of real variable. However, until Section 3 the periodicity is not important. Introduce the polar coordinates $(r, \theta) \in (0, \infty) \times (0, 2\pi)$ by

$$q_1 = r\cos\theta, \quad q_2 = r\sin\theta.$$

The Hamilton's function of the system then reads

$$H(r, \theta, p_r, p_\theta, t) = \frac{1}{2m}\left(p_r^2 + \left(\frac{p_\theta - \frac{e\Phi(t)}{2\pi}}{r} + \frac{eb}{2}r\right)^2\right).$$

The phase space is $(\mathbb{R}^+ \times S_1) \times \mathbb{R}^2$. Obviously the coordinate θ is cyclic, i.e. $\dot{p}_\theta = 0$. Therefore p_θ is an integral of motion (in fact, it is the angular momentum). Thus the system has effectively only one degree of freedom. From Hamiltonian equations of motion it follows that the radial motion of the particle is governed by the equation

$$\ddot{r} + \frac{e^2b^2}{4m^2}r = \frac{\left(p_\theta - \frac{e\Phi(t)}{2\pi}\right)^2}{mr^3}.$$

In the following sections we will investigate the behaviour of solution of this equation in the resonant situation.

2 Transformation to Action-Angle Coordinates

Let us begin with the construction of the action-angle coordinates (for more details confer [2]). In order to simplify the expressions we set the charge and mass equal to one, $e = m = 1$. Note that the Hamilton's function of radial motion is

$$H(r, p, t) = \frac{1}{2}\left(p_r^2 + \left(\frac{a(t)}{r} + \frac{b}{2}r\right)^2\right), \tag{1}$$

where $a(t) = p_\theta - \Phi(t)/2\pi$. Denote

$$V(r) = \frac{1}{2}\left(\frac{a}{r} + \frac{br}{2}\right)^2.$$

The minimum of V for $r > 0$ is

$$V_{min} = \min_{r>0} V(r) = \begin{cases} V\left(\sqrt{\frac{2a}{b}}\right) = ab & a > 0, \\ V\left(\sqrt{\frac{2|a|}{b}}\right) = 0 & a < 0. \end{cases}$$

Now we will construct action-angle coordinates in case when $a(t) = a$ is constant, i.e. our Hamiltonian is independent of time[1]. For a fixed energy level $E > V_{min}$ the motion

[1] This is the case when there is no Aharonov-Bohm flux.

is constrained to the interval $[r_+, r_-]$. These constraints are obtained as a solution of equation $V(r) = E$. Thus we have

$$E - V(r) = \frac{b^2}{8r^2}(r_+^2 - r^2)(r^2 - r_-^2),$$

where

$$r_\pm^2 = \frac{2}{b^2}\left(2E - ab \pm \sqrt{(2E-ab)^2 - a^2b^2}\right).$$

The action is defined by integral

$$\begin{aligned} I(E) &= \frac{1}{\pi}\int_{r_-}^{r_+} \sqrt{2(E - V(r))}\mathrm{d}r = \frac{b}{4\pi}\int_{r_-^2}^{r_+^2} \frac{1}{x}\sqrt{(r_+^2 - x)(x - r_-^2)}\ \mathrm{d}x = \\ &= \frac{b}{8}(r_+ - r_-)^2 = \frac{1}{b}(E - \vartheta(a)ab) = \frac{1}{b}(E - V_{min}). \end{aligned}$$

Generating function of the transformation reads

$$\begin{aligned} S(r, I) &= \int_{r_-}^{r} \sqrt{2(E - V(\rho))}\mathrm{d}\rho = \frac{b}{2}\int_{r_-}^{r} \frac{1}{\rho}\sqrt{(r_+^2 - \rho^2)(\rho^2 - r_-^2)}\mathrm{d}\rho = \\ &= \frac{b}{4}\int_{r_-^2}^{r^2} \frac{1}{x}\sqrt{(r_+^2 - x)(x - r_-^2)}\ \mathrm{d}x. \end{aligned}$$

This integral can be evaluated explicitly, and after some minor adjustments one obtains the expression

$$\begin{aligned} S(r, I) = \frac{1}{4}\sqrt{8bIr^2 - (br^2 - 2|a|)^2} - I\arctan\left(\frac{4I - br^2 + 2|a|}{\sqrt{8bIr^2 - (br^2 - 2|a|)^2}}\right) - \\ - \frac{|a|}{2}\arctan\left(\frac{(br^2 + 2|a|)\sqrt{8bIr^2 - (br^2 - 2|a|)^2}}{b^2r^4 - 4bIr^2 + 4|a|^2}\right). \end{aligned}$$

The induced transformation of variables $(r, p) = \Psi(\varphi, I)$ is defined as follows: $\Psi = F \circ G^{-1}$, where the transformations $(r, p) = F(u, v)$ and $(\varphi, I) = G(u, v)$ are given respectively by the relations

$$r = u,\ p = \frac{\partial S(u, v)}{\partial u} \quad \text{and} \quad \varphi = \frac{\partial S(u, v)}{\partial v},\ I = v.$$

By direct computation we get

$$r = \frac{2}{\sqrt{b}}\sqrt{I + \frac{|a|}{2} + \sqrt{I(I + |a|)}\sin\varphi}, \quad p = \frac{\sqrt{bI(I + |a|)}\cos\varphi}{\sqrt{I + \frac{|a|}{2} + \sqrt{I(I + |a|)}\sin\varphi}},$$

and conversely,

$$\varphi = -\arctan\left(\frac{1}{bpr}\left(p^2 + \frac{a^2}{r^2} - \frac{b^2r^2}{4}\right)\right), \quad I = \frac{1}{b}(H - V_{min}) = \frac{1}{2b}\left(p^2 + \left(\frac{|a|}{r} - \frac{br}{2}\right)^2\right).$$

Let us switch to the time-dependent case with a Hamiltonian $H(r,p,t)$. Seeking the action-angle variables for the frozen Hamiltonian at each moment of time one in fact constructs a time-dependent transformation of variables. Hence the generating function of the transformation, $S(u,v,t)$, is time-dependent as well. One arrives again at a Hamiltonian system with a Hamiltonian $K(\varphi,I,t)$ and it holds

$$K(\varphi,I,t) = H(\Psi(\varphi,I,t),t) + \left.\frac{\partial S(u,I,t)}{\partial t}\right|_{u=\Psi_r(\varphi,I,t)},$$

where Ψ_r denotes component of Ψ belonging to r. Our Hamiltonian depends on time t only through function $a(t)$, cf. (1). New Hamiltonian now reads

$$K(\varphi,I,t) = bI - \arctan\left(\frac{\sqrt{I}\cos\varphi}{\sqrt{I+|a(t)|}+\sqrt{I}\sin\varphi}\right)\dot{a}(t)\mathrm{sgn}a(t)$$

Finally the Hamiltonian equations of motion are

$$\dot{\varphi} = b - \frac{a\dot{a}}{2}\frac{\cos\varphi}{\sqrt{I(I+|a|)}}\frac{1}{2I+|a|+2\sqrt{I(I+|a|)}\sin\varphi}, \tag{2}$$

$$\dot{I} = -\frac{\mathrm{sgn}a}{2}\left(\dot{a} - \frac{|a|\dot{a}}{2I+|a|+2\sqrt{I(I+|a|)}\sin\varphi}\right). \tag{3}$$

3 The Averaged System

Henceforth assume that $a(t) \neq 0$. Further simplification of (2) and (3) is achived by passing to the coordinates ϕ and G given by

$$G = \frac{2I}{|a|} + 1, \quad \phi = \varphi - bt. \tag{4}$$

Obviously $(\phi, G) \in (1,\infty)\times[0,2\pi)$. Hence

$$\dot{G} = \frac{\dot{a}}{a}\left(\frac{1}{G}\frac{1}{1+\sqrt{1-1/G^2}\sin(bt+\phi)} - G\right), \tag{5}$$

$$\dot{\phi} = -\frac{\dot{a}}{a}\frac{\cos(bt+\phi)}{G\sqrt{G^2-1}}\frac{1}{1+\sqrt{1-1/G^2}\sin(bt+\phi)}. \tag{6}$$

Denote $A(t) = \dot{a}(t)/a(t)$. From the assumptions laden on the flux Φ it follows that the Fourier series

$$A(t) = \frac{1}{\sqrt{2\pi/\Omega}}\sum_{n\in\mathbb{Z}} A_n e^{in\Omega t}, \quad A_n = \frac{1}{\sqrt{2\pi/\Omega}}\int_0^{\frac{2\pi}{\Omega}} A(t)e^{-in\Omega t}\,\mathrm{d}t,$$

is uniformly convergent on $\mathbb{R}$. Since $A(t)$ is real it holds $\overline{A_n} = A_{-n}$, moreover it is true that $A_0 = 0$.

Now suppose that $\Omega/b = p/q$, where p, q are coprime natural numbers. Then the right hand sides of (5) and (6) are $2\pi p/\Omega$-periodic with respect to the time t. The averaged system is obtained by computing the time average of the right hand sides of (5) and (6). Further, the averaging principle[2] states, that the solution of the averaged system might be a good approximation to the original system provided the $A(t)$ is small in some sense. The averaged system is given by

$$\dot{G} = \frac{1}{\frac{2\pi}{\Omega}p} \underbrace{\int_0^{\frac{2\pi}{\Omega}p} A(t) \left(\frac{1}{G} \frac{1}{1 + \sqrt{1 - 1/G^2}\sin(bt+\phi)} - G \right) \mathrm{d}t}_{I_1},$$

$$\dot{\phi} = -\frac{1/G^2}{\sqrt{1-1/G^2}} \frac{1}{\frac{2\pi}{\Omega}p} \underbrace{\int_0^{\frac{2\pi}{\Omega}p} A(t) \frac{\cos(bt+\phi)}{1 + \sqrt{1 - 1/G^2}\sin(bt+\phi)} \mathrm{d}t}_{I_2},$$

where we treat G and ϕ as constants. We also denote $\beta = \sqrt{1-1/G^2}$ and keep in mind that $0 < \beta < 1$. Putting $\tau = bt + \phi$ in integrals we arrive to

$$I_1 = \frac{1/b}{\sqrt{2\pi/\Omega}} \frac{1}{G} \sum_{n\in\mathbb{N}} \int_0^{2\pi q} \left[\frac{(A_n + A_{-n})\cos\frac{np}{q}(\tau-\phi) + i(A_n - A_{-n})\sin\frac{np}{q}(\tau-\phi)}{1+\beta\sin\tau} \right] \mathrm{d}\tau,$$

$$I_2 = \frac{1/b}{\sqrt{2\pi/\Omega}} \sum_{n\in\mathbb{N}} \int_0^{2\pi q} \left[\frac{(A_n + A_{-n})\cos\frac{np}{q}(\tau-\phi) + i(A_n - A_{-n})\sin\frac{np}{q}(\tau-\phi)}{1+\beta\sin\tau} \right] \cos\tau \mathrm{d}\tau.$$

Now notice that integrals

$$\int_0^{2\pi q} \frac{\cos\frac{n}{q}\tau}{1+\beta\sin\tau}\mathrm{d}\tau = \int_0^{2\pi q} \frac{\sin\frac{n}{q}\tau}{1+\beta\sin\tau}\mathrm{d}\tau = \int_0^{2\pi q} \frac{\cos\tau\cos\frac{n}{q}\tau}{1+\beta\sin\tau}\mathrm{d}\tau = \int_0^{2\pi q} \frac{\cos\tau\sin\frac{n}{q}\tau}{1+\beta\sin\tau}\mathrm{d}\tau$$

are equal to zero, if $\frac{n}{q} \notin \mathbb{N}$. This can be seen by dividing the domain of integration to pieces of length 2π, shifting all the domains to $(0, 2\pi)$ and summing the integrands. Using relations from Section 4 we finally arrive at

$$\dot{G} = \frac{2}{\sqrt{2\pi/\Omega}} \sum_{\substack{n\in\mathbb{N} \\ (\tilde{n}\in\mathbb{N})}} \Im\left[iA_n e^{-i\tilde{n}(\phi+\pi/2)} \right] \left(\frac{G-1}{G+1} \right)^{\tilde{n}/2},$$

$$\dot{\phi} = \frac{2}{\sqrt{2\pi/\Omega}} \frac{1}{G^2-1} \sum_{\substack{n\in\mathbb{N} \\ (\tilde{n}\in\mathbb{N})}} \Im\left[A_n e^{-i\tilde{n}(\phi+\pi/2)} \right] \left(\frac{G-1}{G+1} \right)^{\tilde{n}/2},$$

where $\tilde{n} = \frac{np}{q}$ and $\Im(c)$ is an imaginary part of a complex number c. This is certainly Hamiltonian system, with Hamilton's function given by

$$\overline{H}(\phi, G) = \frac{2}{\sqrt{2\pi/\Omega}} \sum_{\substack{n\in\mathbb{N} \\ (\tilde{n}\in\mathbb{N})}} \frac{1}{\tilde{n}} \Im\left[A_n e^{-i\tilde{n}(\phi+\pi/2)} \right] \left(\frac{G-1}{G+1} \right)^{\tilde{n}/2}.$$

Note that this series is uniformly convergent.

[2]For more details consult [3] or [4]. The applicability of the averaging principle is an open problem that will not be discussed here. The numerical computation is employed to check our results. See below.

3.1 Example

As an example we choose $a(t) = \gamma + \varepsilon \sin \Omega t$, where $0 < \varepsilon < \gamma$. Or, in other words, we take the flux $\Phi(t) = -2\pi\varepsilon \sin \Omega t$. It can be computed (with the aid of relations from Section 4), that

$$A_n = \sqrt{2\pi\Omega}\; e^{i\frac{\pi}{2}(n-1)} \left(\frac{\varepsilon/\gamma}{1+\sqrt{1-(\varepsilon/\gamma)^2}} \right)^n, \; n \in \mathbb{N}.$$

Therefore

$$\overline{H}(\phi, G) = \frac{2\Omega}{p} \Im \left\{ -i \sum_{m=1}^{\infty} \frac{1}{m} \left[\underbrace{e^{i\frac{\pi}{2}(q-p)-ip\phi} \left(\frac{\varepsilon/\gamma}{1+\sqrt{1-(\varepsilon/\gamma)^2}} \right)^q \left(\frac{G-1}{G+1} \right)^{p/2}}_{z(\phi,G)} \right]^m \right\}.$$

This series can be easily summed with the aid of geometric series, the desired result is

$$\overline{H}(\phi, G) = \frac{2\Omega}{p} \ln |1 - z(\phi, G)|.$$

Since this is two dimensional time-independent Hamiltonian system, it is easy to compute invariant curves, i.e. trajectories in the phase space. However, it turns out that these solutions are good approximation to the original problem only if $p/q \in \mathbb{N}$. For the sake of simplicity we confine ourselves to the case $\Omega = b = p = q = 1$. Thus

$$\overline{H}(\phi, G) = 2 \ln \left| 1 - e^{-i\phi} \underbrace{\frac{\varepsilon/\gamma}{1+\sqrt{1-\varepsilon^2/\gamma^2}}}_{\beta} \sqrt{\frac{G-1}{G+1}} \right|. \tag{7}$$

Obviously $0 < \beta < 1$. Since the Hamiltonian is time independent, it is conserved during the time evolution. The equation

$$\overline{H}(\phi, G) = h \in (-\infty, \ln 4)$$

defines implicitly G as a function of ϕ. It is straightforward to compute these invariant curves, on the other hand the result is not so nice. One must treat cases $0 < h < \ln 4$ and $h \leq 0$ separately. In the latter case, the curve must be sometimes stitched from two pieces. Summary of the results follows:

- $h < 0 < \ln 4$:

$$G(\phi) = \frac{\beta^2 + (\cos\phi + \sqrt{\cos^2\phi - 1 + e^h})^2}{\beta^2 - (\cos\phi + \sqrt{\cos^2\phi - 1 + e^h})^2} \tag{8}$$

  if $\phi \in [0, 2\pi)$ such as $\cos\phi < 1 - \frac{e^h}{2}$ and $\cos\phi + \sqrt{\cos^2\phi - 1 + e^h} < \beta$.

- $-\infty < h \leq 0$:

$$G(\phi) = \frac{\beta^2 + (\cos\phi + \sqrt{\cos^2\phi - 1 + e^h})^2}{\beta^2 - (\cos\phi + \sqrt{\cos^2\phi - 1 + e^h})^2} \tag{9}$$

if $\phi \in [0, 2\pi)$ such as $\sqrt{1-e^h} \leq \cos\phi < 1 - \frac{e^h}{2}$ and $\cos\phi + \sqrt{\cos^2\phi - 1 + e^h} < \beta$.

$$G(\phi) = \frac{\beta^2 + (\cos\phi - \sqrt{\cos^2\phi - 1 + e^h})^2}{\beta^2 - (\cos\phi - \sqrt{\cos^2\phi - 1 + e^h})^2} \tag{10}$$

if $\phi \in [0, 2\pi)$ such as $\sqrt{1-e^h} < \cos\phi$ and $\cos\phi - \sqrt{\cos^2\phi - 1 + e^h} < \beta$.

These invariant curves are depicted in Figure 1. Note that if $h = 0$ then the curve hits line $G = 1$. It seems that this is a pathological feature of the averaged system, i.e. the original system does not posses such behaviour. At least it is not verified by the numerical computation. We see that ϕ tends to constant and G escapes to infinity. This means that (cf. (2) and (4)) in the course of time the particle will get arbitrarily far from and close to the origin.

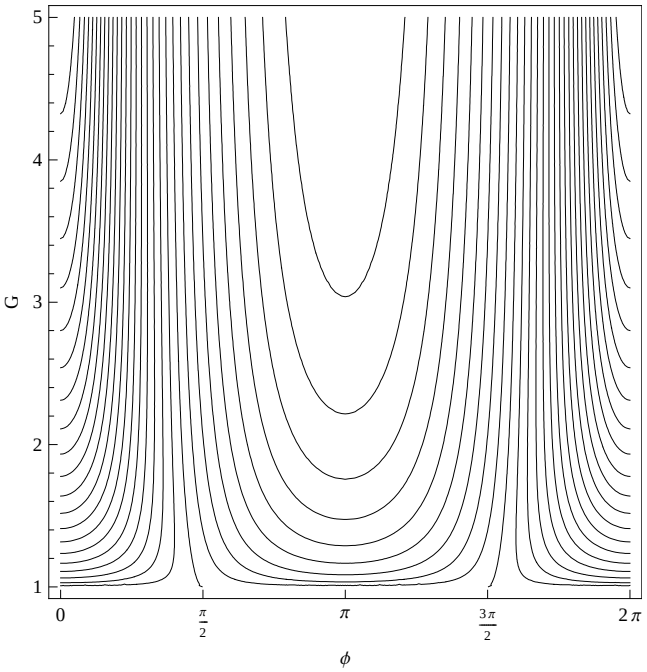


Figure 1: Invariant curves of the averaged Hamiltonian (7) given by formulae (8), (9), and (10). Parameter β is equal to 0.8.

We compare the numerical solution of (5) and (6) (where we use special $a(t)$ and values of parameters mentioned at the beginning of this subsection) with the invariant curves computed above. The value of h is computed from the initial conditions ϕ_0 and G_0. Choice of ε and γ is noted above each frame in Figure 2.

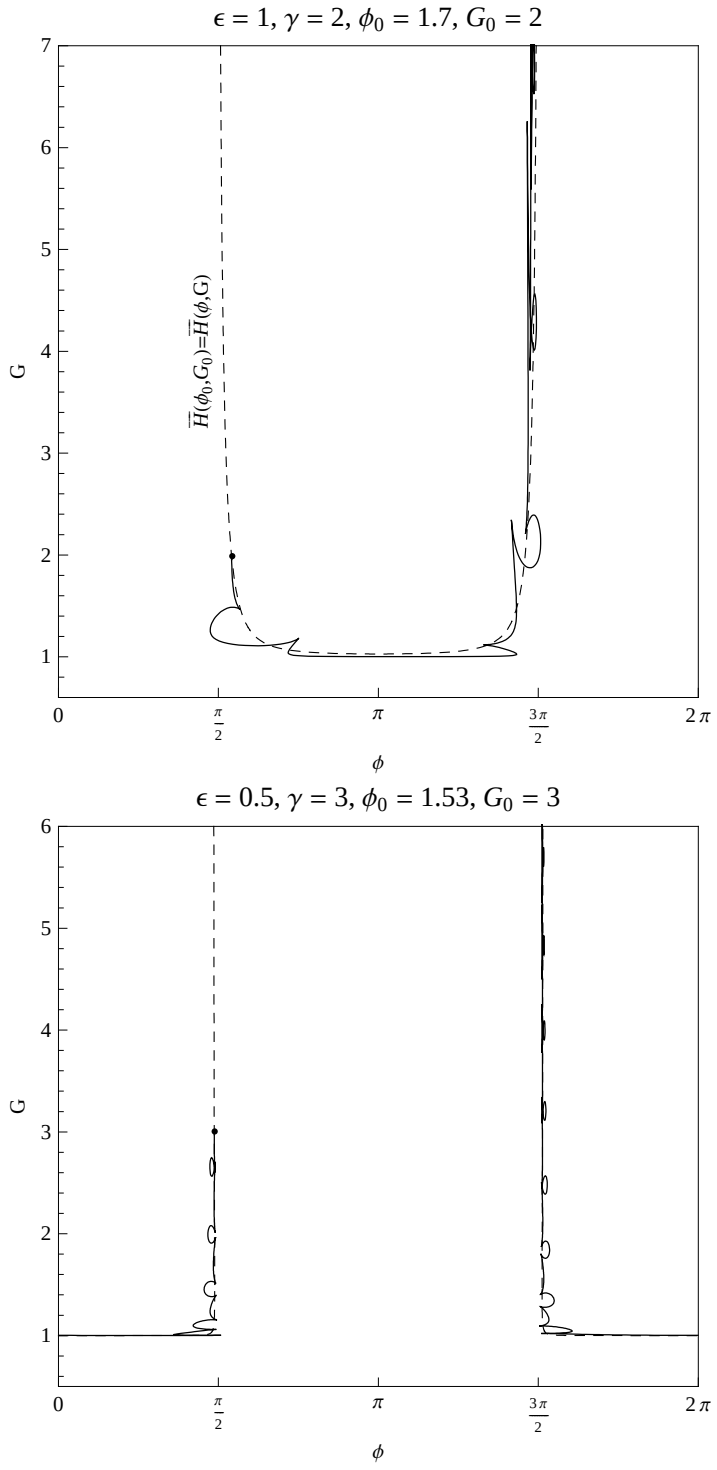


Figure 2: The numerical solution (solid line) of (5) and (6) and the invariant curve (dashed) of the averaged Hamiltonian (7). Small dots denote the initial point.

4 Evaluation of Auxiliary Integrals

This section contains the proof of the following proposition. For $n \in N_0$ and $|\beta| < 1$ it is true that

$$\int_0^{2\pi} \frac{\cos nt}{1+\beta\sin t}\,\mathrm{d}t = 2\pi\frac{\beta^n}{\sqrt{1-\beta^2}(1+\sqrt{1-\beta^2})^n}\cos\frac{\pi n}{2}, \tag{11}$$

$$\int_0^{2\pi} \frac{\sin nt}{1+\beta\sin t}\,\mathrm{d}t = -2\pi\frac{\beta^n}{\sqrt{1-\beta^2}(1+\sqrt{1-\beta^2})^n}\sin\frac{\pi n}{2}, \tag{12}$$

$$\int_0^{2\pi} \frac{\cos nt\cos t}{1+\beta\sin t}\,\mathrm{d}t = 2\pi\frac{\beta^{n-1}}{(1+\sqrt{1-\beta^2})^n}\sin\frac{\pi n}{2}, \tag{13}$$

and for $n \in \mathbb{N}$

$$\int_0^{2\pi} \frac{\sin nt\cos t}{1+\beta\sin t}\,\mathrm{d}t = 2\pi\frac{\beta^{n-1}}{(1+\sqrt{1-\beta^2})^n}\cos\frac{\pi n}{2}. \tag{14}$$

For $n = 0$ this is obviously zero.

We will check the equality (11). The proof of the others is analogous. For the sake of brevity denote the LHS of (11) by symbol I. Using the multiple angle formula for cosine, the geometric series expansion of the denominator, and the relation

$$\int_0^{2\pi} \cos^k t\sin^n t\,\mathrm{d}t = \frac{(1+(-1)^k)(1+(-1)^n)}{2}B\Big(\frac{1+k}{2},\frac{1+n}{k}\Big),\ n,k\in\mathbb{N}_0,$$

one obtains

$$I = \sum_{k=0}^{n}\cos\frac{\pi}{2}(n-k)\sum_{m=0}^{\infty}(-\beta)^m\frac{(1+(-1)^k)(1+(-1)^{m+n})}{2(-1)^k}B\Big(\frac{1+k}{2},\frac{1}{2}(1+m+n-k)\Big).$$

The summands with k or $m+n$ odd are zero. Hence we can assume that k and $m+n$ are even. Furthermore if n is also odd then $\cos(\pi(n-k)/2) = 0$ and therefore $I = 0$. We must investigate the case of $n = 2N$ where $N \in \mathbb{N}_0$. After re-notation of indexes we clearly have

$$I = 2\sum_{k=0}^{N}\binom{2N}{2k}(-1)^{N-k}\sum_{m=0}^{\infty}\beta^{2m}B\Big(\frac{1}{2}k,\frac{1}{2}+N+m-k\Big).$$

Rewriting the Beta function in terms of Gamma function and using the Gauss hypergeometric series

$$F(a,b,c,z) = \frac{\Gamma(c)}{\Gamma(a)\Gamma(b)}\sum_{m=0}^{\infty}\frac{\Gamma(a+m)\Gamma(b+m)}{\Gamma(c+m)}\frac{z^m}{m!},$$

we arrive at

$$I = (-1)^N\frac{2\pi(2N)!}{2^N}\sum_{k=0}^{N}\frac{(-1)^k}{(2N-2k)!!(2k)!!}F^{reg}(1,N-k+1/2,N+1,\beta^2),$$

where $F^{reg}(a,b,c,z) = F(a,b,c,z)/\Gamma(c)$ is regularised hypergeometric function. Next step is to take adventage of the symmetry in interchange of a, b and of the integral representation of hypergeometric function

$$F(a,b,c,z) = \frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)} \int_0^1 t^{b-1}(1-t)^{c-b-1}(1-tz)^{-a}\ \mathrm{d}t.$$

It turns out that

$$I = \pi \frac{(-1)^{N+1}}{2^{2N}} \beta^{2N+1} F(N+1, N+3/2, 2N+2, \beta^2),$$

where the binomial theorem was used. Final step is to look in [1] and find relation 15.1.14:

$$F(a, a+1/2, 2a, z) = 2^{2a-1}(1-z)^{-1/2}(1+\sqrt{1-z})^{1-2a}.$$

Hence

$$I = 2\pi(-1)^N \frac{\beta^{2N}}{\sqrt{1-\beta^2}(1+\sqrt{1-\beta^2})^{2N}}.$$

Combining results for odd and even n one obtains the formula which was to be proved. For the sake of completeness note, that for the computation of the two last integrals one needs formula [1], 15.1.13

$$F(a, a+1/2, 2a+1, z) = 2^{2a}(1+\sqrt{1-z})^{-2a}.$$

5 Conclusion

We studied the dynamics of the classical charged particle placed in the homogeneous magnetic field and influenced by time-periodic Aharonov-Bohm flux. With the aid of the averaging method it is possible to compute a good approximation to the solution of action-angel equations of motion, provided the ratio of the magnetic field strength and the flux frequency is a natural number. In this resonant situation the radial motion of the particle is highly oscillatory, more precisely the particle can be located arbitrary close to and far from origin.

References

[1] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables.* Dover, (1964).

[2] V. Arnold. *Mathematical Methods of Classical Mechanics.* Springer-Verlag, (1989).

[3] J. Guckenheimer and P. Holmes. *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields.* Springer-Verlag, (1983).

[4] J. A. Sanders and F. Verhulst. *Averaging Methods in Nonlinear Dynamical Systems.* Springer-Verlag, (1985).

Invariant Picture Region Detection

Jan Kamenický

4th year of PGS, email: j.kamenicky@sh.cvut.cz
Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU
advisor: Jan Flusser, Institute of Information Theory and Automation, AS CR

Abstract. In the article we present a newly developed method for invariant picture region detection without a priori information. The goal is to detect such picture regions that remain more or less unchanged after various simple transformations have been applied (such as change of brightness, contrast, scale, cropping of the picture, etc.). Such regions can be then used for automatic picture identification.

The method itself is based on block processing of the picture. The importance of a block is given by representation of individual intensities in the block. Individual blocks are then put together and rated according to other criterions. The result is based on one (or more) "stable" region(s) of the picture.

At the end, experimental results on real data are presented, verifying the functionality and practical usage of the method.

Abstrakt. V článku je prezentována nově vyvinutá metoda pro vyhledávání invariantních oblastí obrázků bez apriorních informací. Cílem je nalézt takové oblasti obrázku, které zůstávají víceméně nezměněné po aplikaci různých jednoduchých transformací (změna jasu, kontrastu, měřítka, oříznutí obrázku, apod.). Takové oblasti je dále možné využít pro automatickou identifikaci obrázku.

Metoda samotná je založena na blokovém zpracování obrázku. Významnost bloku je dána zastoupením jednotlivých intenzit v tomto bloku. Jednotlivé bloky jsou poté sdružovány a ohodnoceny podle dalších kritérií. Výsledkem je pak jedna (nebo více) "stabilních" oblastí obrázku.

Na závěr jsou uvedeny výsledky na reálných datech, které potvrzují funkčnost a praktickou použitelnost navržené metody.

1 Introduction

1.1 Background

Large image databases are often run on a commercial basis – browsing through and viewing images is free of charge while downloading and re-using them on web pages or in articles is a subject of a fee. However, some users republish downloaded images without paying the fee, which is a violation of copyright law. The copyright owner thus wants to regularly scan suspicious domains or websites to check if there are some unauthorized copies of their database images.

Detection of such illegal copies is complicated by two principal difficulties – the unauthorized images are usually modified before they are republished on the web, and the response of the system must be extremely fast because of the enormous number of database

images. Although this problem formulation looks like an image retrieval task, this is not the case. In traditional image retrieval, we want to find in the database all *similar* images to the query image, where similarity is evaluated by colors, textures, content, etc. Here we want to identify only the *equivalent* images to the query (we call this task *image identification*). This is why we cannot apply most of standard image retrieval techniques. By the term "equivalent images" we understand any pair of images which differ from one another by the following transformations.

- quality reduction - recompression or resize
- radiometric and color distortions - changes of brightness, contrast, color tone, conversion to gray-scale
- cropping of the image - major part still preserved
- local changes - addition of logos or thin labels
- combinations - reasonable combinations of the modifications mentioned above (however, their increasing complexity will surely impact the algorithm results)

1.2 Motivation

We have developed a new method for the above mentioned image identification. This method consists of several steps. The core of the identification process is an intensity based stochastic tree, which needs whole (rectangular) pictures on input. If we want to overcome problems with some mentioned transformations (especially scaling and cropping) and still provide rectangular image, we need to extract some "important" part of the picture which can then be used as the required input. So, we do an *invariant picture region detection* and use the best one. In this article we will describe this process in more detail.

As we have already said, the problem is to crop the input picture in such way, that it gives as similar result to the crop of the corresponding database picture as possible (provided that the input picture was created by modifying the original). In other words, image A and transformed image A' should give the same (or similar) invariant regions. Otherwise we wouldn't be able to identify the modified image properly.

The only restriction we presume is that the main (most important) part of the picture remains (it is not destroyed be any mentioned transformation). In practice it is a reasonable assumption because what is usually cropped away is background or some not very important information.

Summarizing the main requirements for our desired method:

- **invariance** to simple basic transformations such as change of brightness and/or contrast, repacking (i.e. weak noise), cropping, scaling, etc., and their combinations
- **stability** - picture modified as described above should give same (or at least similar) results compared to the original

- **speed** - the maximum time available for one picture (about 0.3 Mpixel large) is only about five seconds

It is impossible to develop a method exactly fulfilling all the above mentioned requirements. So at first we have to choose the importance of individual requirements and then create a method which is as good as possible with respect to them. Not only we deal with a huge amount of images but also we need to process several images per minute. Therefore, the most important for us is the time available, thus the method should be fast in the first place.

There are several possible basic approaches to this problem. One of them is critical/significant point detection. In this case we are trying to detect stable important points, such as line endings or corners. These methods typically result in many points of interest but what we need is as few stable points as possible, preferably only the best one. So, this approach cannot be used in our case because it is very difficult, if not impossible, to decide what critical point is the best and would be detected in the modified image in the same way.

Another approach is to detect local homogeneous regions (i.e. regions with homogeneous intensity values). Many methods solving this problem exist, an overview of segmentation methods can be found in [1]. Probably the most useful approach for us is region growing. This approach is much better because it is possible to decide which region is better. We can use size of the region, intensity variance, compactness, and so on for this decision. Some region growing algorithms can be found in [2], [3], [4], [5] or [6].

But we have another problem – speed. Region growing methods generally take their time and we need fast response. That's why we have developed a new method based on homogeneous region detection which is more simple but still gives reasonable results.

2 Method description

The presented method is a method for detection of stable regions of an image, i.e. regions detected invariantly to some basic transformations (mentioned above).

Current version of our method is intended for grayscale images only, though further improvements making use of the color information can be done and can potentially lead to better results because more information is available.

Method itself consists of two main steps: stable point detection, and region extraction.

2.1 Stable point detection

Our stable point detection is based on local homogeneous region detection through intensity variance local minimization. The resulting points are then selected as centers of gravity of these regions.

Finding region with minimal intensity variance is very time consuming. Therefore we apply block processing for this task. At first we divide the image into overlapping blocks. The size of these blocks can be chosen either as static (e.g. 20 pixels) or dynamically as further described in experimental section.

On these blocks we compute intensity variance. Now we have much smaller matrix with variance values which is much faster to work with. We call it stability map and we apply thresholding on it resulting in a small binary image.

Now we label different continuous areas of the stability map. As we need stable areas which can be detected even on cropped images, we eliminate areas touching the border of the image. Such areas are either background or too unstable to be used. Especially background is typically present near the image border and the detection of such area is likely to be very unstable.

Now we have a set of stable areas and the only thing left is to select the best of them. For this decision we use:

- the size of the area – the larger area the better (more stable),
- its homogeneity – the less variance the better, we already used this criterion in stability map creation and its thresholding,
- compactness – the more compact the area is the better (again more stable),
- distance from the border – as we already mentioned, areas nearing the border are likely to be misdetected.

So, we have the best stable regions. One possibility is to use it directly as final stable region. The problem is we used block processing for stability map creation. Therefore detected areas are accurate only on the level of these blocks. In case of typical blocks of size about 20 pixels, we can see that usage of these stable areas is very inaccurate on the image level. That's why we compute a stable point as the region representation and then use region extraction back in image domain.

The stable point itself is chosen as the center of gravity of the best stable area.

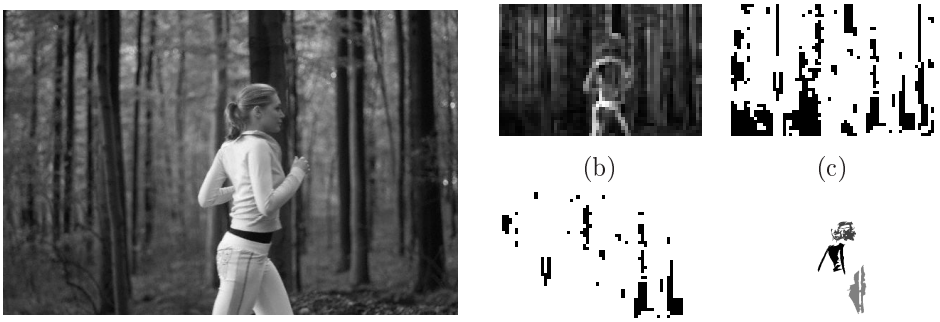
2.2 Region extraction

What we need to do, is to select a region based on the detected stable point with specific requirements. Certainly we want it to correspond to the detected homogeneous area. Using a predefined threshold we select intensity interval from the input image. We take the mean intensity value computed during stable point detection as the interval centre, the range is given by the threshold. Now we extract stable region as a compact (continuous) area with the stable point in its interior.

As described in the motivation section 1.2, we need rectangular picture as a result. Therefore we select final region as frame surrounding the detected stable area.

This frame is defined by the center of gravity and second order central moments of the stable area. This means we take the center as the center of our frame and select width and height having the same second order moments as the original region.

So, the final picture used for the identification through our stochastic tree is the picture cropped by the above computed frame.



(a) (b) (c) (d) (e)

Figure 1: (a) original picture (b) stability map, (c) thresholded stability, (d) border areas removed, (e) best regions segmented from the original image

3 Experimental results

In this section we will describe behavior of the method and then we will show experimental results on real data.

3.1 Method functionality

At first we will demonstrate functionality of the method on a sample picture. In figure 1 you can see:

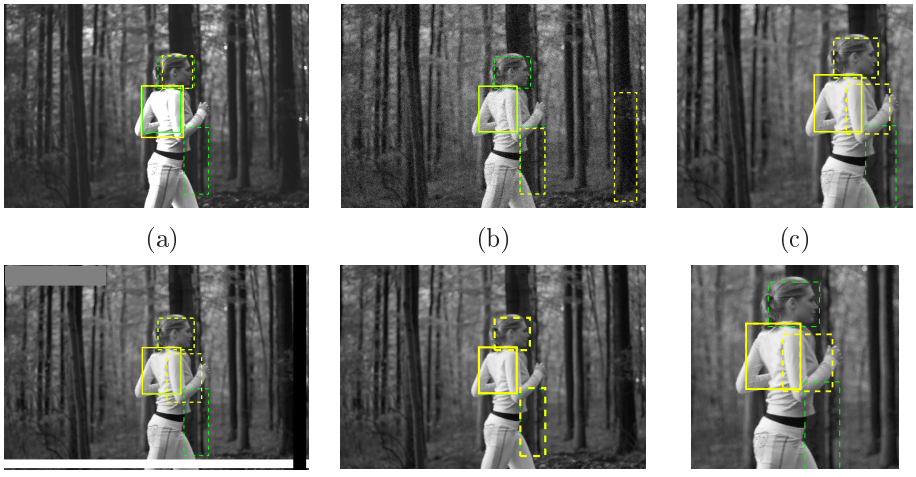
- (a) The original picture.
- (b) The stability map, i.e. the block computed variance. The block size and shift of an image with size $m \times n$ are computed as $B_{size} = 2B_{shift} = 2\sqrt{(m+n)/10}$, in this case it is 20 and 10 pixels respectively.
- (c) Thresholded stability map.
- (d) Thresholded stability map with marginal areas removed.
- (e) Three final best stable regions. These regions are segmented back from the original image based on previously detected stable points.

You can see the final frames in figure 2. The best detected frame is drawn by a solid line, the two next frames are represented by a dashed line.

3.2 Results after modifications

Now we can demonstrate how the method handles required modifications. We will do this on the same image for more simple comparison of the achieved results. You can see best detected regions in the image after applying several modifications in figure 3. The thicker yellow rectangles represent detected regions, while the thinner green rectangles correspond to expected locations of regions detected in the original image.

Figure 2: Best stable regions



(a) (b) (c)

(d) (e) (f)

Figure 3: Best regions in modified images. (a) brightness and contrast, (b) white noise, (d) added artifacts, (e) scaling, (c, f) cropping

Table 1: Statistical results – possibility of successful identification

modification	successfulness
original	100%
JPEG compression	86%
scaling	62%
brightness and contrast change	58%
cropping	47%
added artifacts	42%

The first modification is slight brightness and contrast change and the second is a white noise addition. In both cases, the impact of the intensity changes on the results is quite noticeable. The best region is still detected correctly (with a small difference). However, when the regions are further used for picture identification, the error rate can be rather high.

Next modification we can see in the figure is some artifact addition, namely some rectangles near the border. We can see that the only problem in this case arises when originally detected regions are overlapping (or nearing) the artifacts. These regions are then not detected correctly, however, we expect these artifact to appear only near the border, so important parts of the image should not be affected.

Another important transformation is scaling. In the experiment we used scaling down by the factor of 1.7. As we can see, the impact is very insignificant, the result is nearly the same compared to the original image. Naturally, with increasing scaling factors the results become worse, but we expect reasonable factors to appear most often.

Last modification we have tested here is cropping, which is also quite important for us. Again, we can see that the main problem arises when the originally detected regions approach or even exceed image borders. So, positive identifications can be made only if the important part remains inside the image. However, this is just what we expect to happen.

3.3 Statistical results

We have tested the method on one thousand of pictures, each modified by methods shown in the previous section. We used random coefficients (such as scale factor, brightness change, cropping, etc.) taken from meaningfully restricted intervals. Individual results are similar to the described ones, sometimes slightly worse.

The statistical overview is shown in table 1. These are only very rough numbers indicating the possibility of further successful identification for individual modification types. It should be mentioned that 50% is a very good result for us, as we use the method for automatic identification. In fact, there is much more important for us to keep the number of false alarms as low as possible, than achieving very high successful hit ratio.

As we can see, practical results are acceptable while fulfilling required attributes. The time needed to compute best stable regions for one image (modification) is about one second.

4 Conclusion

In the article we have presented a newly developed method for invariant picture region detection. It is intended for cases where speed is critical. The performance of the method has been demonstrated on real experiments. Quality of the results together with rather fast computing time is quite promising.

Many further modifications of the method can be made according to specific usage conditions.

References

[1] N. R. Pal, S. K. Pal. *A review on image segmentation techniques.* Pattern Recognition **9** (1993), 1277-–1294.

[2] A. Tremeau, N. Bolel. *A region growing and merging algorithm to color segmentation.* Pattern Recognition **7** (1997), 1191—1203.

[3] A. Mehnert, P. Jackway. *An improved seeded region growing algorithm.* Pattern Recognition Letters **10** (1997), 1065—1071.

[4] Ch. Revol, M. Jourlin. *A new minimum variance region growing algorithm for image segmentation.* Pattern Recognition Letters **3** (1997), 249–258.

[5] J. Fan, D. K. Y. Yau, A. K. Elmagarmid, W. G. Aref. *Automatic image segmentation by integrating color–edge extraction and seeded region growing.* IEEE Transactions on Image Processing **10** (2001), 1454-–1466.

[6] F. Y. Shih, S. Cheng. *Automatic seeded region growing for color image segmentation.* Image and Vision Computing **10** (2005), 877–886.

Thermodynamic Model of Bone Adaptation*

Václav Klika

1st year of PGS, email: klika@it.cas.cz
Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU
advisor: František Maršík, Institute of Thermomechanics, AS CR

Abstract. The capacity of bone to adapt to functional mechanical requirements has been known for more than a century, and many theoretical and experimental models have been developed for bone remodelling. However, these models are still not able to sufficiently predict its behaviour. A thermodynamic model based on recent knowledge of biochemical control mechanisms is presented. Despite the complexity of the regulatory system of bone adaptation, the calculated results are in very good correlation with the experimental observations - the inner structure of bone can be elucidated, simulation of the influence of dynamic loading together with biochemical factors is carried out, e.g. the fundamental RANKL-RANK-OPG pathway, and a comparism between model prediction and x-ray pictures of human patients of the effect of bone adaptation to prosthesis insertion is done.

Abstrakt. Už více než sto let je známá schopnost kosti přizpůsobovat se mechanickým požadavkům a též mnoho teoretických a experimentálních modelů pro remodelaci kostí bylo vytvořeno. Stále však nejsou tyto modely dostatečně schopné předpovídat její chování. Zde předkládáme termodynamický model založený na současných znalostech biochemického řízení procesu. Navzdory veliké složitosti řídícího systému adaptace kostí jsou vypočítané výsledky ve velmi dobré shodě s experimentálními pozorováními - můžeme vysvětlit vnitřní strukturu kosti, byly provedeny simulace vlivu dynamické zátěže spolu s biochemickými faktory - např. základního řetězce RANKL-RANK-OPG - a byla porovnána předpověď modelu adaptace kosti na vložení totální endoprotézy s rentgenovými snímky pacientů.

1 Introduction

Bone is biological system which keeps adapting its structure to mechanical environment. In the 19th century Julius Wollf [30] described the fact that the internal trabecular architecture of bone matches trajectories of the mechanical stress (trajectorial hypothesis). At the same time Wilhelm Roux suggested a quantitative self-regulating mechanism of trabecular formation and functional adaptation. Mechanical stimuli to local cells was considered critical for the bone adaptation process [24] and this interaction was later described by Heřt in 1970s [7]. In 1987, Frost [5] suggested a feedback mechanism, the "mechanostat", controlling bone mass behaviour in response to mechanical loading.

When mechanical stresses are placed upon bone, it remodels in order to withstand the stresses. This process may also be considered to be structural optimisation. The optimisation process systematically, iteratively and continually eliminates and redistributes

*This research has been supported by the Grant agency of the Czech Republic no. 106/03/1073 and by the project 1M06031 of Ministry of Education, Youth and Sports of Czech Republic.

osseous material throughout the domain to obtain an optimal arrangement of internal bony structures.

With the development of computer-aided strategies and based on the knowledge of bone geometry, applied forces and elastic properties of the tissue, it may be possible to calculate mechanical stress transfer inside the bone (FE-analysis). Assuming the above mentioned structural optimisation process the change of stress in particular compartments of the bone should further be followed by internal bone density distribution. This logical consequence allows us to think about mathematical models that can be used to study functional adaptation quantitatively and furthermore to create the mineral bone density distribution patterns [8, 29]. Similar mathematical models have been built in the past. Since they calculate just mechanical transmission inside the bone and not considering humoral cell-biologic factors of bone physiology, they just partially correspond to reality seen in living organisms. We realize that biochemical reactions are initiated and influenced primary by genetic effects and the external biomechanical effects (stress changes). The aim of following mathematical model is to combine the biological factors with biomechanical ones[11, 15, 16]. Such model may also reflect changes in remodelling behaviour corresponding to pathological changes of the bone metabolism [13, 12].

Biology of bone remodelling

Bone remodelling (BR) occurs when the populations of bone cells break down old bone and replace it with new bone. This reformation results in the reorientation of internal bone structure and eventually in changing the shape of the bone, which means that bone can better adapt to the loads that are being placed upon it. Loads on bone cause mechanical strains and even micro-damage generating signals that specific cells can detect and to which they or other cellular populations respond. Actually remodelling depends on time-varying straining. Because of the viscoelastic properties of bones, the strains vary not only at varying loading but the strain changes continue and fade as the elastic after-effects at constant loads and after unloading. In this manner, the existence of remodelling effects even at rest can be explained [26].

The signalling and subsequent change in cellular phenotype may be called activation and represents the first stage of the remodelling process. The aim of activation is to prepare sufficient pool of executive cells concentrated in the domain of the bone that is to be repaired. The original bony structures (Old_B) infracted by the initiating biomechanical stimuli are intended to be absorbed and subsequently replaced by the new bone (New_B). The generated bone mass will be structurally and morphologically adjusted to new mechanical loads. These two phases described as resorption and formation, accomplish the whole process of remodelling.

Biology of the bone remodelling itself is not completely understood in this moment. Frost has defined the minimum effective strain neither apposition nor resorption below 1500-2500 microstrains. According to Frost [3], the strains above that threshold level affect modelling and remodelling activities in ways that change the size and configuration of growing bones, tendons, ligaments and fascia to their new mechanical usage and return their strains to the threshold level. Recently, the control mechanism between resorption and formation of bone was described, by so called RANKL-RANK-OPG pathway [17, 18, 19] and our mathematical model covers the crucial moments (based on chemical

description of bone remodelling process [23]).

System of basic multicellular units (BMU) is widely accepted for bone remodelling description [4, 21]. It represents local populations of osteocytes (OC), osteoblasts (OB) and of mononuclear precursor osteoclasts ($MCELL$). Osteocytes are presumed to react to mechanical strain either piezoelectrically through ionic currents induced when bone is deformed or by detecting fluid flow in the periosteocytic lacunas. They respond to this strain by sending signals that activate bone formation or existing bone removal. During the activation $MCELL$ turn to multinucleated osteoclasts ($MNOC$) having high metabolic activity. $MNOC$s are charged with resorption of the old bone and the defect is subsequently filled with osteoid, non mineralised bone matrix produced by activated osteoblasts (OB) that during next 5-15 days becomes mineralised.

2 Methods

2.1 Thermodynamic BR model

The only cells that are able to resorb bone tissue are osteoclasts (as mentioned in section 1). To be active they need to be coupled in multinucleated complex, which formation can be described as follows:

$$D_1 + MCELL \overset{k_{\pm 1}}{\rightleftarrows} MNOC + D_4, \tag{1}$$

where D_1 is mixture of substances that are initiating reaction with mononuclear cells ($MCELL$). $MNOC$ is abbreviation for multinucleated osteoclast and D_4 is a remaining product from reaction (1).

Bone decomposition can be characterised by following chemical reaction:

$$MNOC + Old\_B \overset{k_{\pm 2}}{\rightleftarrows} D_6 + D_7, \tag{2}$$

where Old_B denotes old bone, D_6 and D_7 are products made during degradation of an old bone. The end product in reaction (2) is divided into two parts because one of them (D_7) participates in activation of osteoblast as will be elucidated in subsequent paragraph. The chain RANKL-RANK-OPG which is important as the control mechanism for bone remodelling is substituted by the concentration level of the mixture of substances D_1.

Before osteoblasts (OB) secret collagen in hollowed cavity they need first to be activated. This activator ($Activ\_OB$) is being produced after resorption in given volume (cavity). Thus behaviour of osteoblasts at specific site can be represented by following reaction scheme:

$$D_7 + Old\_B \overset{k_{\pm 3}}{\rightleftarrows} Activ\_OB + D_9 \tag{3}$$

$$Activ\_OB + OB \overset{k_{\pm 4}}{\rightleftarrows} Osteoid + D_{12}, \tag{4}$$

where D_{12} is remaining substratum.

The longest period in bone remodelling process pertains to mineralisation (deposing calcium, etc. – D_{13} – into matrix) of osteoid:

$$D_{13} + Osteoid \overset{k_{\pm 5}}{\rightleftarrows} New\_B + D_{15}, \tag{5}$$

where New_B denotes new bone formed by remodelling process and D_{15} is the residuum of bone formation reaction.

These chemical equations (1)-(5) describe the essential processes of bone remodelling. There are 15 substances involved and by using the law of active mass and adding the external fluxes, 5 differential equations describing the whole system can be obtained:

$$\frac{\partial N_{MCELL}}{\partial \tau} = -\delta_1(\beta_1 + N_{MCELL})N_{MCELL} + \mathcal{J}_3 + \mathcal{J}_{New\_B} - \mathcal{D}_1 \tag{6}$$

$$\frac{\partial N_{Old\_B}}{\partial \tau} = -(\beta_3 - N_{MCELL} + N_{Old\_B} + N_{Activ\_OB} + N_{Osteoid} + N_{New\_B})N_{Old\_B} - \\ - \delta_3(\beta_7 - N_{Old\_B} - 2(N_{Activ\_OB} + N_{Osteoid} + N_{14}))N_{Old\_B} + 2\mathcal{J}_{New\_B} - \mathcal{D}_2 - \mathcal{D}_3 \tag{7}$$

$$\frac{\partial N_{Activ\_OB}}{\partial \tau} = \delta_3(\beta_7 - N_{Old\_B} - 2(N_{Activ\_OB} + N_{Osteoid} + N_{New\_B}))N_{Old\_B} - \\ - \delta_4(\beta_{10} - N_{Osteoid} - N_{New\_B})N_{Activ\_OB} + \mathcal{D}_3 - \mathcal{D}_4 \tag{8}$$

$$\frac{\partial N_{Osteoid}}{\partial \tau} = \delta_4(\beta_{10} - N_{Osteoid} - N_{New\_B})N_{Activ\_OB} - \\ - \delta_5(\beta_{13} - N_{New\_B})N_{Osteoid} + \mathcal{D}_4 - \mathcal{D}_5 \tag{9}$$

$$\frac{\partial N_{New\_B}}{\partial \tau} = \delta_5(\beta_{13} - N_{New\_B})N_{Osteoid} - \mathcal{J}_{New\_B} + \mathcal{D}_5, \tag{10}$$

where $\tau = t \cdot k_{+2} \cdot n_{B_0}$, $N_i = \frac{n_i}{n_{B_0}}$, $\delta_\rho = \frac{k_{+\rho}}{k_{+2}}$, $\beta_i = \frac{\mathcal{B}_i}{n_{B_0}}$, $\mathcal{D}_\rho = \frac{l_{\rho v} d_{(1)}}{k_{+2} n_{B_0}^2}$, $\mathcal{J}_i = \frac{J_i}{k_{+2} n_{B_0}^2}$. In other words δ_ρ is ratio of rate of ρ-th reaction to second reaction, $\mathcal{D}_\rho$ is a parameter that describes the influence of dynamic loading on rate of ρ-th chemical reaction, β_i is a sum of initial molar concentration of relevant substances and N_i is a normalised concentration of i-th substance.

By solving these kinetic equations, time evolution of $MCELL$, Old_B, $Activ\_OB$, $Osteoid$, New_B concentrations are obtained. All remaining can be calculated (for more details and for detailed mathematical analysis see [10]).

3 Results

It can be shown thatAs was mentioned in section the key role in presented model plays the coupling of the dynamic loading and chemical reaction rates[1]. Thus for the validation of the model a following simulation in ANSYS FE software (ANSYS 10.0, Ansys inc.) was used: rate of deformation in each element and consequently the rates of chemical reactions in each element ($\mathcal{D}_\rho$) were calculated. Then with the aid of kinetic equations (6)-(10) we can describe time dependency of concentrations of each substance. Thus density changes (precisely changes in concentration of Old_B, New_B) can be calculated in each element of bone.

Hence after running several iterations we may simulate the density distribution throughout the bone according to the presented model. By comparing these calculated results with density distribution in living bone validation can be carried out.

[1]Moreover, reaction may run even in a case when some chemical reactions have negative affinity $\mathcal{A}_\rho$. Then the influence of such reaction $w_\rho \mathcal{A}_\rho < 0$ is compensated by the enhanced efficiency of the other reactions

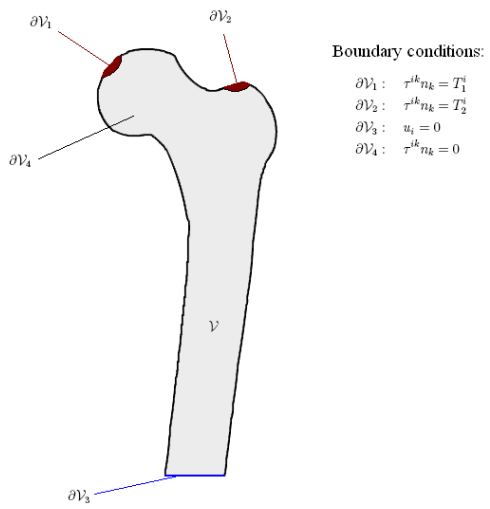


Figure 1: Geometry and boundary conditions used for calculations. Elastic constants depend on bone components, especially N_{Old_b} and N_{New_b}.

Ansys program was used to calculate strains (sum of principal strains) and stresses in each element of bone during walking. Real geometry was gained from CT-scan and external forces were applied as is shown in Fig.1.

Values of $\mathcal{D}_\rho$ parameters are derived from deformation rate tensor. Since ANSYS calculates just deformations (strains), to determine the spheric part of deformation rate tensor $d_{(1)}$ following approximation was used:

$$d_{(1)}(I) = \frac{\mathrm{d}\, e_{(1)}(I)}{\mathrm{d}\, t} \approx \frac{\Delta e_{(1)}(I)}{\Delta t} = \frac{e_{(1)}(I)}{\Delta t}, \quad (11)$$

where $e_{(1)}(I)$ is the trace of deformation tensor in I-th element. Provided that no deformation exists at the beginning of each iteration, last equality in (11) is valid. Thus Δt is the time interval between loaded and unloaded state. The change of Δt enables to include the influence of frequency of the loading (e.g. the setting of the paces) on bone remodelling which will later be discussed.

The density in each element can be calculated as follows:

$$Dens(I) = \hat{\rho} \cdot \left(N_{New\_b}(I) + N_{Old\_b}(I)\right), \quad (12)$$

where $\hat{\rho}$ is a reference (apparent) density and N_{New_b}, N_{Old_b} are normalised concentrations of old and new bone in I-th element, respectively.

We are not interested only in density distribution but also in withstanding to applied load after effects of bone remodelling. Thus in each element of bone material properties are modified according to changes in density as power to three:

$$\begin{aligned} E_{zz}(I) = \Big(&E_{zz\ old} \cdot frac\_N_{Old\_b} + \\ &+ E_{zz\ new} \cdot frac\_N_{New\_b}\Big) \cdot \left(\frac{Dens(I)}{\hat{\rho}}\right)^3, \end{aligned} \quad (13)$$

where

$$frac\_N_{Old\_b} = \frac{N_{Old\_b}(I)}{N_{New\_b}(I) + N_{Old\_b}(I)}$$

and

$$frac\_N_{New\_b} = \frac{N_{New\_b}(I)}{N_{New\_b}(I) + N_{Old\_b}(I)}$$

are ratios of N_{Old_b} and N_{New_b} in I-th element, respectively. $E_{zz}(I)$ is the Young modulus in direction of axis 'z' in I-th element, $E_{zz\ old}$ and $E_{zz\ new}$ are the material properties of old and new bone. We expect stiffness of bone to vary not only with density but also with the old/new bone ratio. Similarly were calculated E_{yy}, E_{xx}, G_{xy}, G_{yz} andG_{xz}.

At first, the impact of mechanic loading on density will be elucidated on a single element.

Presented model calculates (molar) concentrations in considered volume of bone. By means of eq. (1)-(5) are described chemical reactions that are assumed to run in each part of bone independently.

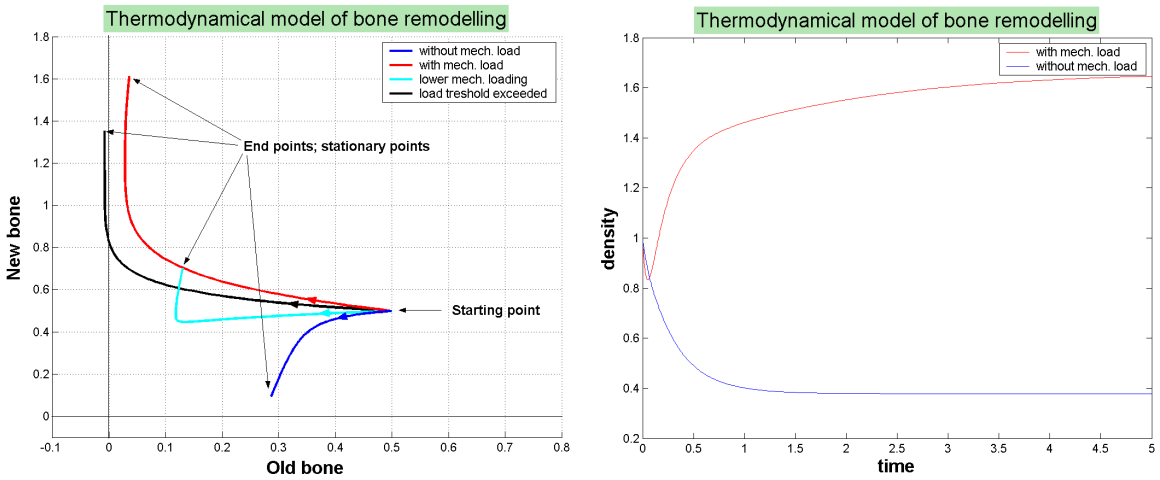


Figure 2: Evolution of structure in several cases of mechanical loading - Fig. (a) - and density of bone in 1 (isolated) element - Fig. (b).

To illustrate the effect of dynamic loading we varied the $\mathcal{D}_\rho$ parameters of the model. If no stimuli is present, the bone resorbs (concentration of both old bone and new bone is decreasing) and after some period of time reaches equilibrium where almost no new bone is being produced (Fig.2a, blue line). On the other hand if proper exercise is applied (yet not possible to determine what type of exercise it represents) firstly a moderate decrease in density may be observed. This decrease is soon shifted into significant rise where formation predominates resorption. When the activity is increased furthermore (by higher frequency or higher load) the stationary solution may become unreal (negative value of Old_B) - Fig.2a, black line. According to the model, there is a threshold for dynamic loading. Exceeding this threshold leads to bone fracture.

3.1 Simulation results

The only parameters of mathematical model of BR that vary throughout the bone are $D_\rho, \rho \in \hat{5}$ (all the other parameters are assumed to be constant throughout the whole bone and independent on time). In other words, all the calculated results here presented are consequence of solely dynamic loading as a control factor.

On the other hand if mechanical stimuli in given element is small, the bone mass is not zero. In this case the biochemical factors prevails, such as hormones and nutrition.

As a initial state a homogeneous distribution of density throughout the whole bone was used ($Dens(I) = \hat{\rho}, \forall I$). Since each iteration is calculated by solving differential equation (6)-(10) describing "thermodynamic BR model", each iteration is not just approximation

of correct solution but is actually a time evolution of bone remodelling in bone. We are trying to simulate nature - to create the bone organ from homogeneous tissue.

Each step shows density changes (impact of remodelling) in each part of bone. Steady state is reached after a few tens of iterations, one iteration step corresponds approximately to 10 thousands strides. Approach to a steady state is checked by mean and maximal density change. After 35 iterations the values are following: max difference $\approx 1e-03$ and mean diff. $\approx 1e-04$.

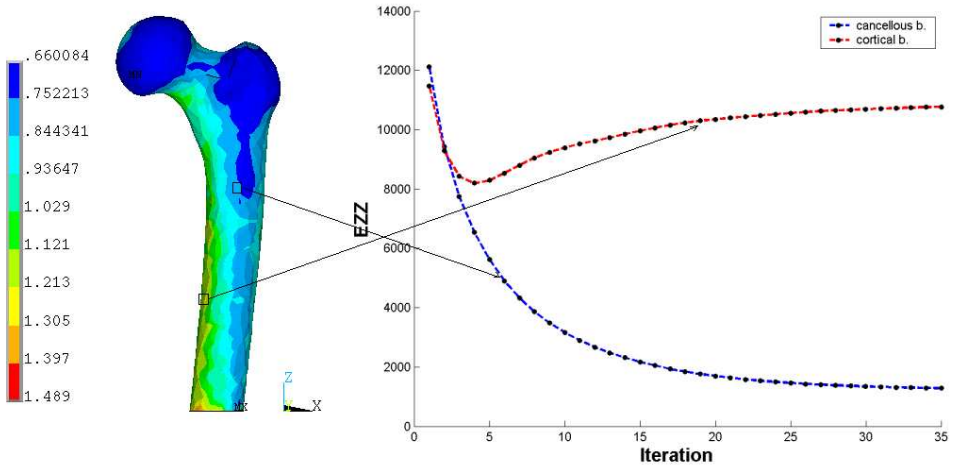


Figure 3: 'Healthy state' - result from ANSYS computation using thermodynamic BR model calculated from initial homogeneous density distribution. Notice the clear formation of cortical and cancellous bone.

The Fig.3 depicts a calculated result from initial homogeneous distribution of density after 35 iterations of BR according to presented model. It refers to healthy person - similar as in Fig.2a, red line.

Notice the eminent correspondence in Fig.4 where you may compare calculated results with the inner structure of proximal femur as it is known from human anatomy. BR process (according to our model) creates cortical and cancellous bone (even from homogeneous distribution) even when *only* influence of external forces is considered. From here is patent how mechanic (dynamic) loading not only significantly influences the bone remodelling process - resorption or formation of bone in a given element - but also determines the shape, thickness and emplacement of cortical bone. See enclosed graphs of time evolution of Young modulus in different parts of bone, where can be clearly seen the formation of spongy and cortical bone.

The main goal for every theoretical model in biological sciences is of course the application and possibly prediction of evolution of the particular process. One of the very important applications of bone remodelling model is the prediction of adaptation of bone to different mechanical conditions. Such a change occurs e.g. when degenerative arthrosis disables a proper function of joint and a surgical total replacement of joint is needed. This replacement is done by prosthesis made from various materials (mainly steel, composites,...) which are very different from bone tissue from mechanic point of view. And

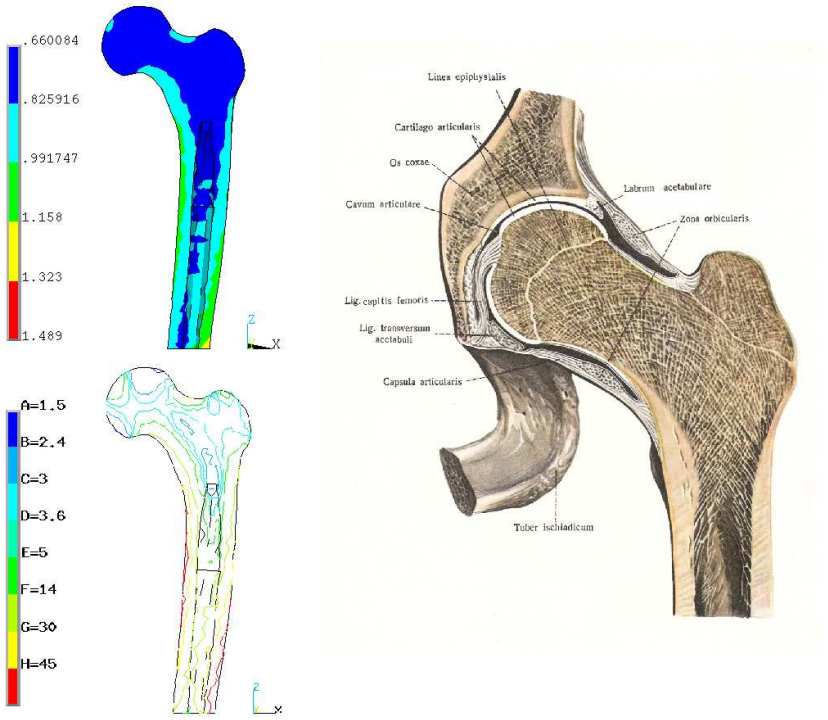


Figure 4: Cuts of a proximal femur - compare the calculated distribution of density in healthy case and isovalues of von Mises stress (probable direction of osteons in bone) with figure from anatomical atlas.

this of course causes a great changes in stress and strains magnitudes and its distribution in bone resulting in changes of density distribution.

On Fig.5 are depicted the X-ray pictures of human proximal femur right after operation, 6 years after operation, and density distribution predicted by the presented model. Great unknown in the joint-replacement problem is how will the bone respond in terms of remodelling to a new stress-strain field in bone after the replacement. Usually there is considerable resorption in the vicinity of implant (especially in proximal-medial and proximal-lateral part of femur) but in some cases there is also a significant deposition of bone in specific sites that strengthens the imposition of prosthesis in bone. Fig.5 shows one example when adequate physical activity (50-year-old man at the time of operation, approx. 10 thousand gaits per day) guarantees sufficient bone density for a long time (Fig.5b).

Our research group tries to give some insight into this problematics. Despite the complexity , when not only person-specific gene expression together with diet and activity that he performs but also the choice of material for prosthesis, angle of insertion and hollow created plays a role, the same type of response – the same pattern of density distribution - after month from operation is obtained as in clinical observation - Fig.5c.

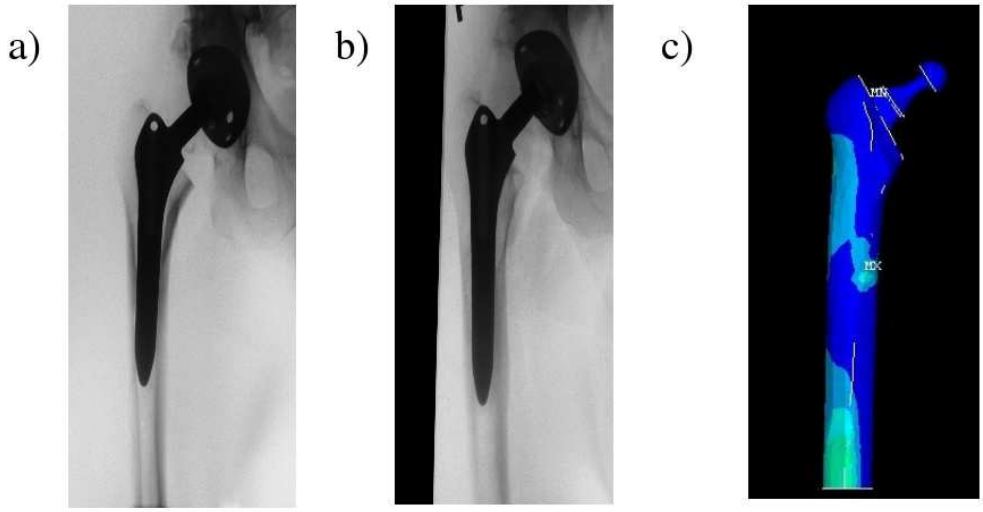


Figure 5: The change of human bone (right femur) density distribution after hip joint replacement. During the course of time the denser regions in the bone shift in the distal direction towards the implant tip. The bone is markedly thicker near the implant tip. a) X-ray image immediately after operation, b) X-ray image 6 years after operation, c) bone density evolution calculated by our method.

4 Discussion

There are nowadays several bone adaptation theories (remodelling models) but vast majority of them is based only on influence of mechanical loading [9, 20, 25, 27, 28, 31]. These models predict that bone mass will be zero if stresses in bone are zero. This is not in agreement with *in vivo* observances. When limbs are casted (and the bone loading is minimal), *bone is rapidly lost but it reaches new steady state with lower bone density.* Notice that the bone loss is not complete. This result suggests that there are some effects, e.g. hormones, nutrition, that are missing. On the other hand there are a few biological models [14] but with no influence of mechanical loading, which is as it is known very important stimuli for bone remodelling [2]. The magnitude and direction of the applied forces used in simulation are common for walking [1].

Model proposed in this paper combine both - the mechanical stimuli and also the biological background. Using FEM software we are able to calculate the density distribution and consequently also material properties (the dependency of E on density was experimentally determined as power to three [9]) throughout the whole bone. The calculated pattern is in great agreement with the knowledge of bone anatomy. Moreover, calculated isostress lines do correspond to osteon direction as [6, 22] claims. Using this approach we may simulate different loading cases (exercise, patients) but also nutrition or other biological features (and possibly the influence of several hormones involved in bone remodelling). These simulations (and possible predictions) are being studied nowadays.

As was shown in this article, locomotion or *dynamic* loading is crucial for correct bone development and remodelling. The only parameters that were varying throughout the bone were D_ρ - the effect of dynamical loading on rate of running chemical reactions.

Nevertheless the agreement with structure of bone is remarkable. Frequency of loading is also found as an very important factor.

The limitation of "thermodynamic BR"' model is mainly caused by the difficulty to adjust the different parameters. Up to now it would be uneasy to determine many of parameters experimentally even though they actually are characterising chemical reactions (1)-(5) (and thus they have a real meaning). Partial solution of this problem may be found in comparing calculated *results* for given set of parameters with real data from clinical observation. Parameters of model were chosen so that concentrations of all of the substances were positive and so that model showed in fundamental aspects the same behaviour as clinically observed. If the presented thermodynamic model fits on clinical data, it can be used for predictions or even treatment of skeletal disorders connected with bone remodelling, e.g. osteoporosis, osteomalatia and inborn skeletal defects called bone dysplasias. Some of these inborn defects of locomotor apparatus appear to us like an experiment of nature and making it possible to study pathobiomechanics of skeleton directly. Also other possible usage may be in designing implants of hip joints. Thus the application of presented thermodynamic bone remodelling model may reach the clinically broad domain.

This model was originally intended as a simplified model of bone metabolism modelling. But even such a model, which does not contain the detailed mechanisms of bone remodelling control, gives results that are very challenging.

Acknowledgement
I would like to thank to my supervisor Professor F. Maršík, Eng., D.Sc. for gratuitous passing of know-how, for advice, pleasant consultations, and help with writing this article.

This research has been supported by the Grant agency of the Czech Republic no. 106/03/1073 and by the project 1M06031 of Ministry of Education, Youth and Sports of Czech Republic.

References

[1] G. Bergmenn, F. Graichen, and A. Rohlmann. *Hip joint loading during walking and running, measured in two patients.* Journal of Biomechanics **26** (1993), 969–990. doi:10.1016/0021-9290(93)90058-M.

[2] P. J. Ehrlich and L. E. Lanyon. *Mechanical strain and bone cell function: A review.* Osteoporosis International (2002), 688–700.

[3] H. M. Frost. *Osteogenesis imperfecta. the set point proposal (a possible causative mechanism).* Clinical orthopaedics (1987), 280–296.

[4] H. M. Frost. *Tetracycline-base histological analysis of bone remodelling.* Calcif Tissue Res (1969), 211–237.

[5] H. M. Frost. *The mechanostat: a proposed pathogenetic mechanism of osteoporoses and the bone mass effects of mechanical and nonmechanical agents.* Bone and mineral (1987), 73–85.

[6] H. M. Frost. *The Utah paradigm of skeletal physiology*, volume second. ISMNI, Greece, first edition, (2004).

[7] J. Heřt, E. Přibylová, and M. Lišková. *Reaction of bone to mechanical stimuli. part 3: Microstructure of compact bone of rabbit tibia after intermittent loading.* Acta Anat **82** (1972), 218–230.

[8] R. Huiskes and S. J. Hollister. *From structure to process, from organ to cell: recent developments of fe-analysis in orthopaedic biomechanics.* Journal of biomechanical engineering (1993), 520–527.

[9] R. Huiskes, H. Weinans, H. Grootenboer, M. Dalstra, B. Fudala, and T. Slooff. *Adaptive bone-remodeling theory applied to prosthetic-design analysis.* Journal of Biomechanics **20** (1987), 1135–1150. doi:10.1016/0021-9290(87)90030-3.

[10] V. Klika. Mathematical and numerical analysis of differential equations of bone remodelling. Master thesis, Czech Technical University, Faculty of Nuclear Sciences and Physical Engineering, Prague, (May 2006).

[11] V. Klika, F. Maršík, and P. Barsa. *Remodelling of a living bone - numerical simulation.* Locomotor System **14** (2007), 112–117.

[12] V. Klika, F. Maršík, and V. Bobro. Vliv dynamické zátěže na mechanické vlastnosti kostí. In 'Locomotor System, Supplementum', volume 11, 120, (2004).

[13] V. Klika, F. Maršík, and I. Mařík. Vznik kortikális jako důsledek mechanické zátěže - osteoporóza jako důsledek biochemické nerovnováhy. In 'Locomotor System, Supplementum', volume 12, 17–18, (2005).

[14] V. Lemaire, F. Tobin, L. Greller, C. Cho, and L. Suva. *Modeling the interactions between osteoblast and osteoclast activities in bone remodeling.* Journal of Theoretical Biology (2004), 293–309. doi:10.1016/j.jtbi.2004.03.023.

[15] F. Maršík, V. Klika, and P. Barsa. Remodelling of a living bone - numerical simulation. In 'ICMOSPS proceedings on CD'. School of Mechanical Engineering UKZN, Durban, South Africa, (2007). ISBN: 1-86840-643-1.

[16] F. Maršík, V. Klika, I. Mařík, H. A. Bougherara, and L. Yahia. Bone remodeling induced by dynamical loading. In 'Book of Abstracts of the 1st International Conference on Mechanics of Biomaterials & Tissues', volume 11, 120, (December 2005).

[17] F. Maršík, I. Mařík, and V. Klika. *Chemical kinetics of bone remodeling based on RANK-RANKL-OPG biology, preliminary study.* Locomotor System **11** (2004), 266–269.

[18] F. Maršík, I. Mařík, and V. Klika. *Chemistry of bone remodeling processes.* Locomotor System **12** (2005), 51–61.

[19] T. Martin. *Paracrine regulation of osteoclast formation and activity: Milestones in discovery.* Journal of Musculoskel Neuron Interact (2004), 243–253.

[20] M. G. Mullender, R. Huiskes, and H. Weinans. *A physiological approach to the simulation of bone remodeling as a self-organizational control process.* Journal of Biomechanics **27** (1994), 1389–1394. doi:10.1016/0021-9290(94)90049-3.

[21] A. Parfitt. *Osteonal and hemi-osteonal remodeling: the spatial and temporal framework for signal traffic in adult human bone.* Journal of cellular biochemistry (1994), 273–286.

[22] M. Petrtýl, J. Heřt, and P. Fiala. *Spatial organization of the haversian bone in man.* Journal of Biomechanics **29** (1996), 161–169. doi:10.1016/0021-9290(94)00035-2.

[23] M. Petrtýl and J. Danešová. *Principles of bone remodelling - the limit cycles of bone remodelling.* Acta of Bioengineering and Biomechanics **3** (2001), 75–91.

[24] W. Roux. *Der kampf der teile im organismus.* Engelmann, Leipzig, (1881).

[25] R. Ruimerman, P. Hilbers, B. van Rietbergen, and R. Huiskes. *A theoretical framework for strain-related trabecular bone maintenance and adaptation.* Journal of Biomechanics (2005), 931–941. doi:10.1016/j.jbiomech.2004.03.037.

[26] Z. Sobotka and I. Mařík. *Remodelation and regeneration of bone tissue at some bone dysplasias.* Locomotor System **2** (1995), 15–24.

[27] K.-i. Tezuka, Y. Wada, and A. Takahashi. *Computer-simulated bone architecture in a simple bone-remodeling model based on a reaction-diffusion system.* Journal of Bone and Mineral Metabolism (2005), 1–7.

[28] C. H. Turner, V. Anne, and R. M. V. Pidaparti. *A uniform strain criterion for trabecular bone adaptation: Do continuum-level strain gradients drive adaptation?* Journal of Biomechanics **30** (1997), 555–563. doi:10.1016/S0021-9290(97)84505-8.

[29] H. Weinans, R. Huiskes, and H. Grootenboer. *The behaviour of adaptive bone-remodeling simulation models.* Journal of Biomechanics (1992), 1425–1441. doi:10.1016/0021-9290(92)90056-7.

[30] J. Wolff. *Das Gesetz Der Transformation Der Knochen.* A Hirchwild, Berlin, (1892). Translated as: The low of bone remodeling (Maquet P, Furlong R) Berlin: Springer, 1986.

[31] X. Zhu, H. Gong, and B. Gao. *The application of topology optimization on the quantitative description of the external shape of bone structure.* Journal of Biomechanics (2005), 1612–1620. doi:10.1016/j.jbiomech.2004.06.029.

An Effective Algorithm for Search Reductions in Compositional Models*

Václav Kratochvíl

2nd year of PGS, email: velorex@utia.cas.cz
Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU
advisor: Radim Jiroušek, Institute of Information Theory and Automation, AS CR

Abstract. This paper deals with the problem of marginalization of multidimensional probability distributions represented by a compositional model. By the perfect one in this case. From the computational point of view this solution is more efficient than any known marginalization process for Bayesian models. This is because the process mentioned in the paper in a form of an algorithm and takes an advantage of the fact that the perfect sequence models have some information encoded; if can be obtained from the Bayesian networks by an application of rather computationally expensive procedures. One part of that algorithm is marginalization by means of reduction. This paper describe a new faster algorithm to find a reduction in a compositional model.

Abstrakt. Marginalizace multidimenzionálních distribucí reprezentovaných perfektními kompozicionálními modely je mnohem efektivnější než jakýkoli marginalizační proces v bayesovských sítích. Důvod je prostý. Marginalizační algoritmus, zmíněný v tomto článku, využívá informací zakódovaných ve struktuře kompozicionálních modelů, které se v bayseovských sítích musí složitě vypočítat. V tomto článku se zabýváme jednou podsekcí marginalizačního algoritmu - marginalizací redukcí. Je zde představen nový rychlejší způsob hledání redukcí v kompozicionálních modelech.

1 Introduction

The ability to represent and process multidimensional probability distributions is a necessary condition for application of probabilistic methods in Artificial Intelligence. Among the most popular approaches are the methods based on Graphical Markov Models, e.g., Bayesian Networks. This paper deals with an alternative approach to Graphical Markov Models, so called the Compositional Models. The presented algorithm enables us to effectively compute marginal distributions from really multidimensional models.

One possible solution of this task for Bayesian Networks is given in papers by R. Shachter [6, 7]. His well known procedure is based on two rules: *node deletion* and *edge reversal*. Roughly speaking, the effectiveness of his algorithm is like the effectiveness of our algorithm without the accelerating procedures. This advantage consists in the fact that compositional models express explicitly some marginals, whose computation in the

*The research was partially supported by Grant Agency of the Academy of Sciences of ČR under grant A2075302, Austrian-Czech grant AKTION 45p16 and Ministry of Education of the Czech Republic under grants no 1M0572 and 2C06019.

Bayesian network may be computationally expensive. One of the accelerating procedures is marginalization by means of reduction. In this paper is presented an alternative algorithm to search such reduction. This algorithm is currently used in a system MUDIM[1].

2 Notation and Basic Properties

In this paper we will consider a system of finite-valued random variables with indices from a non-empty finite set N. All the probability distributions discussed in the paper will be denoted by Greek letters. For $K \subset N$, $\pi(x_K)$ denotes a distribution of variables $\{X_i\}_{i\in K}$.

Having a distribution $\pi(x_K)$ and $L \subset K$, we will denote its corresponding marginal distribution either $\pi(x_L)$, or $\pi^{\downarrow L}$. These symbols are used when we want to highlight the variables, for which the marginal distribution is defined.

To describe how to compose low-dimensional distributions to get a distribution of a higher dimension we will use the following operator of composition.

Definition 1. For arbitrary two distributions $\pi(x_K)$ and $\kappa(x_L)$ their *composition* is given by the formula

$$\pi(x_K) \rhd \kappa(x_L) = \begin{cases} \frac{\pi(x_K)\kappa(x_L)}{\kappa(x_{K\cap L})} & \text{when } \pi^{\downarrow K\cap L} \ll \kappa^{\downarrow K\cap L}, \\ \text{undefined} & \text{otherwise,} \end{cases}$$

where the symbol $\pi(x_M) \ll \kappa(x_M)$ denotes that $\pi(x_M)$ is *dominated* by $\kappa(x_M)$, which means (in the considered finite setting)

$$\forall x_M \in \times_{i\in M}\mathbf{X}_i \ (\kappa(x_M) = 0 \Longrightarrow \pi(x_M) = 0).$$

The result of the composition (if it is defined) is a new distribution. We can iteratively repeat the application of this operator composing a multidimensional model. This is why these multidimensional distributions are called *compositional models*. To describe such a model it is enough to introduce an ordered system of low-dimensional distributions $\pi_1, \pi_2, \ldots, \pi_n$. We denote this ordered system as a *generating sequence*, to which the operator is applied from left to right:

$$\pi_1 \rhd \pi_2 \rhd \pi_3 \rhd \ldots \rhd \pi_{n-1} \rhd \pi_n = (\ldots((\pi_1 \rhd \pi_2) \rhd \pi_3) \rhd \ldots \rhd \pi_{n-1}) \rhd \pi_n.$$

Then we say that a generating sequence defines (or represents) a multidimensional compositional model.

In the process of marginalization we will also need another important operator.

3 Perfect Sequence Models

Now the attention will be focused on marginalization of distributions given by a special subclass of generating sequences. From now, we will consider generating sequences

$$\pi_1(x_{K_1}) \rhd \pi_2(x_{K_2}) \rhd \ldots \pi_n(x_{K_n}).$$

Therefore, whenever distribution π_j is used, we assume it is defined for variables $\{X_i\}_{i\in K_j}$.

[1] Experimental system based on Multidimensional models.

Definition 2. We call a generating sequence $\pi_1, \pi_2, \ldots, \pi_n$ *perfect* if for all $j = 2, \ldots, n$

$$(\pi_1 \triangleright \ldots \triangleright \pi_{j-1}) \triangleright \pi_j = \pi_j \triangleright (\pi_1 \triangleright \ldots \triangleright \pi_{j-1})$$

hold true.

Perfect sequences have many pleasant properties, which are advantageous for multidimensional distributions representation. The most important one is expressed in the following assertion.

Theorem 3. *A generating sequence $\pi_1, \pi_2, \ldots, \pi_n$ is perfect* iff *all the distributions π_i are marginal to the represented distribution, i.e., for all $i = 1, 2, \ldots, n$*

$$(\pi_1 \triangleright \ldots \triangleright \pi_n)^{\downarrow K_i} = \pi_i.$$

Now, let us formulate universal rules which make it possible to decrease the dimensionality of compositional models by one. By iterative application of these rules may be obtained any required marginal. The proof of the following assertion, which holds not only for perfect but for all generating sequences, can be found in [5].

Theorem 4. *Let $\pi_1, \pi_2, \ldots, \pi_n$ be a generating sequence and*

$$\ell \in K_{i_1} \cap K_{i_2} \cap \ldots \cap K_{i_m}$$

for a subsequence $(i_1, i_2, \ldots, i_m)$ of $(1, 2, \ldots, n)$ such that $\ell \notin K_j$ for all $j \notin \{i_1, i_2, \ldots, i_m\}$. Then

$$(\pi_1 \triangleright \pi_2 \triangleright \ldots \triangleright \pi_n)^{-\{\ell\}} = \kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n,$$

where

$$\begin{aligned}
\kappa_j &= \pi_j, \qquad \forall j \notin \{i_1, i_2, \ldots, i_m\}, \\
\kappa_{i_1} &= \pi_{i_1}^{-\{\ell\}}, \\
\kappa_{i_2} &= (\pi_{i_1} \circledcirc_{L_{i_2-1}} \pi_{i_2})^{-\{\ell\}}, \\
\kappa_{i_3} &= (\pi_{i_1} \circledcirc_{L_{i_2-1}} \pi_{i_2} \circledcirc_{L_{i_3-1}} \pi_{i_3})^{-\{\ell\}}, \\
&\vdots \\
\kappa_{i_m} &= (\pi_{i_1} \circledcirc_{L_{i_2-1}} \pi_{i_2} \circledcirc_{L_{i_3-1}} \ldots \circledcirc_{L_{i_m-1}} \pi_{i_m})^{-\{\ell\}},
\end{aligned}$$

and $L_{i_k-1} = (K_1 \cup K_2 \cup \ldots \cup K_{i_k-1}) \setminus \{\ell\}$.

The iterative application of this theorem always leads to the desired marginal distribution.

More effective marginalizing procedures are, however, based on the different properties which reminds rather of graph's algorithms than statistical processes. One of them is denoted like *marginalization by reduction*

However, these effective marginalization procedures could be used only in a special case and, therefore, they are used as accelerating supplement only together with the general algorithm shown in Theorem 4.

First, let us define an auxiliary notion of a reduction of a generating sequence, which will simplify formulations in the following text.

Definition 5. Let $\pi_1, \pi_2, \ldots, \pi_n$ be a generating sequence, $(j_1, j_2, \ldots, j_m)$ a subsequence of $\{1, 2, \ldots, n\}$ and $s \in Z = \{j_1, \ldots, j_m\}$ be such that

$$(\bigcup_{j \in Z} K_j) \cap (\bigcup_{j \notin Z} K_j) \subseteq K_s.$$

Then s and Z determine a *reduction* of generating sequence $\pi_1, \ldots, \pi_n$ (or simply that (s, Z) is a reduction).

Theorem 6. *Let $s \in Z$ and $Z = \{j_1, \ldots, j_m\}$ determine a reduction of a perfect sequence $\pi_1, \pi_2, \ldots, \pi_n$. Then, denoting*

$$\bar{L}_j = \bigcup_{i \in \{1,\ldots,j\} \setminus Z} K_i; L = \bigcup_{j \in Z} K_j$$

for all $j \notin Z$, marginal distribution $(\pi_1 \triangleright \pi_2 \triangleright \ldots \triangleright \pi_n)^{\downarrow L}$ can be expressed

$$(\pi_1 \triangleright \pi_2 \triangleright \ldots \triangleright \pi_n)^{\downarrow L} = \kappa_1 \triangleright \kappa_2 \triangleright \ldots \triangleright \kappa_n,$$

where

$$\begin{array}{ll} \kappa_j = \pi_j & \text{for} \quad j \in Z, \\ \kappa_j = \pi_s^{\downarrow K_s \cap \bar{L}_j} & \text{for} \quad j \notin Z. \end{array}$$

A functional algorithm based on the proof of Theorem 6 was presented in [1].

The most difficult part of this algorithm is how to find the reduction (s, Z). An old solution is based on an iterative extension of the set Z with recounting $W(Z, j)$ in every step and testing validity of reduction (s, Z). This extension could be time-consuming and has to be done regardless of effect. (As mentioned above, reduction by Z does not have to exist) Because of this, the advantage gained by accelerating algorithm is wasted by searching the set Z. The possible solution of this situation is described at the end of the following section.

4 Marginalization Algorithm

Marginalization is one of the basic operations computing with multidimensional models.

Considering a perfect generating sequence $\pi_1(x_{K_1}), \pi_2(x_{K_2}), \ldots, \pi_n(x_{K_n})$ and a set of indices $L \subset (K_1 \cup K_2 \cup \ldots \cup K_n)$. The marginalization algorithm performs computation of a generating sequence $\kappa_1(x_{L_1}), \ldots, \kappa_m(x_{L_m})$, representing the required marginal distribution:

$$(\pi_1(x_{K_1}) \triangleright \pi_2(x_{K_2}) \triangleright \ldots \triangleright \pi_n(x_{K_n}))^{\downarrow L} = \kappa_1(x_{L_1}) \triangleright \kappa_2(x_{L_2}) \triangleright \ldots \triangleright \kappa_m(x_{L_m}).$$

The marginalization algorithm itself, as is implemented in MUDIM system, is depicted in [1]. The algorithm consists in (cyclical) employment of five procedures.

- Truncation of an unavailing tail,
- Deletion of redundant elements,
- Simple marginalization $[j]$,

- Marginalization by means of reduction,
- General marginalization $[j]$.

First four of them try to accelerate the computation. If these procedures fails, the general marginalization procedure will be applied. As the reader can see from theorem 4, the general marginalization procedure can be applied anytime but is very time consuming. Therefore, accelerating procedures such as marginalization by means of reduction are needful.

4.1 Marginalization by means of Reduction

The simple procedures could decrease dimension of the multidimensional distribution either by one, or by more than one, but only when the variables to be deleted appeared "in the tail" of the generating sequence. A more complex algorithm - Marginalization by means of Reduction, which proves to be very efficient in many situations, especially when the number of the variables to be deleted is really high, is described in [1]. This algorithm is not dependent on the order of distributions in a generating sequence. For this, one has to find a reduction (s, Z) such that Z contains all indices of the variables for which the computed marginal distribution should be defined ($L \subseteq Z$). And it is this search for reduction what makes the process rather complicated. There was published one algorithm in the paper [1]. Unfortunately, time demand factor of that algorithm is very high and this algorithm does not fulfil basic properties of accelerating procedures: If possible, accelerate computation. If not possible, do not cause any additional delay. This one is referred to as Full-Scan algorithm with regards to its structure and functionality.

4.2 Full-Scan algorithm

The structure of the Full-Scan algorithm is depicted in the Figure 1. It employs four relatively simple procedures described in [1]. Quite naturally, all these procedures work with the generating sequence in question $\pi_1(x_{K_1}), \pi_2(x_{K_2}), \ldots, \pi_n(x_{K_n})$ (but only one of them - Marginalization $[s, Z']$ - modifies it). To find a reduction, the process employs sets $W(Z, j)$ (defined below) and their properties, which were proven in [2].

Having a set $Z \subset \{1, \ldots, n\}$ and $j \notin Z$ the symbol $W(Z, j)$ denotes the following subset of indices:

$$W(Z,j) = \left\{ s \in Z : \left(\bigcup_{i \in Z} K_i \right) \cap K_j \subseteq K_s \right\}$$

(notice that sets $W(Z, j)$ depend not only on Z and j but naturally also on the considered generating sequence).

The basic idea of this algorithm is as follows: The algorithm extends set Z until there exists $s \in \bigcap_{j \notin Z} W(Z, j)$ or $Z = (1, .., n)$. Let us try to evaluate the time complexity of the particular steps of this algorithm. The complexity will be present in multiples of $|Z|$ of set functions.

- $W(Z, j)$: $O(|Z|)$

- Extension of Z: $O(|Z|^2)$
- Constr. of a conn. set $\bar{Z}$: $O(|Z|^3)$
- Constr. of a bridge $\hat{Z}$: $O(|Z|^3)$

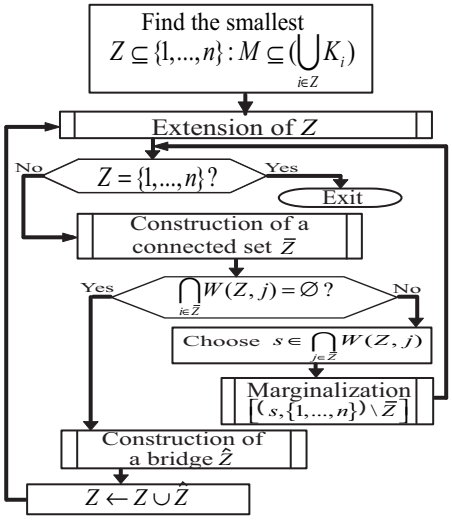


Figure 1: Marginalization by means of reduction

Because we know that $|Z|$ is a function of n where n is a number of distributions in a model; we can assume that $O(n^i) \approx O(|Z|^i), i \in \mathbb{N}$. The total complexity of this reduction search algorithm is **minimally** $O(n^3)$ of **set functions**. The space complexity of this algorithm has not been computed.

4.3 DFS algorithm

Let us present a slightly modified algorithm to find a reduction which does not discover all reductions but is very fast. This algorithm will be denoted as the DFS algorithm. The time and space complexity of Full-Scan algorithm were the reason why DFS algorithm was created. The rest of marginalization algorithm remains same as it was published in [1].

Marginalization by meas of reduction does not depend on the order of distributions in generating sequence. Therefore, if we ignore that order, we can treat a multidimensional model as a group of subsets (distributions). That is exactly how hypergraph can be defined. In other words, consider a generating sequence of compositional model as a hypergraph (see the definition and an illustration bellow). Reduction K_5 in a model 3 seems like a bridge in a hypergraph. We can convert the problem of finding a reduction (s, Z) to the problem of searching for a bridge in a hypergraph. Every bridge in a hypergraph corresponds to a reduction in corresponding multidimensional model. Nevertheless, there are reductions in a multidimensional model which can not be represented by bridges in a corresponding hypergraph.

Let us defined a hypergraph H on n vertices to be an ordered pair (V, E), where V is the set of vertices, with $|V| = n$, and E is a multiset of subsets(hyperedges) of V. For an

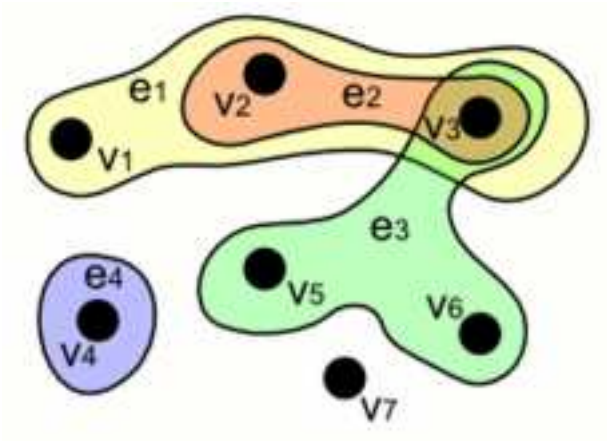


Figure 2: Simple Hypergraph: $V = \{v_1, v_2, v_3, v_4, v_5, v_6, v_7\}, E = \{e_1 = \{v_1, v_2, v_3\}, e_2 = \{v_2, v_3\}, e_3 = \{v_3, v_5, v_6\}, e_4 = \{v_4\}\}$

.

arbitrary hypergraph H, we let $v(H)$ denote the number of vertices of H and $e(H)$ denote the number of hyperedges of H. In general, we will consider hypergraphs labeled so that if the hypergraph has n vertices, they are labeled by the elements of $\hat{n} = 1, 2, 3, ..., n$, and if the hypergraph has m edges, they are labeled by the elements of $\hat{m}$. For simplicity, we will call such objects *labeled hypergraphs*.

We define a *walk* in a hypergraph to be a sequence $v_0, e_1, v_1, ..., v_{n-1}, e_n, v_n$, where for all i, $v_i \in V$, $e \in E$, and for each e_i, $v_{i-1}, v_i \subseteq e_i$. We define a *path* in a hypergraph to be a walk in which all v_i are distinct and all e_i are distinct. A walk is a *cycle* if the walk contains at least two edges, all e_i are distinct, and all v_i are distinct except $v_0 = v_n$.

A hypergraph is *connected* if for every pair of vertices v, v' in the hypergraph, there is a path starting at v and ending at v'. The hypergraph in Figure 2 is not connected in contrast to the hypergraph in Figure 3. More information about hypergraphs can be found in [4].

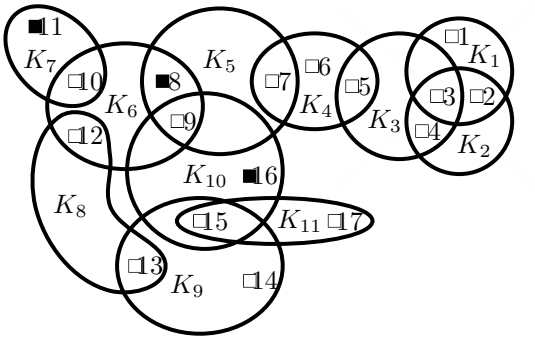


Figure 3: Visualization of a Multidimensional Model

Definition 7. A hyperedge in connected hypergraph is called a bridge if the deletion of that hyperedge disconnects the hypergraph.

Whilst an effective algorithm of searching for bridges in hypergraphs does not exist,

the algorithm for an articulation point does. For that we convert the hypergraph into a simple graph. We define a graph G on n vertices to be an ordered pair (V, E), where V is the set of vertices, with $|V| = n$, and E is a subset $E \subseteq V \times V$. Let connected graph be defined as usual. Exact definition can be found in [3].

Definition 8. A node in a connected graph is called an articulation point if the deletion of that node disconnects the graph.

To convert the generating sequence (hypergraph) to a respective graph, we are using the *representative graph of hypergraph* defined below.

Definition 9. Given a hypergraph $H = (V, E)$, its representative graph $G = (E^*, E^{**})$ is a graph whose vertices are points $e_1^*, ..., e_m^*$ representing the edges $e_1, ..., e_m$ of H, the vertices e_i^*, e_j^* being adjacent if and only if $e_i \bigcap e_j \neq \emptyset$.

Because not every reduction can be represented by a bridge in a hypergraph, the modified algorithm does not discover all reductions in a model. We may loose some reductions during the conversion of a reduction search problem to a bridge search problem. Nevertheless, our measuring of time consumption proves that this loss is sufficiently compensated.

Hence, considering a multidimensional model $(\pi_1, ..., \pi_n)$ as a hypergraph, we convert it into its representative graph $G(V, E)$ as follows:

1. Each distribution is considered to be a vertex $(\pi_1, ..., \pi_n \rightarrow V)$
2. $(\pi_i, \pi_j) \in E \Leftrightarrow (K_i \bigcap K_j \neq \emptyset)$ (if two arbitrary distributions in the generating sequence have non-empty intersect, then we join them by an edge.)

Then, G is the representative graph of the multidimensional model $(\pi_1, ..., \pi_n)$.

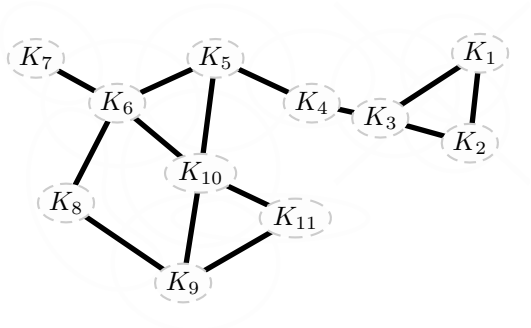


Figure 4: Representative graph

When we apply this conversion procedure to the model visible in Figure 3 we gain the graph as in Figure 4.

Both problems of searching a reduction and searching a bridge in the hypergraph are equivalent in a way. Both bridge and articulation point split a graph (hypergraph) into two independent parts. Nevertheless, if we think about reduction as we did, several reductions could be lost during the searching of bridges. We consider this loss as

unimportant. After all, the reduction search procedure should serve as an accelerating procedure and therefore it should be as fast as possible.

The conversion is justified (in light of reduction) by the following lemma.

Lemma 10. *Let Pi_s is an arbitrary but fixed articulation found in a representative graph derived from generating sequence. Then there exists a corresponding reduction (Z, s) in generating sequence of multidimensional model where Z is given by representative graph.*

Proof: Let π_s be an articulation in the *representative graph* G of a multidimensional model. By removing vertex π_s from graph G, the graph is split into several independent parts. Let denote them Z and the rest as $\hat{n} \setminus Z$. Then $(\bigcup_{j \in Z} K_j) \cap (\bigcup_{j \notin Z} K_j) \subseteq K_s$ holds because π_s is the articulation point. This formula is consistent with Definition 5 and the lemma is proved. □

The whole procedure of marginalization by means of reduction employs three simple steps that we are going to describe now. You can see that this approach is much easier than the old one, regrettably, it is not so powerful.

1. The generating sequence (considered as a hypergraph) is read from the model as a simple graph. Data are returned as an associative array which presents the simple graph as an adjacency list[2].

2. The adjacency list is processed by standard **DFS procedure**. (That is the reason why the new approach is referred as DFS algorithm)

3. All found articulations are deleted according to Lemma 1.

The complexity of the DFS algorithm is computed in basic functions, in contradiction to the previous algorithm.

- Conversion to simple graph: $O(n)$
- DFS procedure: $O(n + m)$

A total (time and space) complexity of this algorithm is $O(n + m)$. It is much less then in previous algorithm.

5 Conclusions

In this paper a slightly modified algorithm for marginalization in compositional models is described. More precisely, an algorithm for models which are represented by perfect sequences. The algorithm is based on theoretical properties proven in several assertions published in [2]. The algorithm is currently realized in the system MUDIM[3]. Its efficiency is being tested on artificially generated data.

In Table 1 we refer to computations with two models constructed for artificially generated data. Let us stress that it would be easy to construct a model, for which the

[2] An *adjacency list* is the representation of all edges or arcs in a graph as a list.

[3] The system is realized in **R** and **C** language

Table 1: Computational time in seconds

	OFF	DFS	FullScan
Model 1 marginalization from 24 to 3 variables	1.67 s	0.53. s	0.95 s
Model 1 marginalization from 24 to 3 variables	1.39 s	1.38 s	2.15 s
Model 2 marginalization from 30 to 3 variables	>30 s	4.21 s	4.02 s
Model 2 marginalization from 30 to 3 variables	>30 s	15.56 s	12.25 s

reduction substantially decreases the computational time. Nevertheless, on purpose we are presenting examples which are, in a way, from this point of view inconvenient. They represent distributions of 24 and 30 variables. The difference between the first two rows of Table 1 shows that the efficiency of the process does not depend only on the model – probability distribution, but also on which variables are to be marginalized out. This also explains the difference between the 3rd and 4th rows. First model illustrates the following situation: Reduction exists in the first row while no reduction can be found for the variables in the second row. The delay caused by Full-Scan algorithm is for this number of variables approximately $0,5$s. DFS algorithm does not delay the process compared to the speed when the reduction is switched off. (The lower time is caused by incorrectness of measuring tool.) The following model observed this situation: DFS algorithm did not discover the same reduction as the Full-Scan algorithm. Nevertheless, a computation is still faster than without this accelerating procedure. The last row illustrates a similar situation.

During the testing of the original algorithm published in [1] several problems appeared. The application of the accelerating reduction procedure (realized by Full-Scan algorithm) was very time-consuming because of its nature and also because of the *reduction condition*. With the help of hypergraphs another approach was discovered. The created DFS algorithm is not as powerful as the original Full-Scan algorithm but can be briskly performed by computer.

The way of treating multidimensional models as hypergraphs opens new extensions

in the theory of multidimensional models for the future. Nevertheless, in this paper hypergraphs are only a bridge to connect the multidimensional model to the articulation in simple graphs. In fact, the hypergraphs are not used in DFS algorithm.

References

[1] V. Kratochvíl, R. Jiroušek: Marginalization Algorithm for Compositional Models. In: Information Processing and Management of Uncertainty in Knowledge-based Systems. 3 Éditions EDK, Paris 2006, pp. 2300-2307.

[2] Vl. Bína, R. Jiroušek: Marginalization in Multidimensional Compositional Models. Submitted for publication to Kybernetika.

[3] R. Merris: Graph Theory. Wiley Interscience, New York 2001.

[4] C. Berge: Graphs and Hypergraphs. North Holland, Amsterdam 1973.

[5] R. Jiroušek: Marginalization in composed probabilistic models. In: Proc. of the 16th Conf. Uncertainty in Artificial Intelligence UAI'00 (C. Boutilier and M. Goldszmidt eds.), Morgan Kaufmann Publ., San Francisco, California 2000, pp. 301–308.

[6] R. D. Shachter: Evaluating Influence Diagrams. Oper. Res. **34** (1986), 871–890.

[7] R. D. Shachter: Probabilistic inference and influence diagrams. Oper. Res. **36** (1988), 589–604.

[8] Hypergraph. World wide web document.
http://en.wikipedia.org/wiki/Hypergraph

Mapování schémat v prostředí Sémantického webu*

Zdeňka Linková

3. ročník PGS, email: linkova@cs.cas.cz
Katedra matematiky, Fakulta jaderná a fyzikálně inženýrská, ČVUT
školitel: Július Štuller, Ústav informatiky, AV ČR

Abstract. The paper deals with steps of the non-materialized data integration, and focuses on schema matching and schema mapping issues. The proposal is for data sources on the Semantic Web; the crucial assumption for the considered task is the availability of the ontologies describing data to integrate. These ontologies are used to find correspondences between source schemas elements, and also for found mapping expression.

Abstrakt. Článek se zabývá úlohami, které je třeba řešit při nematerializované integraci dat. Zaměřuje se na hledání korespondencí mezi schématy a mapování schémat. Návrh přístupu řešení těchto úloh na Sémantickém webu těží z dostupných ontologiích popisujících integrované zdroje. Ontologie jsou využity jak k hledání mapování, tak i při jejich popisu.

1 Úvod

Integrace dat [1] je úloha, která se zabývá sloučením dat. Jejím cílem je prezentovat data pocházející z různých datových zdrojů jako jediný celek a umožnit je zpracovávat, jako by z jediného datového zdroje opravdu pocházela. V případě tzv. nematerializovaného přístupu [25] bývá řešením úlohy poskytnutí unifikovaného pohledu na zdroje dat. Tento pohled je využíván jako nový zdroj obsahující všechna data. Ve skutečnosti jde o pohled virtuální a data zůstávají fyzicky uložena v původních zdrojích.

Aby bylo možné integraci založit na využití virtuálního pohledu, neboli aby bylo možné k datům přes tento pohled přistupovat, je nezbytné definovat jeho vazby na fyzická data. Proto je třeba se v tomto přístupu zabývat schémata dat. Vazby mezi pohledem a daty se pak zajistí definováním vztahů mezi jednotlivými částmi schématu pohledu a částmi schémat původních zdrojů. Ty jsou pak dále využity při zpracování dat, například při dotazování.

Proces integrace je možné nahlížet jako kolekci několika úloh, které společně přinášejí požadovaný výsledek. Základní kroky při řešení integrace dat pomocí virtuálního pohledu jsou:

- **Hledáni korespondencí mezi schématy** (schema matching) - Za předpokladu, že datové zdroje, které mají být integrovány, byly vytvořeny nezávisle, různými designéry

*Práce byla podpořena projektem 1ET100300419 programu Informační společnost (Tématického programu II Národního programu výzkumu v ČR: "Inteligentní modely, algoritmy, metody a nástroje pro vytváření sémantického webu") a výzkumným záměrem AV0Z10300504 "Informatika pro informační společnost: Modely, algoritmy, aplikace".

a pro různé účely, jsou jejich schémata obecně heterogenní. Proto je důležitou úlohou nalezení jejich vzájemných korespondencí. Problém hledání korespondencí mezi schématy bývá označován jako schema matching [20], [21].

- **Mapování schémat** (schema mapping) - Obvyklým způsobem jak vyjádřit nalezené souvislosti mezi schématy zdrojů je použití tzv. *mapování*. Mapování je struktura, např. množina tvrzení, která popisuje vazbu mezi elementy schématu pohledu (obvykle označovaného jako globální schéma) a schémat datových zdrojů (označovaných jako lokální schémata).
  Při tvorbě mapování jsou využívány dva základní přístupy [14], [3]: *Global As View* (GAV) přístup, který spočívá v definování globálního schématu jako množiny pohledů nad lokálními schématy, a *Local As View* (LAV) přístup, který definuje lokální schéma zdroje pomocí pohledů nad globálním schématem. Je samozřejmě možné oba přístupy kombinovat.

- **Zpracování dotazů** (query processing) - Vytvoření mapování je stěžejní úloha, jejíž výsledek má důležitou roli při přístupu k datům pomocí dotazů.
  Při použití systému, který integruje data, klade uživatel dotazy tvořené nad poskytovaným pohledem, tj. využívá jeho jazyk, schéma atd. Pro vyhodnocení dotazu nad daty je třeba původní (globální) dotaz nějakým způsobem zpracovat [18].
  Tím se zabývají dva základní přístupy. Prvním je *query rewriting* - dotaz je dekomponován na části odpovídající lokálním zdrojům. Ty jsou dále přepsány tak, aby byly vyjádřeny v prostředí příslušného lokálního zdroje. Nad lokálními zdroji jsou pak vzniklé lokální dotazy vyhodnoceny a ze získaných lokálních odpovědí je opět sestavena globální odpověď, která je vrácena jako výsledek na původní dotaz.
  Druhou možností je *query answering*, která nijak nespecifikuje, jak má být daný dotaz zpracován. Jejím cílem je využít všechny dostupné informace k získání odpovědi na dotaz. Příkladem může být hledání takových dat, u nichž lze dle dostupných znalostí usuzovat, že jsou hledaným výsledkem.

Tento článek se dále zabývá prvními dvěma kroky integrace, tedy hledání korespondencí mezi schématy a jejich popisem pomocí vhodné struktury. Zaměřuje se na datové zdroje Sémantického webu.

Sémantický web [2], [13], [8] je zamýšlen jako sémantické rozšíření webu současného. V současné době jsou hlavními technikami při popisu dat Sémantického webu především:

- jazyk XML [33] pro strukturování dat
- RDF(S) [30], [31] pro popis metadat
- OWL [28] pro specifikaci ontologií.

Omezení na data Sémantického webu spočívá v požadavku vyjádření dat pomocí RDF/XML a dostupných ontologií [6] popisující jednotlivé zdroje.

Článek je členěn následovně: Kapitola 2 představuje obecně úlohu hledání korespondencí spolu s přístupy, které se touto úlohou zabývaly. Kapitola 3 se zaměřuje na ontologický přístup hledání korespondence mezi schématy na Sémantickém webu. Vyjádření mapování se věnuje Kapitola 4.

2 Úloha hledání korespondencí mezi schématy

Vstupem v úloze hledání korespondencí jsou dvě schémata, mezi nimiž je třeba nalézt vzájemné vztahy. Tato úloha je předmětem mnoha výzkumů. Bohužel však je v různých projektech či implementacích řešena především manuálně [15], tzn. je založena na lidském zásahu, uživatel - člověk je ten, kdo vztahy nalezne. To s sebou přináší mnohá omezení, je to například časově náročné, náchylné k chybám, drahé. Přirozená snaha o zautomatizování prováděné operace má ovšem většinou za následek pouze tzv. *kandidáty* možných korespondencí a je to opět člověk, kdo musí rozhodnout, zda nalezené možné korespondence skutečně platí.

Způsoby, kterými bývají korespondence hledány, lze rozdělit na základě úrovně informací, které jsou při porovnávání schémat využívány:

- **Na úrovni instancí** - Srovnávací přístupy pracují s vlastními daty ze zdrojů, aby nalezly korespondence mezi jejich schématy.
- **Na úrovni použitých pojmů** - Přístupy pracující na této úrovni bývají lingvisticky založené (např. jsou založené na jménech a textových popisech elementů schémat). Mohou pracovat se známými vztahy mezi použitými pojmy (synonyma, homonyma, apod.) nebo mohou pojmy zpracovávat jako řetězec znaků (a využívat vztahů jako je prefix, sufix, kořen apod.)
- **Na úrovni struktury** - Při hledání korespondencí (především mezi schématy, které mají složitější strukturu) bývá brána v úvahu i vlastní struktura zdroje. K porovnávání struktur mohou být využity například metody z oblasti teorie grafů.

Tyto techniky mohou být samozřejmě kombinovánay. Například při porovnávání jednotlivých elementů schémat je možné brát v úvahu jak jejich jména, datové typy, aktivní domény, ale i jejich strukturu.

Možnost existence mapování, ke kterému se nalezený kandidát vztahuje, bývá často vyjádřena pomocí nějaké funkce, která podobnost porovnávaných elementů vyjadřuje. Je možné ji založit na pravděpodobnosti [16], kosinové míře příznakových vektorů [23], nebo míře vyjadřující počet shodných zkoumaných aspektů [27]. Použitá míra může být využita při výběru skutečných korespondencí z kandidátů, čímž je možné lidský zásah více eliminovat. Někdy jsou navíc použity i další techniky, jako například zpřesňování kandidátů [7] či machine learning [26].

3 Hledání korespondencí mezi schématy na Sémantickém webu

V prezentovaném přístupu se předpokládá, že spolu s integrovanými zdroji jsou k dispozici také ontologie, které popisují data uložená ve zdrojích. Pomocí nich jsou vyvozovány požadované korespondence mezi jednotlivými elementy schémat. Jelikož je přístup orientován na Sémantický web, předpokládá, že ontologie zdrojů jsou vyjádřeny v jazyce OWL [28].

V obecném případě může jeden element korespondovat s jedním nebo více jinými elementy, může korespondovat s kombinací elementů, nebo nemusí korespondovat s žádným

jiným elementem. V této souvislosti se obvykle při hledání korespondencí používá pojem *kardinalita*, která pro určitou korespondenci vyjadřuje, kolik elementů mapovaných schémat do vztahu vstupuje. Kardinalita korespondence může být 1:1, 1:N, N:1, N:M. Ovšem většina existujících přístupů využívá kardinalit 1:1 or 1:N.

Prezentovaný přístup uvažuje vztahy následujících kardinalit:

- **1:1** - při vzájemném porovnávání dvou schémat. Tento případ vyjadřuje, že element jednoho schématu je ve vztahu s jedním elementem druhého schématu.

- **1:N** - při porovnávání jednoho schématu s více dalšími schématy. Tento případ je možné vidět jako množinu korespondencí kardinalit 1:1. Kardinality 1:N se často využívá v integrci dat pro vyjádření korespondencí mezi schématem globálního virtuálního pohledu a schématy lokálních zdrojů.

Pojetí korespondence při porovnávání schémat je formalizováno následovně:

- **Korespondence** kardinality **1:1** je tvrzení:

$$\varepsilon_1 \ \rho \ \varepsilon_2$$

kde

ε_1 je element jednoho schématu
ε_2 je element druhého schématu
ρ je vztah mezi ε_1 a ε_2, který vyjadřuje jejich vzájemnou korespondenci.

- **Korespondence** kardinality **1:N** je množina tvrzení kardinalit 1:1:

$$\{\varepsilon_1 \ \rho_i \ \varepsilon_i\}$$

kde

ε_1 je element jednoho schématu
ε_i je element druhého schématu
ρ_i je vztah mezi ε_1 a ε_i, který vyjadřuje jejich vzájemnou korespondenci.

Vztahem ρ mohou být následující druhy korespondencí:

- **Is-a** hierarchický vztah (tj. jeden element je obecnější než druhý, nebo naopak). Tento druh je označen jako $\supseteq$, resp. $\subseteq$.

- **Ekvivalence** mezi elementy.
Tento druh je označen jako $=$.

- **Disjunktnost**, tj. mezi elementy není žádná souvislost.

3.1 Hledání korespondencí v případě sdílené ontologie

V nejjednodušším případě je popis všech zdrojů dostupný v jediné ontologii. Tato ontologie je lokálními zdroji sdílena a pokrývá popis všech lokálních dat. Vztahy mezi elementy jednotlivých schémat mohou být nalezeny přímo v této ontologii.

Pro to je použito pravidlo:

Sémantický vztah mezi pojmy definovaný v ontologii implikuje stejný vztah mezi elementy schemat, které jsou těmito pojmy označené.

Uvažujeme-li dříve zmíněné typy korespondencí, je možné přístup založit na *is-a hierarchii* definované sdílenou ontologií. Jsou-li porovnávána dvě schémata, pro každý element jednoho schématu a každý element druhého schématu je jejich vztah hledán v této ontologii. Je-li mezi nimi vztah nelezen, je příslušná korespondence i mezi uvažovanými elementy.

Některé vztahy nemusí být v ontologii vyjádřeny přímo, ale je možné je z ontologie získat využitím tranzitivity is-a vztahu. Je-li například použit přístup k ontologii jako grafu s třídami popisujícími jednotlivé pojmy jako uzly a s orientovanými hranami vyjařujícími existenci is-a vztahu, nalezenou korespondenci neznamená pouze existující hrana, ale také příslušně značená cesta.

V případě, že jsou elementy disjuntkní, znamená to, že by v is-a hierarchii neměla být žádná cesta a není tedy nutné nějaký vztah hledat. V praxi vede tato situace ke stejnému efektu, jako když je vztah hledán, ale žádný není nalezen. Ovšem je vhodné tuto informaci o disjunktnosti dále uchovávat, protože může být dále využita při rozšiřování přístupu například o další usuzovavání apod.

Všechny korespondence, které jsou ze sdílené ontologie získány, jsou přijaty. Není na ně nahlíženo nejprve jako na kandidáty, neboť zde není žádný odhad korespondencí - všechny z nich jsou v dané ontologii definovány. Tento krok tedy nevyžaduje žádný zásah (lidského) uživatele.

3.2 Obecný případ hledání korespondencí založený na ontologiích

Obecně nemusí být ontologie, která by popisovala všechna zpracovávaná data, dostupná. Některé zdroje mohou sdílet některé pojmy, avšak sdílení všech pojmů všemi zdroji nelze předpokládat. Je třeba pracovat obecně s více ontologiemi.

Sloučením všech ontologií, které popisují integrované datové zdroje, získáme "novou" sdílenou ontologii, a tak je tento obecný případ převeden na předchozí. Sloučováním ontologií se zabývá řada výzkumů z oblastí ontology alignment a ontology merging a je tedy pro toto možné využít některou ze známých metod.

V souvislosti s ontologemi, pojmy alignment a merging spolu úzce souvisí [10]. Pro oba jsou také relevantní úlohy hledání korespondencí (matching) a mapování (mapping). *Ontology alignment* obvykle označuje stanovení binárních vztahů mezi dvěma ontologiemi. To umožňuje definovat způsob, jak tyto ontologie sloučit. Výsledkem *ontology merging* je nová integrovaná ontologie.

Metody ontology alignment a ontology merging jsou, podobně jako metody při porovnávání schémat, provozovány na několika úrovních: *instance* (např. srovnání množiny instancí popisovaného pojmu), *element* (např. lexikální techniky) a *struktura* (např. grafové techniky), a také využívají nejen sémantické, ale i syntaktické přístupy.

V obou oblastech lze najít i podobnost s používání kandidátů. Metody vyžadují lidskou interakci nebo jsou založeny na heuristikách z předchozích rozhodnutí. Ačkoliv při odvozování vztahů schémat ze sdílené ontologie žádní kandidáti nevznikají a korespondence jsou přímo určeny, v obecném případě mohou vznikat právě při využívání existujících metod při řešení podúlohy jak sdílenou ontologii najít.

Je patrné, že metody pro hledání korespondencí v ontology merging a ontology alignment jsou založeny na podobných principech jako metody pro hledání korespondencí mezi schématy. Důvodem toho je, že ontologie a datová schémata spolu úzce souvisí. Hlavním důvodem je účel, ke kterému jsou použity. Ontologie jsou vytvářeny, aby popisovaly pojmy používané v nějaké oblasti, zatímco schémata jsou vytvářena, aby modelovala nějaká konkrétní data. Speciálně pro schémata využívající sémantický model není často patrný rozdíl a není zřejmý způsob, jak identifikovat, která reprezentace je schéma a která je ontologie. V praxi mají často schémata i ontologie dobře definované použité pojmy. Protože schémata obecně neposkytují explicitní sémantiku pro data, používají se při hledání korespondencí techniky, pomocí nichž se odhaduje význam užívaných pojmů. Předpokládáme-li, že datové zdroje jsou popsány v dostupných ontologiích, použití takových technik není nutné, neboť potřebnou informaci máme.

Metodymi pro ontology merging, jež je například možné při hledání sdílené ontologie použít, se zabývá mnoho výzkumných projektů:

- **Chimaera** [12] - Systém Chimaera poskytuje nástroj pro slučování ontologií. Je založen na ontologickém editoru Ontolingua [9]. Uvažuje pouze hierarchický is-a vztah. Chimaera je interaktivní nástroj, který vyžaduje interakci uživatele: generuje seznam pojmů (kandidátů pro vztah), což pomáhá uživateli při určování pojmů ke sloučení. Chimaera ponechává rozhodnutí plně na uživateli, sám nenabízí žádné návrhy.

- **PROMPT** [17] - PROMPT je algoritmus pro semiautomatické sloučení ontologií. Provádí některé akce automaticky. Také determinuje možné nekonzistence plynoucí z uživatelových rozhodnutí a nabízí, jak je vyřešit.
  PROMPT nejprve vytvoří iniciální seznam pro korespondence založený na pojmech. Následuje cyklus výběru kandidátů uživatelem a automaticky prováděných akcí - algoritmus využívá datové typy, lingvistické techniky a is-a hierarchii.
  Algoritmus PROMPT byl implementován jako rozšíření ontologického editoru Protégé-2000 [29].

- **FCA-MERGE** [22] - FCA-MERGE je metoda pro slučování ontologií, která nabízí strukturální popis. Pro zdrojové ontologie extrahuje instance z relevantních textových dokumentů dané domény a aplikuje techniky zpracování přirozeného jazyka.
  Po extrakci instancí, jsou použity techniky FCA (Formal Concept Analysis) [19] a je získán strukturální výsledek FCA-MERGE. Extrakce instancí a FCA-MERGE algoritmus jsou plně automatické. Vygenerovaný výsledek je transformován do sloučené ontologie se zásahem uživatele.

- **HCONE** [11] - Přístup HCONE využívá WordNet [32], externí informační zdroj. HCONE z WordNetu získává lexikální informace. Lingvistické a strukturální informace o ontologiích jsou získány pomocí metody LSI - Latent Semantics Indexing

[5]. Jednotlivé koncepty jsou asociovány s jejich neformálními, lidsky orientovynými interpretacemi z WordNetu.
Metoda překládá formální definice pojmů do běžného slovníku, které pak vyšetřuje s využitím deskripční logiky. Cílem je ověřit mapování mezi ontologiemi a najít minimální množinu axiomů pro výslednou sloučenou ontologii. Není plně automatický, lidský zásah je nutný v počátečních fázích procesu.

4 Mapování na Sémantickém webu

Výsledek úlohy hledání vzájemných vztahů mezi schématy, tedy nalezené korespondence, se často označuje jako mapování. Obecně může mapování představovat libovolná struktura. K vyjádření mapování lze použít od jednoduchých 1-1 mapovacích pravidel vyjadřujících přímou korespondenci mezi elementy, přes mapování konceptu na dotaz nebo pohled [4], až po pomocné mapovací struktury (například referenční model v [24]). Různé projekty obvykle používají vlastní pojetí mapování.

Kromě například používání mapovacích pravidel jako tvrzení pro elementy globálních a lokálních schémat, které jsou orientovány na konkrétní řešenou úlohu, je možné využít složitější a dokonce standardizovanou strukturu, jenž by pokrývala všechna mapování. K popisu mapování mezi elementy schématu globálního pohledu a schémat lokálních zdrojů bude sloužit *ontologie OWL*.

Užití ontologie pro mapování přináší možnost znovupoužití také v jiných úlohách či situacích. Je také možné při odvozování dalších korespondencí, například při integrování dalšího zdroje, využít mapování v ontologii jako další ontologii, která integrované zdroje popisuje. Tak je možné dále využívat již jednou zjištěné skutečnosti. Navíc, bude-li v budoucnu třeba zachytit i další typy vztahů mezi elementy, může být ontologie dále využita, neboť je schopna zachytit různé typy vztahů.

K popisu mapování bude v závislosti na typu vztahu využit odpovídající [28] konstrukt. Abstraktním mechanismem pro seskupování popisovaných zdrojů v OWL je třída (class). Zdrojem je na webu jakákoli identifikovatelná entita. Proto bude pojetí `owl:Class` použito pro korespondenci elementů:

- **Is-a** hierarchický vztah, tj. $\varepsilon_1 \subseteq \varepsilon_2$, lze vyjádřit pomocí podtříd.
  Příslušným rysem OWL je `rdfs:subClassOf`, který umožňuje vyjádřit, že extenze popisu jedné třídy je podmnožinou extenze popisu jiné třídy.

- Vztah **ekvivalence**, tj. $\varepsilon_1 = \varepsilon_2$, lze v OWL vyjádřit s `owl:equivalentClass`.
  `owl:equivalentClass` umožňuje vyjádřit, že dvě třídy mají stejnou extenzi.
  V tomto případě může být také použit `rdfs:subClassOf` tak, že definujeme ε_1 jako podtřídu třídy ε_2 a současně ε_2 jako podtřídu třídy ε_1, říkáme, že ε_1 a ε_2 jsou ekvivalentní třídy.

- **Disjunktnost** (neboli tvrzení, že extenze popisu jedné třídy nemá žádné společné prvky s extenzí popisu jiné třídy) lze vyjádřit pomocí `owl:disjointWith`.

K vyjádření mapování slouží ontologie OWL. Zdrojem, ze kterého je mapování získáváno je ontologie sdílená zdroji, také ontologie OWL. Daná sdílená ontologie je"nadontologií" hledané ontologie v tom významu, že popisuje všechny třídy a jejich vztahy obsažené v mapování.

5 Shrnutí a závěr

Hledání korespondencí mazi schématy (schema matching) je stěžejní částí integračního procesu. Jeho výsledkem je mapování, které je dále využíváno při zpracovávání integrovaných dat. Při hledání vztahů je možné využít různých technik založených na různých informacích o datech. Jsou-li dostupné ontologie zdrojů, je možné odvodit hledané korespondence také z nich.

Důležitou otázkou je také způsob, jak nalezené mapování zaznamenat. V popsaném přístupu je k tomuto využita ontologie OWL. To přináší možnost mapování sdílet či znovu používat. Navíc mapování, které je vyjádřeno pomocí standardizované struktury může být dále využíváno i v jiných situacích a lze jej zpracovávat různými nástroji. Pro toto mapování je například možné používat metody vyvinuté pro zpracovávání ontologií.

Je-li k dispozici jediná ontologie, která popisuje data v integrovaných datových zdrojích, lze mapování v podstatě snadno získat přímo z této ontologie. V obecném případě, kdy je pro popis dat využito více ontologií, jsou tyto ontologie integrovány. Výsledkem integrace ontologií je sdílená ontologie a úloha je převedena na předchozí případ. Tímto způsobem je úloha hledání korespondencí mezi schématy převedena na úlohu slučování ontologií, pro kterou je možné využít některou z dostupných metod.

Mapování schémat založené na ontologiích je podúlohou celého procesu integrace. V budoucnu je proto plánováno zaměřit se také na následující fázi, tj. využití mapování pro zpracování dotazů.

Literatura

[1] Z. Bellahsene, "Data integration over the Web", *Data & Knowledge Engineering*, 44 (2003), pp. 265-266.

[2] T. Berners-Lee, J. Hendler and O. Lassila, "The Semantic Web", *Scientific American*, vol. 284, 5, pp. 35-43, 2001.

[3] A. Cali, D. Calvanese, G. De Giacomo, and M. Lenzerini, "On the Expressive Power of Data Integration Systems", In Proceedings of the *21st Int. Conf. On Conceptual Modeling (ER 2002)*, LNCS 2503, Springer, pp. 338-350, 2002.

[4] D. Calvanese, G. De Giacomo, and M. Lenzerini, "Ontology of integration and integration of ontologies", In Proceedings of the *2001 Description Logic Workshop (DL 2001)*, 2001.

[5] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman, "Indexing by Latent Semantic Analysis", *Journal of the American Society of Information Science* 41(6), pp. 391-407, 1990.

[6] Y. Ding, D. Fensel, M. Klein, and B. Omelayenko, "The semantic web: yet another hip?", *Data & Knowledge Engineering*, 41 (2002), pp. 205-227.

[7] H.-H. Doa and E. Rahmb, "Matching large schemas: Approaches and evaluation", *Information Systems*, (in print), 2007.

[8] J. Euzenat, "Research challenges and perspectives of the Semantic Web". Report of the EU-NSF Strategic Research Workshop, Sophia-Antipolis, France, October, 2001.

[9] A. Farquhar, R. Fikes, and J. Rice, "The Ontolingua Server: a Tool for Collaborative Ontology Construction", Technical report, Stanford KSL 96-26, 1996.

[10] Y. Kalfoglou and M. Schorlemmer, "Ontology mapping: the state of the art", *The Knowledge Engineering Review* 18(1), pp. 1-31, 2003.

[11] K. Kotis and G. A. Vouros, "The HCONE Approach to Ontology Merging", In *ESWS*, LNCS 3053, Springer, pp. 137-151, 2004.

[12] D. L. McGuinness, R. Fikes, J. Rice, and S. Wilder, "An Environment for Merging and Testing Large Ontologies", In Proceedings of the *Seventh International Conference*, 2000.

[13] M.-R. Koivunen and E. Miller, "W3C Semantic Web Activity", in the proceedings of the *Semantic Web Kick/off Seminar*, Finland, 2001.

[14] M. Lenzerini, "Data Integration: A Theoretical Perspective", In Proceedings of the *21st ACM SIGMOD - SIGACT - SIGART symposium on Principles of database systems*, ACM Press, pp. 233-246, 2002.

[15] P. Mitra, G. Wiederhold, and J. Jannink, "Semi-automatic integration of knowledge sources", In Proceeding of the *2nd Int. Conf. On Information FUSION'99*, 1999.

[16] H. Nottelmann and U. Straccia, "Information retrieval and machine learning for probabilistic schema matching", *Inf. Process. Manage.*, 43(3), pp. 552-576, 2007.

[17] N. F. Noy and M. A. Musen, "PROMPT: Algorithm and Tool for Automated Ontology Merging and Alignment", In *AAAI/IAAI*, pp. 450-455, 2000.

[18] R. Pottinger and A. Levy, "A Scalable Algorithm for Answering Queries Using Views", In the Proceedings of the *26th VLDB Conference*, Cairo, Egypt (2000).

[19] U. Priss, "Formal Concept Analysis in Information Science (draft)", `http://www.upriss.org.uk/papers/arist.pdf`.

[20] E. Rahm and P. A. Bernstein, "A survey of approaches to automatic schema matching", In *VLDB Journal: Very Large Data Bases*, 10(4), pp. 334-350, 2001.

[21] P. Shvaiko and J. Euzenat, "A survey of schema-based matching approaches", 3730, pp. 146-171, 2005.

[22] G. Stumme and A. Maedche "FCA-MERGE: Bottom-Up Merging of Ontologies", In *IJCAI*, pp. 225-234, 2001.

[23] X. Su and J. A. Gulla, "An information retrieval approach to ontology mapping", *Data & Knowledge Engineering* 58(1), pp. 47-69, 2006.

[24] H. T. Uitermark, P. J. M. van Oosterom, N. J. I. Mars, and M. Molenaar, "Ontology-based integration of topographic data sets", *International Journal of Applied Earth Observation and Geoinformation* 7 (2005), pp. 97-106.

[25] J. D. Ullman, "Information integration using logical views", *Theoretical Computer Science* 239 (2000), pp. 189-210.

[26] L. Xu and D. W. Embley, "A composite approach to automating direct and indirect schema mappings", *Inf. Syst.*, 31(8), pp. 697-732, 2006.

[27] S. Yi, B. Huang, and W. T. Chan, "Xml application schema matching using similarity measure and relaxation labeling", *Inf. Sci.*, 169(1-2), pp. 27-46, 2005.

[28] Web Ontology Language (OWL), `http://www.w3.org/2004/OWL`.

[29] The Protégé Ontology Editor and Knowledge Acquisition System, `http://protege.stanford.edu/`.

[30] Resource Description Framework (RDF), `http://www.w3.org/RDF/`.

[31] RDF Vocabulary Description Language 1.0: RDF Schema, *W3C Recommendation*, `http://www.w3.org/TR/2004/REC-rdf-schema-20040210`.

[32] WordNet, a lexical database for the English language, `http://wordnet.princeton.edu/`.

[33] Extensible Markup Language (XML), `http://www.w3.org/XML/`.

Computational Study of the Gray-Scott Model

Jan Mach

1st year of PGS, email: machj1@kmlinux.fjfi.cvut.cz
Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU
advisor: Michal Beneš, Department of mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU

Abstract. This contribution deals with numerical solution of the Gray-Scott (GS) model. We introduce two numerical schemes for the 2D GS model based on the method of lines. To perform spatial discretization we use FDM in first case and FEM in the second case. Resulting systems of ODEs are solved using the Runge-Kutta-Merson method. We present some of our numerical simulations.

Abstrakt. V tomto příspěvku se věnujeme numerického řešení Grayova-Scottova (GS) modelu. Představujeme dvě numerická schémata pro 2D GS model založená na metodě přímek. K prostorové diskretizaci používáme v prvním případě FDM, ve druhém FEM. Vzniklé systémy ODEs řešíme metodou Runge-Kutta-Merson. Uvádíme výsledky numerických simulací.

1 Introduction

Reaction-diffusion systems are a class of systems of partial differential equations of parabolic type. It includes mathematical models describing various phenomena in the field of physics, biology and chemistry. Gray-Scott model is one of these models. It was first introduced in 1985 in an article by P. Gray and S. K. Scott. It is a mathematical description of autocatalytic chemical reaction

$$\begin{aligned} U + 2V &\longrightarrow 3V \\ V &\longrightarrow P \end{aligned} \tag{1}$$

and can be written in this form

$$\begin{aligned} \frac{\partial u}{\partial t} &= a\nabla^2 u - uv^2 + F(1-u), \\ \frac{\partial v}{\partial t} &= b\nabla^2 v + uv^2 - (F+k)v. \end{aligned} \tag{2}$$

Here u, v are unknown functions representing concentrations of chemical substances U, V. Parameter F denotes the rate at which the chemical substance U is being added during the chemical reaction, $F + k$ is the rate of $V \to P$ transformation and a, b are constants characterizing the environment where the chemical reaction takes place (see [2, 3, 6]).

2 Problem formulation

Assume that $\Omega \equiv (0, L) \times (0, L)$ is an open square representing the square reactor where the chemical reaction (1) takes place, $\partial\Omega$ is its boundary and ν is its outer normal. Then initial-boundary value problem for the Gray-Scott model we solve is a system (2) of two partial differential equations with initial conditions and zero Neumann boundary conditions

$$\begin{aligned}
\frac{\partial u}{\partial t} &= a\nabla^2 u - uv^2 + F(1-u), \\
\frac{\partial v}{\partial t} &= b\nabla^2 v + uv^2 - (F+k)v \quad in\ \Omega \times (0,T), \\
u(\cdot,0) &= u_{ini}, \\
v(\cdot,0) &= v_{ini}, \\
\frac{\partial u}{\partial \nu}\,|_{\partial\Omega} &= 0, \\
\frac{\partial v}{\partial \nu}\,|_{\partial\Omega} &= 0.
\end{aligned} \tag{3}$$

3 Numerical schemes

We use two numerical schemes to solve initial boundary value problem (3). Both of them are based on the method of lines. For spatial discretization we used finite difference method (FDM) in the first case and finite elements method (FEM) in the second case. We use structured numerical grids (see Fig. 1). To solve resulting systems of ordinary differential equations Runge-Kutta-Merson method is used.

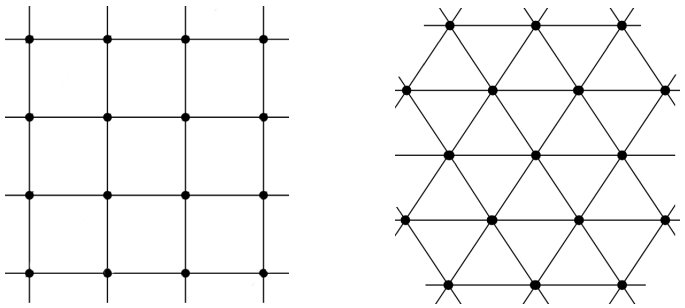


Figure 1: Numerical grids we used for our numerical simulations.

3.1 FDM based numerical scheme

Let h be mesh size such that $h = \frac{L}{N-1}$ for some $N \in \mathbb{N}^+$. We define numerical grid as a set

$$\begin{aligned}
\omega_h &= \{(ih, jh) \mid i = 1, \ldots, N-2, j = 1, \ldots, N-2\}, \\
\overline{\omega}_h &= \{(ih, jh) \mid i = 0, \ldots, N-1, j = 0, \ldots, N-1\}.
\end{aligned}$$

For function $u : \mathbb{R}^2 \to \mathbb{R}$ we define a projection on $\overline{\omega}_h$ as $u_{ij} = u\,(ih, jh)$. We introduce finite differences

$$u_{x_1,ij} = \frac{u_{i+1,j} - u_{i,j}}{h}, u_{\overline{x}_1,ij} = \frac{u_{i,j} - u_{i-1,j}}{h}$$
$$u_{x_2,ij} = \frac{u_{i,j+1} - u_{i,j}}{h}, u_{\overline{x}_2,ij} = \frac{u_{i,j} - u_{i,j-1}}{h},$$

and define approximation Δ_h of the Laplace operator Δ as follows

$$\Delta_h u_{ij} = u_{\overline{x}_1 x_1,ij} + u_{\overline{x}_2 x_2,ij}.$$

Then semi-discrete scheme has the following form

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}u_{ij}(t) &= \frac{a}{h^2}\Delta_h u_{ij} + F(1 - u_{ij}) - u_{ij}v_{ij}^2, \\ \frac{\mathrm{d}}{\mathrm{d}t}v_{ij}(t) &= \frac{b}{h^2}\Delta_h v_{ij} - (F + k)v_{ij} + u_{ij}v_{ij}^2, \end{aligned} \tag{4}$$

plus corresponding initial and boundary conditions.

3.2 FEM based numerical scheme

To induce the semi-discrete scheme we begin with variation formulation of the problem (3). Let

$$\begin{aligned} \varphi_1(x), \varphi_2(x) &\in C_0^\infty(\Omega), \\ \psi_1(t), \psi_2(t) &\in C_0^\infty(0, T) \end{aligned}$$

are test functions and

$$\begin{aligned} f_1(u, v) &= F(1 - u) - uv^2, \\ f_2(u, v) &= -(F + k)v + uv^2 \end{aligned}$$

denote right-hand sides of differential equations (2). Using standard approach (see [1]) we induce weak formulation of the problem

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}(u, \varphi_1) + a(\nabla u, \nabla\varphi_1) &= (f_1, \varphi_1), \\ \frac{\mathrm{d}}{\mathrm{d}t}(v, \varphi_2) + b(\nabla v, \nabla\varphi_2) &= (f_2, \varphi_2), \\ u(\cdot, 0) &= u_{ini}, \\ v(\cdot, 0) &= v_{ini}, \end{aligned} \tag{5}$$

with solution u, v from the Sobolev space $W_2^{(1)}(\Omega)$. We are looking for Galerkin approximation

$$\begin{aligned} u_h(t) &= \sum_{i=1}^{N} \alpha_i(t)\Phi_i, \\ v_h(t) &= \sum_{i=1}^{N} \beta_i(t)\Phi_i \end{aligned}$$

of this weak solution in the finite dimensional space $S_h \subset W_2^{(1)}(\Omega)$, where $\Phi_1, \ldots, \Phi_N$ are its basis functions. Functions α_i, β_i are real functions which we get using common technique as solutions of initial value problems. Choosing basis functions Φ_i in the form of pyramidal functions

$$\Phi_i(P_j) = \delta_{ij} \quad \text{for all grid nodes} \quad P_j,$$

and using mass-lumping we can rewrite the problem for finding functions α_i, β_i in the following form

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}u_{ij}(t) &= \frac{2a}{3h^2}[u_{i+1,j} + u_{i+1,j+1} + u_{i,j-1} + u_{i,j+1} + u_{i-1,j} + \\ &\quad + u_{i-1,j+1} - 6u_{ij}] + F(1 - u_{ij}) - u_{ij}v_{ij}^2 \\ \frac{\mathrm{d}}{\mathrm{d}t}v_{ij}(t) &= \frac{2b}{3h^2}[v_{i+1,j} + v_{i+1,j+1} + v_{i,j-1} + v_{i,j+1} + v_{i-1,j} + \\ &\quad + v_{i-1,j+1} - 6v_{ij}] - (F + k)v_{ij} + u_{ij}v_{ij}^2 \end{aligned} \tag{6}$$

plus corresponding initial and boundary conditions.

For details on induction of presented semi-discrete schemes we refer reader to [5].

4 Numerical experiments

4.1 EOC measurements

To determine the order of convergence of our numerical algorithm based on the FDM based semi-discrete scheme (4) we use experimental order of convergence (EOC). For our measurements we used formula

$$\frac{\| v - v_{h2} \|}{\| v - v_{h1} \|} = \left(\frac{h2}{h1}\right)^{\alpha},$$

where v is numerical solution computed on the grid of size 2000×2000 and substitutes the analytical solution, v_{h2}, v_{h1} are numerical solutions computed on courser grids with mesh sizes $h2$, $h1$ and α is the EOC coefficient. We present some of our measurements for different GS model parameter values and initial conditions (see Tab. 1, Tab. 2, Tab. 3). According to the presented results, the question about the EOC do not have easy answer. Our results vary between the values of 1 and 2. More research into this problem is needed including EOC measurement for the FEM based numerical algorithm.

4.2 Comparison of FDM and FEM

We performed a series of computations to compare our numerical schemes. According to our results the GS model is very sensitive on the numerical scheme used for numerical simulation. FEM based scheme (6) provides results less dependent on the numerical grid size (see Fig. 2). Here, spatial distributions of chemical V concentration over the domain Ω are visualized. Lighter color means higher concentration. For concentrations u,v of chemicals U,V a relation $u = 1 - v$ applies in each point of the domain Ω.

$N_x \times N_y$	h	EOC L_2	EOC L_∞
100x100	0.0050505	-	-
150x150	0.0033557	1.6479179	1.6364127
200x200	0.0025125	1.8042298	1.5663398
250x250	0.0020080	1.9112146	1.7531840
300x300	0.0016722	1.9725610	1.8660718
350x350	0.0014326	2.0089377	1.8995297
400x400	0.0012531	2.0336490	1.9882238

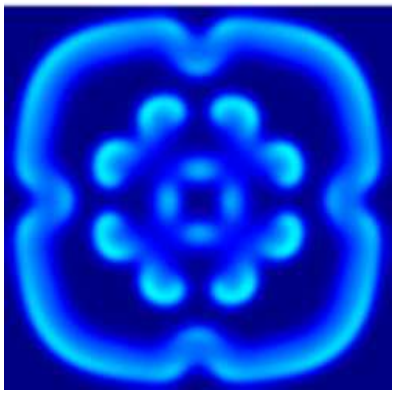
Table 1: Table of EOC coefficients.

$N_x \times N_y$	h	EOC L_2	EOC L_∞
100x100	0.0101010	-	-
150x150	0.0067114	0.8225371	0.5550153
200x200	0.0050251	0.9222231	0.7584173
250x250	0.0040160	0.9995422	0.9052681
300x300	0.0033444	1.0667171	1.0124643
350x350	0.0028653	1.1237827	1.0727512
400x400	0.0025062	1.1754085	1.1689477

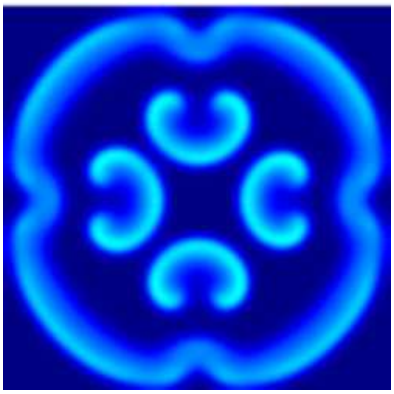
Table 2: Table of EOC coefficients.

$N_x \times N_y$	h	EOC L_2	EOC L_∞
100x100	0.0050505	-	-
150x150	0.0033557	2.0466270	1.0203486
200x200	0.0025125	2.0460521	0.9659226
250x250	0.0020080	2.0512043	1.1006299
300x300	0.0016722	1.9143909	0.9491632
350x350	0.0014326	1.5423185	1.0946135
400x400	0.0012531	1.5552072	0.9893100

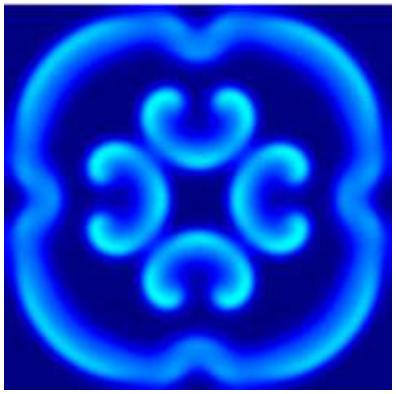
Table 3: Table of EOC coefficients.



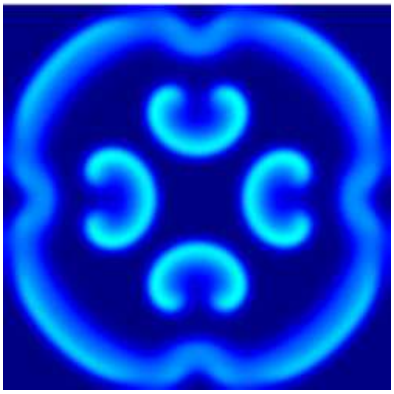
FDM,100×100,$t = 1000$



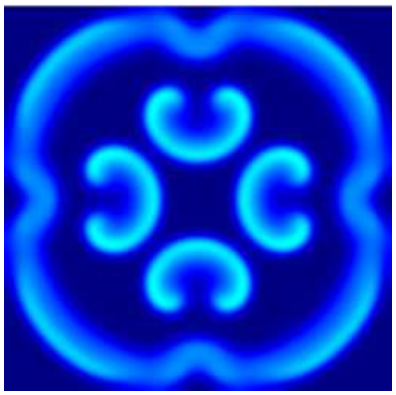
FEM,100×116,$t = 1000$



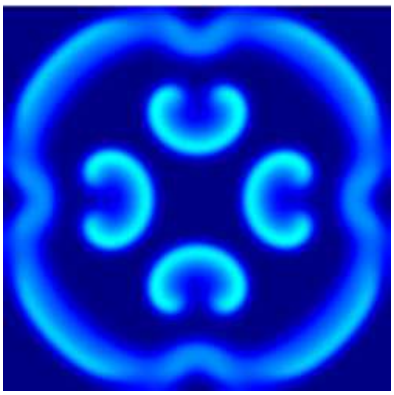
FDM,150×150,$t = 1000$



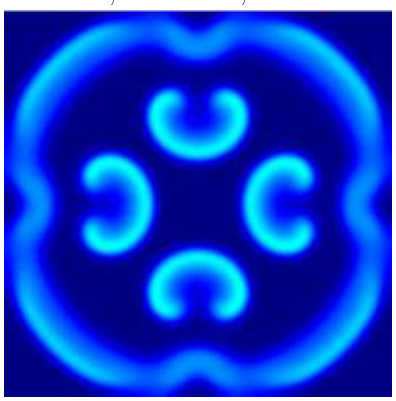
FEM,150×174,$t = 1000$



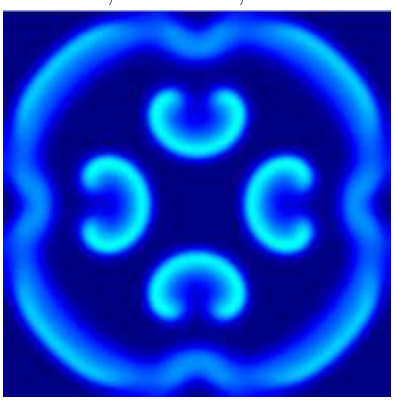
FDM,200×200,$t = 1000$



FEM,200×230,$t = 1000$



FDM,400×400,$t = 1000$



FEM,400×462,$t = 1000$

Figure 2: Dependence of numerical solution on numerical scheme and grid size. GS model parameter values: $a = 2e - 5$, $b = 1e - 5$, $F = 0.024$, $k = 0.054$, $L = 1.0$.

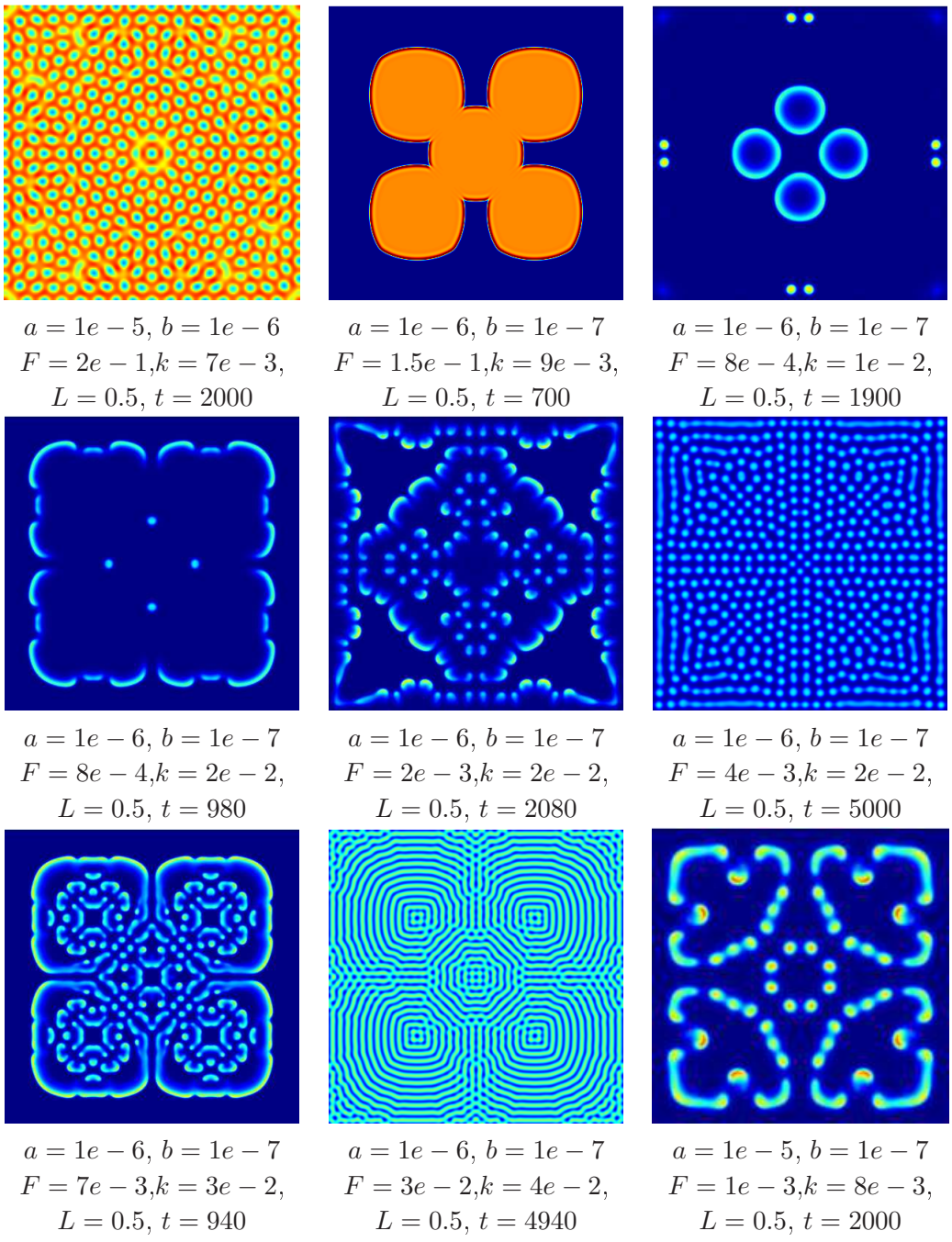


Figure 3: Results demonstrating diversity of solutions of the GS model computed using FDM based numerical scheme (4) and grid size 400×400 for different parameter value combinations.

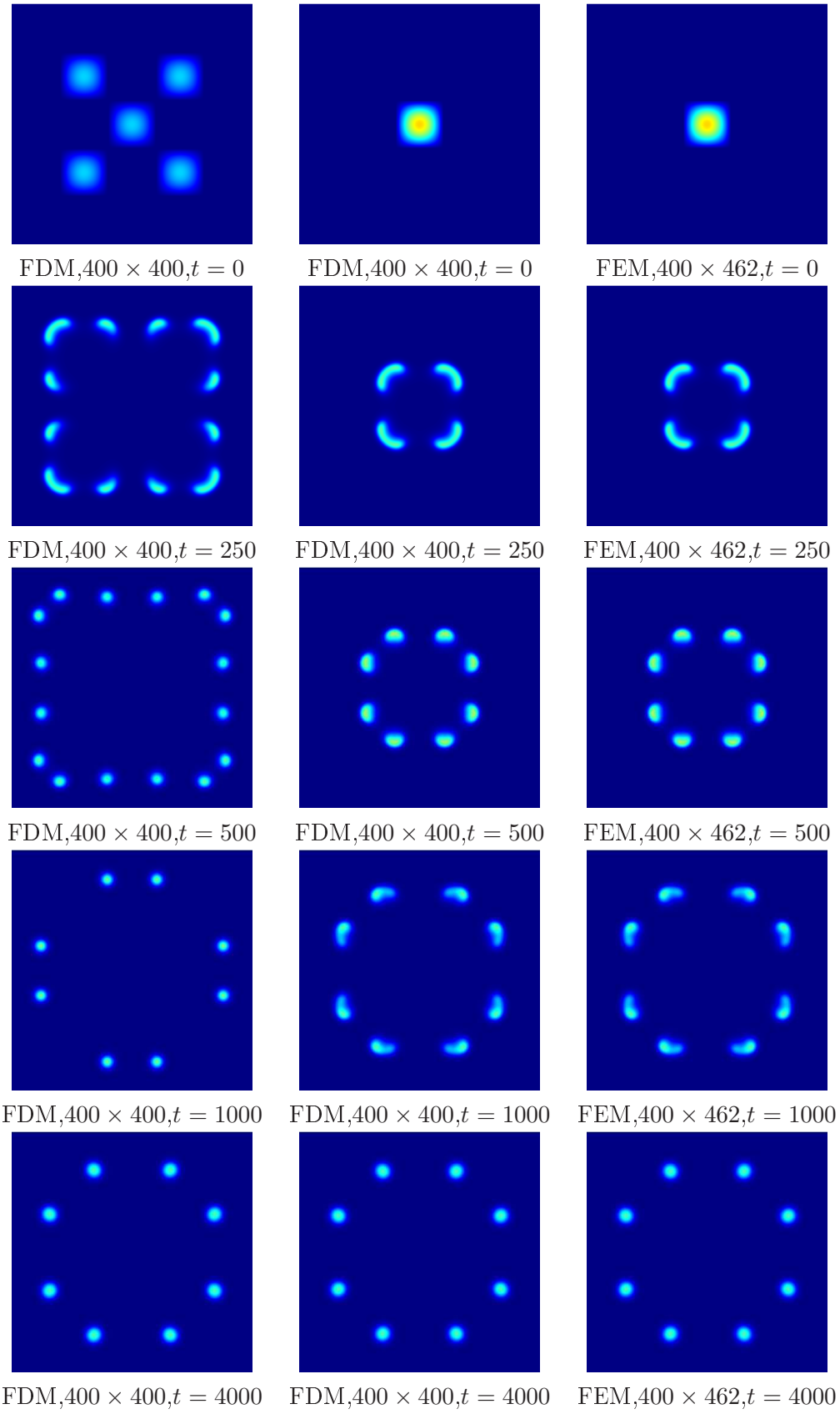


Figure 4: For some parameter value combinations solutions are becoming more and more similar even when using different initial conditions. GS model parameter values: $a = 1e-5$, $b = 1e-6$, $F = 1e-3$, $k = 4e-2$, $L = 0.5$. FEM used to verify results (3^{rd} column).

4.3 Diversity of solutions

On the Fig. 3 and the Fig. 4 we present some of our numerical results. The meaning of images is the same as in the previous text. These results demonstrate the diversity of GS model solutions and some interesting phenomena we found. We can see that patterns are changing from geometrically simple ones to those which are much more complex.

Acknowledgement

The work has been performed under the Project HPC-EUROPA (RII3-CT-2003-506079), with the support of the European Community - Research Infrastructure Action under the FP6 "Structuring the European Research Area" Programme.

Partial support of the project of "Necas Center for Mathematical modeling", No. LC06052 and of the project "Applied Mathematics in Physics and Technical Sciences", No. MSN6840770010 of the Ministry of Education, Youth and Sports of the Czech Republic is acknowledged.

References

[1] V. Thomée. *Galerkin Finite Element Methods for Parabolic Problems*. Springer-Verlag Berlin Heidelberg, 1997.

[2] P. Gray and S. K. Scott. *Chemical Oscillations and Instabilities*. Oxford University Press, Oxford, 1990.

[3] J. Wei. *Pattern formation in two-dimensional Gray-Scott model: existence of single-spot solutions and their stability*. Physica D 148: 20-48 (2001).

[4] J. Wei and M. Winter. *Asymmetric spotty patterns for the Gray-Scott model in* $\mathbb{R}^2$. Stud. Appl. Math. 110 (2003), no. 1, 63-102.

[5] J. Kodovský. *Dynamics of reaction-diffusion equations, mathematical and numerical analysis*. Master thesis, FNSPE CTU, Prague, 2006.

[6] P. Gray and S. K. Scott. *Autocatalytic reactions in the isothermal, continuous stirred tank reactor: oscillations and instabilities int the system* $A + 2B \rightarrow 3B$, $B \rightarrow C$. Chem. Eng. Sci. 39:1087-1097 (1984).

Detecting Traces of Affine Transformation Based on Periodic Properties Of Resampled Images

Babak Mahdian

4th year of PGS, email: mahdian@utia.cas.cz
Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU
advisor: Stanislav Saic, Institute of Information Theory and Automation, AS CR

Abstract. In this work we analyze and analytically describe the specific statistical changes brought into the covariance structure of signal by the interpolation process. We show that interpolated signals and their derivatives contain specific detectable periodic properties. Based on this, we propose a blind, efficient and automatic method capable to find traces of resampling and interpolation. The proposed method can be very useful in many areas, especially in image security and authentication. For instance, when two or more images are spliced together, to create high quality and consistent image forgeries, almost always geometric transformations such as scaling, rotation or skewing are needed. These procedures, typically, are based on a resampling and interpolation step. By having a capable method of detecting the traces of resampling, we can significantly reduce the successful usage of such forgeries.

Abstrakt. Tento příspěvek se zabývá statistickými změnami přinášenými do signálu interpolačním procesem. Analyticky ukážeme, že interpolované signály, obsahují specifické detekovatelné periodické vlastnosti. Dále představíme efektivní slepou metodu, která dokáže detekovat v digitálním signálu a jeho derivacích stopy po převzorkování a interpolace. Tato metoda muže být velmi užitečná v oblasti ověření pravosti digitálních fotografií. Když jsou ve fotomontáži dva či více snímku nakombinované navzájem, k vytvoření jednoho kvalitního padělku, jsou skoro vždy potřebné geometrické transformace jako je změna rozměru či rotace. Tyto operace jsou obvykle založeny na převzorkování a interpolaci. Proto, nabízená metoda muže být efektivní ve snižování úspešného zneužívání tohoto typu padělku.

1 Introduction

Despite of importance, massive usage[1] and history[2] of interpolation, to our knowledge, there exist only a few published works concerned with the specific and detectable statistical changes brought into the signal by this process. In this paper we analytically describe specific periodic properties presence in the covariance structure of interpolated signals and their nth derivatives. Without the detailed knowledge of how the statistics of the signal is changed by the interpolation process, applications based on statistical ap-

[1]For instance, almost every image resizing or rotation operation requires an interpolation process (nearest neighbor, linear, cubic, etc.).

[2]Interpolation has a long history and probably started to being used as early as 2000BC by ancient Babylonian mathematicians. For instance, it had an important role in astronomy which in those days was all about time keeping and making predictions concerning astronomical events [1].

proaches working with resampled/interpolated signals or with their derivatives can yield miscalculations and unexpected results.

Furthermore, we propose a blind, efficient and automatic method capable to detect the traces of resampling and interpolation. The method is based on a derivative operator and radon transformation. The knowledge whether the given signal or some of its portions have been resampled can play an essential role in many fields, especially in image security and authentication.

When two or more images are spliced together (for an example, see Figure 1), to create high quality and consistent image forgeries, almost always geometric transformations such as scaling, rotation or skewing are needed. Geometric transformations typically require a resampling and interpolation step. Therefore, by having sophisticated resampling/interpolation detectors, altered images containing resampled portions can be easily identified and their successful usage significantly reduced.

Existing digital forgery detection methods are divided into active [2, 3], and passive (blind) [6, 7, 4, 5, 8] approaches. The passive (blind) approach is regarded as the new direction. In contrast to active approaches, passive approaches do not need any explicit priori information about the image. They work in the absence of any digital watermark or signature. Passive approaches have not yet been thoroughly researched by many. Different methods for identifying each type of forgery must be developed. Then, by fusing the results from each analysis, a decisive conclusion may be drawn.

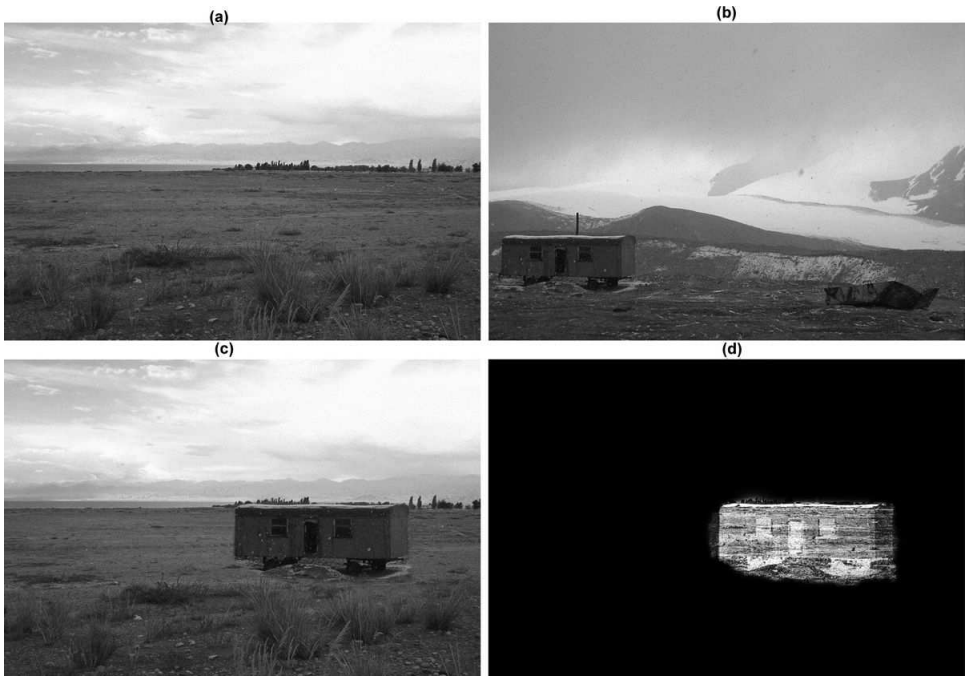


Figure 1: An example of image forgery based on resampling and interpolation. Shown are: source image (a), source image (b), tampered image (c). In (d) is shown the adjusted difference between image (a) and the tampered image (c). The tampered image has been created by splicing source image (a) with a resized part of source image (b). This part has been resized by scaling factor 1.30 using the bicubic interpolation.

In this work, we study and analytically describe the periodic properties of the covariance structure of interpolated signals and their derivatives. Using the theory we bring the main contribution of this paper which is a fast, blind and efficient method capable to detect traces of arbitrary affine transformation. The method can be used for estimating the scaling factors or rotation angles as well as skewing factors. The core of our method is a radon transformation applied to the derivative of the investigated signal. We briefly extend the theory for two-dimensional cases as well. Also we analyze and show periodic patterns of interpolation by an application of Taylor series to the interpolated signals.

2 Basic Notations and Preliminaries

First, a proper mathematical model simulating the acquisition system is required. Periodic properties of interpolation can be effectively studied by using the following simple, linear and stochastic model and assumptions:

$$f(x) = (u * h)(x) + n(x) \tag{1}$$

where f, u, h, $*$, and n are the measured image, original image, system PSF, convolution operator, and random variable representing the influence of noise sources statistically independent from the signal part of the image. For simplicity without loss of generality we assume that $E\{(u*h)(x)\} = 0$ and $E\{n(x)\} = 0$. The covariance of (1) can be shown to be $R_f(x_1, x_2) = E\{(f(x_1) - \overline{f}(x_1))(f(x_2) - \overline{f}(x_2))\} = Cov\{n(x_1), n(x_2)\} = R_n(x_1, x_2)$, where R_f is the covariance matrix of measured image $f(x)$, and R_n is the covariance of random process $n(x)$.

We will denote by f_k a discrete signal representing the samples of $f(x)$ at the locations $k\Delta_x$, $f_k = f(k\Delta_x)$, where $\Delta_x \in \mathcal{R}^+$, is the sampling step and $k \in \mathcal{N}_0$. Furthermore, we assume that the sampling process satisfies the Nyquist criteria. Note that inherent to mentioned assumptions is that $f(x)$ is bandlimited and all derivatives exist at all points. We assume $f(x)$ is bandlimited to $\pm\frac{1}{2\Delta_x}$.

For the sake of simplicity we introduce the operator $\mathcal{D}^n\{\bullet\}$, $n \in \mathcal{N}_0$, which is defined in the following way: $\mathcal{D}^n\{f\}(x) = f(x)$ for $n = 0$ and $\mathcal{D}^n\{f\}(x) = \frac{\partial^n f(x)}{\partial x^n}$ for $n \in \mathcal{N}$. In other words, $\mathcal{D}^0\{f\}(x)$ is identical to $f(x)$ and $\mathcal{D}^n\{f\}(x)$, where $n > 0$, is the nth derivative of $f(x)$. In discrete signals derivative is typically approximated by computing the finite difference between adjacent samples.

3 Periodic Properties of Interpolation

There are two basic steps in geometric transformations. In the first step a spatial transformation of the physical rearrangement of pixels in the image is done. The second step is called the interpolation step. Here pixels intensity values of the transformed image are assigned using a constructed low-pass interpolation filter, w. To compute signal values at arbitrary locations, as the word interpolation signifies[3] discrete samples of f_k are multiplied with the proper filter weights when convolving them with w.

[3]The word "interpolation" originates from the Latin word "inter", meaning "between", and verb "polare", meaning "to polish" [1].

Following the sampling theory, if the Nyquist criterion is satisfied, the spectrum $F(\omega)$ do not overlap in the Fourier domain. The original signal $f(x)$ can be reconstructed perfectly from its samples f_k using the optimal *sinc* interpolator. The sinc function is hard to implement in practice because of its infinite extent. Thus, many different simpler interpolation kernels of bounded support have been investigated and proposed so far [10, 11]. We will be concerned mainly with following low-order piecewise local polynomials: nearest-neighbor, linear, cubic and truncated sinc. These polynomials are used extensively because of their simplicity and implementation unassuming properties.

Combining the derivative theorem with the convolution theorem leads to the conclusion that by convolution of f_k with a derivative kernel $\mathcal{D}^n\{w\}$, it is possible to reconstruct the nth derivative of $f(x)$. We denote the result of interpolation operation by $f^w(x)$, respectively by $\mathcal{D}\{f^w\}(x)$. Formally,

$$\mathcal{D}^n\{f^w\}(x) = \sum_{k=-\infty}^{\infty} f_k \mathcal{D}^n\{w\}(\frac{x}{\Delta_x} - k) \tag{2}$$

As pointed out in [12], it is easy to show that the covariance function of an interpolated image or its derivative is given by:

$$R_{\mathcal{D}^n\{f^w\}}(x, x+\xi) = \sum_{k_1=-\infty}^{\infty} \sum_{k_2=-\infty}^{\infty} \mathcal{D}^n\{w\}(\frac{x}{\Delta_x} - k_1)\mathcal{D}^n\{w\}(\frac{x+\xi}{\Delta_x} - k_2)R_f(k_1, k_2)$$

If we assume band-limited white noise then the variance of $\mathcal{D}^n\{f^w\}$, $var\{\mathcal{D}^n\{f^w\}(x)\}$, as a function of the position x can be represented in the following way:

$$var\{\mathcal{D}^n\{f^w\}(x)\} = R_{\mathcal{D}^n\{f^w\}}(x, x) = \sigma^2 \sum_{k=-\infty}^{\infty} \mathcal{D}^n\{w\}(\frac{x}{\Delta_x} - k)^2 \tag{3}$$

where $\sigma^2 = R_n(k_1, k_2)$. Similarly, the covariance can be represented like:

$$R_{\mathcal{D}^n\{f^w\}}(x, x+\xi) = \sigma^2 \sum_{k=-\infty}^{\infty} \mathcal{D}^n\{w\}(\frac{x}{\Delta_x} - k)\mathcal{D}^n\{w\}(\frac{x+\xi}{\Delta_x} - k)$$

Now, we can notice that

$$var\{\mathcal{D}^n\{f^w\}(x)\} = var\{\mathcal{D}^n\{f^w\}(x + \vartheta\Delta_x)\}, \vartheta \in \mathcal{Z} \tag{4}$$

Thus, $var\{\mathcal{D}^n\{f^w\}(x)\}$ is periodic over x with period Δ_x (as aforementioned, Δ_x is the sampling step). We verify this in the following way:

$$var\{\mathcal{D}^n\{f^w\}(x + \vartheta\Delta_x)\} = \sigma^2 \sum_{k=-\infty}^{\infty} \mathcal{D}^n\{w\}(\frac{x + \vartheta\Delta_x}{\Delta_x} - k)^2$$

$$= \sigma^2 \sum_{k=-\infty}^{\infty} \mathcal{D}^n\{w\}(\frac{x}{\Delta_x} - (k - \vartheta))^2 = var\{\mathcal{D}^n\{f^w\}(x)\}$$

In other words we have shown that interpolation brings into the signal and their derivatives a specific periodicity. This periodicity is dependant on the interpolation kernel used. Several widely used interpolation kernels will be studied in the next section.

Similarly, it can be shown that the covariance of f^w, $R_{\mathcal{D}^n\{f^w\}}(x, x+\xi)$, is periodic as well. The periodicity is apparent for offset $\xi = \vartheta\Delta_x, \vartheta \in \mathcal{Z}$.

$$R_{\mathcal{D}^n\{f^w\}}(x, x+\xi) = R_{\mathcal{D}^n\{f^w\}}(x, x+\vartheta\Delta_x)$$

Before going on, it can be interesting to have a look on application of Taylor series on $\mathcal{D}^n\{f^w\}(x)$. By assuming that the first $(m+1)$ derivatives of $f(x)$ exist, we can rewrite Equation (2) as following:

$$\mathcal{D}^n\{f^w\}(x) = \sum_{k=-\infty}^{\infty} \Big\{ \sum_{m=0}^{m} \frac{\mathcal{D}^m\{f\}(x)}{m!}(k\Delta_x - x)^m + R_{m+1}(x, k\Delta_x) \Big\} \mathcal{D}^n\{w\}(\frac{x}{\Delta_x} - k) \quad (5)$$

By defining

$$\tilde{T}_m(x) = \sum_{k=-\infty}^{\infty} \frac{(k\Delta_x - x)^m}{m!} \mathcal{D}^n\{w\}(\frac{x}{\Delta_x} - k)$$

$$\tilde{R}_{m+1}(x, k\Delta_x) = \sum_{k=-\infty}^{\infty} R_{m+1}(x, k\Delta_x) \mathcal{D}^n\{w(\frac{x}{\Delta_x} - k)\}$$

we can rewrite (5) as:

$$\mathcal{D}^n\{f^w(x)\} = \sum_{m=0}^{m} \tilde{T}_m(x) \mathcal{D}^n\{f\}(x) + \tilde{R}_{m+1}(x, k\Delta_x)$$

Now, by analyzing $\tilde{T}_m(x)$ we can notice that it is periodic with period Δ_x as well:

$$\begin{aligned} \tilde{T}_m(x + \vartheta\Delta_x) &= \sum_{k=-\infty}^{\infty} \frac{(k\Delta_x - (x + \vartheta\Delta_x))^m}{m!} \cdot \mathcal{D}^n\{w\}(\frac{x + \vartheta\Delta_x}{\Delta_x} - k) \\ &= \sum_{k=-\infty}^{\infty} \frac{(\Delta_x(k - \vartheta) - x)^m}{m!} \cdot \mathcal{D}^n\{w\}(\frac{x}{\Delta_x} - (k - \vartheta)) = \tilde{T}_m(x) \end{aligned}$$

3.1 Multidimensional Extension

The theory studied in this section can be analogously extended for the multidimensional cases. If we assume that f_s is a constant variance two-dimensional signal with variance one and $\vartheta \in \mathcal{Z}$, then the Equations (3) and (4) becomes:

$$var\{\mathcal{D}^n\{f^w\}(x, y)\} = \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} \mathcal{D}^n\{w\}(\frac{x}{\Delta_x} - k, \frac{y}{\Delta_y} - l)^2 \quad (6)$$

$$var\{\mathcal{D}^n\{f^w\}(x, y)\} = var\{\mathcal{D}^n\{f^w\}(x + \vartheta\Delta_x, y + \vartheta\Delta_y)\} \quad (7)$$

3.2 Interpolation Kernels

As it is apparent from Equation (3) different interpolators, see Figure 2, change the statistical structure of the signal in different ways. The nearest neighbor interpolator is a zero-degree kernel and the simplest of all piecewise, local polynomials. Its variance function is a constant function. Note that derivatives of the nearest neighbor polynomial are zero. Therefore, signals interpolated by this interpolator can be easily recognized by applying a derivative operator to them.

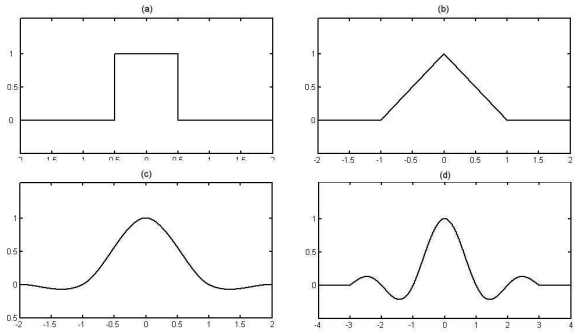


Figure 2: Several popular interpolation kernels : (a) nearest-neighbor, (b) linear, (c) Catmull-Rom cubic, (d) truncated sinc (N=6).

Figure 3 shows periodic variance functions generated via Equation (3) with $\sigma = 1$ for linear interpolation and linear first and second derivative filter. The linear interpolation is a first-degree member of piecewise, local polynomials. It results in an interpolated signal which is continuous, but its first derivative is discontinuous. In Figure 4 the generated periodic variance functions for Catmull-Rom cubic interpolation and cubic first and second order derivative interpolation filter ($\sigma = 1$) are illustrated. Cubic interpolation is a very frequently used interpolation technique and has been widely studied. It uses a third-order interpolation polynomial as kernel. In Figure 5 the variance functions for truncated sinc (with 6 supporting points) interpolation and derivative interpolation filter ($\sigma = 1$) are shown.

4 Detection of Periodic Properties of Resampled Images

The proposed method is based on a few main steps: ROI selection, signal derivative computation, radon transformation and search for periodicity. Each step is explained separately in the following sections.

4.1 Region of Interest Selection

In general, a typical image, $f(x, y)$, consists of several consistent regions. To investigate if any of these regions has been resampled we select this region by a block of $R \times R$ pixels (we denote this block by $b(x, y)$) and apply the method to this image subset. If we are not able to define any ROI in the given image or there is a need to find all resampled

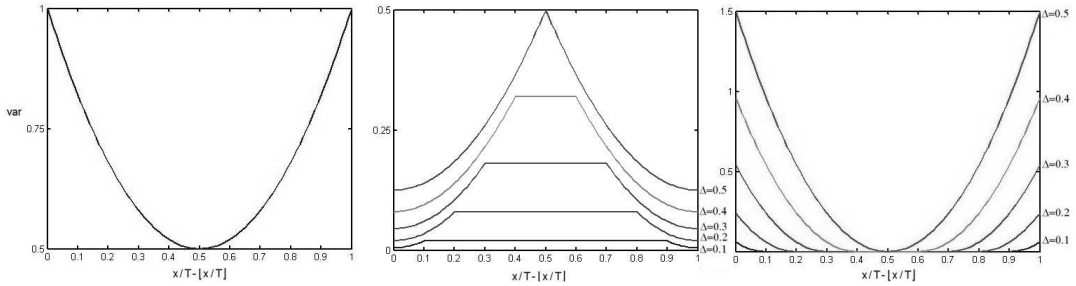


Figure 3: The periodic variance of the linear and linear first and second derivative filter.

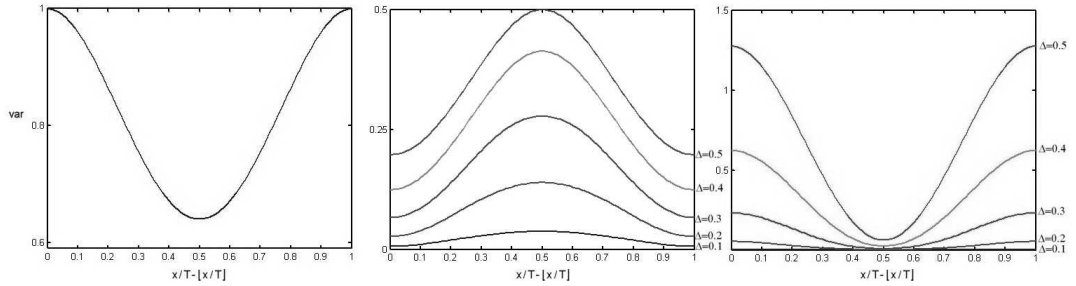


Figure 4: The periodic variance of the cubic and cubic first and second derivative filter.

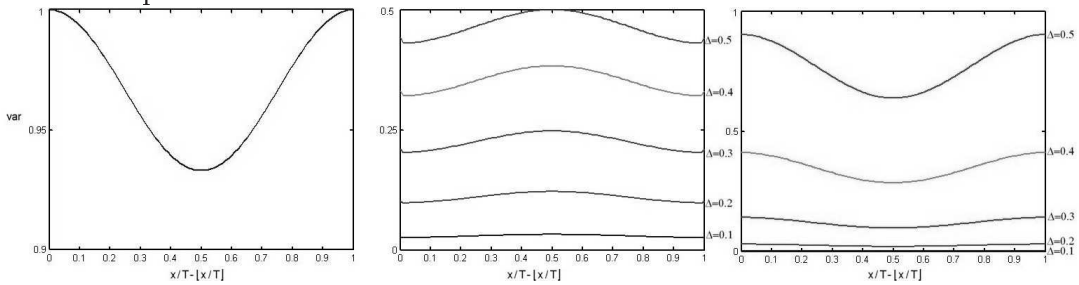


Figure 5: The periodic variance of the truncated sinc and first and second derivative truncated sinc filter ($N = 6$).

regions, the image can be tiled by overlapping blocks, $b_i(x, y)$, of $R \times R$ pixels. Blocks can be horizontally slid by $N, N \in \mathcal{N}$ pixels rightwards starting with the upper left corner and ending with the bottom right corner. Each block can be analyzed by the method separately. In our experiments R is mostly set to 128 pixels.

4.2 Signal Derivative Computation

To emphasize the periodic properties presence in an interpolated image, the nth derivative of $b(x, y)$, $D^n\{b(x, y)\}$, is computed. The derivative operator is applied to the rows of $b(x, y)$. In our experiments the derivative order, n, is set to 2. Similar results can be achieved by other derivative orders or using a laplace operator as well as Gabor filters.

4.3 Radon Transformation

To find traces of affine transform we apply the radon transformation to $|D^n\{b(x,y)\}|$. Radon transformation computes projections of $|D^n\{b(x,y)\}|$ along specified directions determined by angle θ. A projection of $|D^n\{b(x,y)\}|$ is a line integral in a certain direction. By assuming that

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

it is possible to represent the radon transform in the following way:

$$\rho_\theta(x') = \int_{-\infty}^{\infty} D^n\{b(x,y)\} \cdot (x'\cos\theta - y'\sin\theta, x'\sin\theta + y'\cos\theta)dy'.$$

The proposed method computes the radon transformation at angles from 0 to 180°, in 1° increments. Hence, the output of this section is 180 one-dimensional vectors.

4.4 Search for Periodicity

The radon transformation step results in 180 vectors, ρ_θ. If the investigated image has been interpolated, typically some of the auto-covariance sequences of ρ_θ contain a specific strong periodicity. As mentioned previously, our goal is only to determine if the image being investigated has undergone a geometric transformation. Therefore we focus only on the strongest periodic patterns presence in auto-covariance sequences $R_{\rho_\theta}(k)$. This can effect that when the analyzed image has undergone several geometric transformations, our method may not detect all particular transformations presence in this signal, but only those what have the clearest and strongest periodic properties.

To exhibit and detect the searched periodicity, the magnitudes of the Fast Fourier transformation of the autocovariance sequences are computed and all plotted together, $|\mathrm{FFT}(R_{\rho_\theta})|$. This is the main output of the proposed method. In order to easily find strong peaks signifying interpolation, a derivative filter of order one is applied to vectors ρ_θ before computing the $|\mathrm{FFT}(R_{\rho_\theta})|$. If the analyzed signal contains interpolation, peaks in the spectrum are mostly clear and cannot be missed. The spectrum of such a signal has totally different properties compared to non-interpolated signals. To automatically detect interpolation peaks, we apply a simple peak detector searching for the local maximum.

4.5 Experimental Results

Figure 6 shows several outputs of the presented method applied to different TIFF format images that have undergone various transformations. The size of the investigated region in all cases is 128×128 pixels (denoted by a black box). As it is apparent, peaks signifying interpolation are clearly detectable. Note that the way in which we process the outputs of the presented method does not propose a description of all concrete transformation which the investigated image has undergone. For example, in Figure 6(c), only peaks representing the scaling transformation are visible.

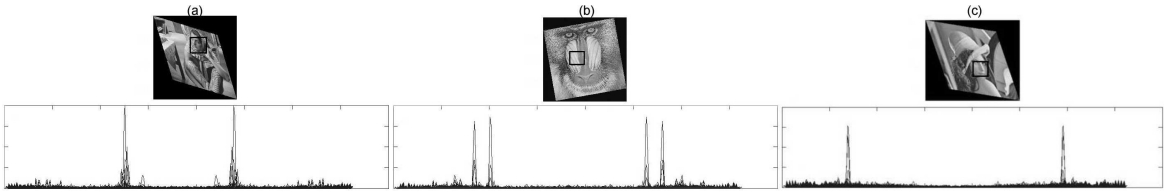


Figure 6: (a) Skewing factor=0.3 (bicubic); (b) scaling factor 1.3; rotation anlge=10° (bicubic); (c) scaling factor 1.2; skewing factor in x-direction=0.2; skewing factor in y-direction=0.4 (bicubic).

5 Discussion

Results obtained show that it is possible in a simple and fast way to find traces of geometric transformation when a low order interpolation polynomial has been used. But, please note that not all resampling factors bring detectable changes in the covariance structure of the signal. For instance, the scaling factor 0.5 does not introduce any periodic correlation into the signal.

The proposed method works well for low order interpolation polynomials: nearest neighbor, linear or cubic. These interpolators have a strong detectable effect on the covariance structure of the signal. The detection performance decreases as the order of interpolation polynomial increases. Different interpolation orders introduce correlations of varying degrees between neighboring samples. These correlations become more difficult to detect as each interpolated sample value is obtained as a function of more samples. Note that when the ideal sinc interpolator is used, the covariance structure of the signal does not change and therefore this interpolator is not detectable. Also, it must be noted that the presented method is highly sensitive to noise.

By applying the proposed method to JPEG compressed images, the detection performance decreases. Experiments show that the presented method works well for JPEG compression quality of 96 - 100. But, generally, obtained results are based on image properties and distribution.

It must be mentioned that obtained results can be affected by spatial correlations presence in the signal. The best results are obtained by applying the method to an interpolated white noise signal (the autocorrelation of a white noise signal have a strong peak at $x = 0$ and is close to 0 elsewhere).

References

[1] E. Meijering, "A chronology of interpolation: From ancient astronomy to modern signal and image processing," *Proceedings of the IEEE*, vol. 90, no. 3, pp. 319–342, March 2002.

[2] M. M. Yeung, "Digital watermarking." *Commun. ACM*, vol. 41, no. 7, pp. 30–33, 1998.

[3] J. Fridrich, "Methods for tamper detection in digital images," *Proceedings of Multimedia and Security Workshop at ACM Multimedia*, pp. 19–23.

[4] J. Fridrich, D. Soukal, and J. Lukas, "Detection of copy–move forgery in digital images," in *Proceedings of Digital Forensic Research Workshop*. Cleveland, OH, USA: IEEE Computer Society, August 2003, pp. 55–61.

[5] A. Popescu and H. Farid, "Exposing digital forgeries by detecting traces of resampling," *IEEE Transactions on Signal Processing*, vol. 53, no. 2, pp. 758–767, 2005.

[6] B. Mahdian and S. Saic, "Detection of copy-move forgery using a method based on invariants," *Forensic Science International*, vol. 171, pp. 180–189, September 2007.

[7] B. Mahdian and S. Saic, "On periodic properties of interpolation and their application to image authentication," *Third International Symposium on Information Assurance and Security. IEEE Computer Society*, pp. 439–446, August 2007.

[8] A. Popescu and H. Farid, "Statistical tools for digital forensics," in *6th International Workshop on Information Hiding*, Toronto, Cananda, 2004, pp. 128–147.

[9] A. C. Gallagher, "Detection of linear and cubic interpolation in jpeg compressed images," in *CRV '05: Proceedings of the The 2nd Canadian Conference on Computer and Robot Vision (CRV'05)*. Washington, DC, USA: IEEE Computer Society, 2005, pp. 65–72.

[10] J. A. Parker, R. V. Kenyon, and D. E. Troxel, "Comparison of interpolation methods for image resampling," *IEEE Transactions on Medical Imaging*, vol. 2, no. 1, pp. 31–39, 1983.

[11] E. H. W. Meijering, W. J. Niessen, and M. A. Viergever, "Piecewise polynomial kernels for image interpolation: A generalization of cubic convolution." in *ICIP (3)*, 1999, pp. 647–651.

[12] G. Rohde, C. Berenstein, and D. Healy, "Measuring image similarity in the presence of noise," *Proceedings of the SPIE Medical Imaging: Image Processing*, vol. 5747, pp. 132–143, February 2005.

Irreversible-thermodynamic Analysis of Proton-exchange-membrane Transport Parameters

Ondřej Mičan

4th year of PGS, email: omican@gmail.com
Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU
advisor: František Maršík, Mathematical Institute, Faculty of Mathematics and Physics, Charles University

Abstract. Irreversible thermodynamics is used to investigate the dependency of maximum possible efficiency of energy conversion in a polymer electrolyte membrane fuel cell (PEMFC) on membrane water content. The importance and plausibility of choice of the value of a proportional parameter in a linear model describing swelling of Nafion 117 membranes is examined. An experiment for simultaneous in-situ measurement of all transport parameters is described.

Abstrakt. Pomocí termodynamiky nerovnovážných stavů bude vyšetřována závislost maximální účinnosti konverze energie v palivovém článku s polymerickou membránou (PEMFC) na obsahu vody v membráně. Zároveň bude anylyzována důležitost správné volby multiplikativního parametru v lineárním modelu pro popis objemových změn membrán z Nafionu 117. Dále bude popsán experiment pro současné stanovení všech transportních parametrů ve funkčním palivovém článku.

1 Introduction

1.1 Model equations and transport parameters

In [17], a simple isothermal, diffusion-type model of a hydrogen polymer-electrolyte membrane (PEM) fuel cell was introduced. The model was based on mass balance equations for H_2O and H_3O^+,

$$\frac{\partial c_{H_2O}}{\partial t} = -4r_a + 6r_c - div\boldsymbol{j}_{H_2O}, \qquad \frac{\partial c_{H_3O^+}}{\partial t} = 4r_a - 4r_c - div\boldsymbol{j}_{H_3O^+}, \tag{1}$$

where c_{H_2O} and $c_{H_3O^+}$ are concentrations of the respective species, $\boldsymbol{j}_{H_2O}$ and $\boldsymbol{j}_{H_3O^+}$ are molar flux densities of the respective species, r_a is the anode reaction rate, and r_c is the cathode reaction rate. The molar flux densities can be expressed as linear combinations of gradients of the species' electrochemical potentials:

$$\boldsymbol{j}_{H_2O} = -L_{ww}\nabla\mu_{H_2O} - L_{we}\nabla\mu_{H_3O^+}, \tag{2}$$

$$\boldsymbol{j}_{H_3O^+} = -L_{ew}\,\nabla\mu_{H_2O} - L_{ee}\,\nabla\mu_{H_3O^+}, \tag{3}$$

where μ_α denotes the electrochemical potential of a species α, and L_{ww}, L_{we}, L_{ew}, and L_{ee} denote phenomenological transport coefficients. It follows from Onsager reciprocity relations that the "cross" coefficients are equal to each other, i. e.

$$L_{ew} = L_{we}, \tag{4}$$

which reduces the number of unknown transport parameters in our model to three. Note that all of them are of the same physical dimensions ($\mathrm{mol^2 J^{-1} m^{-1} s^{-1}}$). If we neglect the effect of concentration changes of H_3O^+, and assume that the reaction mixture behaves as an ideal solution, we can write:

$$\boldsymbol{j}_{\mathrm{H_2O}} = -L_{\mathrm{ww}} \frac{RT}{c_{\mathrm{H_2O}}} \nabla c_{\mathrm{H_2O}} - L_{\mathrm{we}} F \nabla \phi, \tag{5}$$

$$\boldsymbol{j}_{\mathrm{H_3O^+}} = -L_{\mathrm{ew}} \frac{RT}{c_{\mathrm{H_2O}}} \nabla c_{\mathrm{H_2O}} - L_{\mathrm{ee}} F \nabla \phi, \tag{6}$$

where R is the universal gas constant, T is the temperature, F is the Faraday constant and ϕ is the electrostatic potential. From Ohm's law and the obvious fact that current density $\boldsymbol{j} = F\boldsymbol{j}_{\mathrm{H_3O^+}}$, it follows that

$$L_{\mathrm{ee}} = \frac{\sigma}{F^2}, \tag{7}$$

where σ is conductivity. This can easily be seen by putting $\nabla c_{\mathrm{H_2O}} = \mathbf{0}$ in (6). Now, let us eliminate $\nabla\phi$ from (5), (6). Thus we get

$$\begin{aligned} \boldsymbol{j}_{\mathrm{H_2O}} &= -L_{\mathrm{ww}} \frac{RT}{c_{\mathrm{H_2O}}} \nabla c_{\mathrm{H_2O}} - \frac{L_{\mathrm{we}}}{L_{\mathrm{ee}}} \left(-L_{\mathrm{ew}} \frac{RT}{c_{\mathrm{H_2O}}} \nabla c_{\mathrm{H_2O}} - \boldsymbol{j}_{\mathrm{H_3O^+}} \right) = \\ &= -L_{\mathrm{ww}} \frac{RT}{c_{\mathrm{H_2O}}} \nabla c_{\mathrm{H_2O}} + \frac{L_{\mathrm{we}}}{L_{\mathrm{ee}}} \frac{\boldsymbol{j}}{F} + \frac{L_{\mathrm{we}} L_{\mathrm{ew}}}{L_{\mathrm{ee}}} \frac{RT}{c_{\mathrm{H_2O}}} \nabla c_{\mathrm{H_2O}}. \end{aligned} \tag{8}$$

This equation says that water transport through the cell is a superposition of two phenomena, namely diffusion and electro-osmotic drag,

$$\boldsymbol{j}_{\mathrm{H_2O}} = -D_{\mathrm{H_2O}} \nabla c_{\mathrm{H_2O}} + \frac{n^{\mathrm{d}}}{F} \boldsymbol{j}, \tag{9}$$

with diffusion coefficient $D_{\mathrm{H_2O}}$ and electro-osmotic drag coefficient n^{d}. These two coefficients can be related to the phenomenological coefficients by comparing eqs. (8) and (9):

$$D_{\mathrm{H_2O}} = \left(L_{\mathrm{ww}} - \frac{L_{\mathrm{we}} L_{\mathrm{ew}}}{L_{\mathrm{ee}}} \right) \frac{RT}{c_{\mathrm{H_2O}}}, \qquad n^{\mathrm{d}} = \frac{L_{\mathrm{we}}}{L_{\mathrm{ee}}}. \tag{10}$$

Note that very often, the last term in eq. (8) is neglected. Such step yields a more straightforward expression for the diffusion coefficient,

$$D_{\mathrm{H_2O}} = L_{\mathrm{ww}} \frac{RT}{c_{\mathrm{H_2O}}}, \tag{11}$$

and is acceptable since usually $L_{\mathrm{we}} \ll L_{\mathrm{ee}}$.

The electro-osmotic drag coefficient has a physical meaning of the average number of water molecules dragged per H_3O^+ ion moved by electric field through the cell. Note that

$$L_{\mathrm{we}} = L_{\mathrm{ee}} n^{\mathrm{d}} = \frac{\sigma}{F^2} n^{\mathrm{d}}. \tag{12}$$

As was shown in [17], there is a relation between efficiency of energy conversion and the transport parameters or, more precisely, the degree of coupling: The maximum possible efficiency of the conversion of chemical energy into electrical energy in a hydrogen-oxygen fuel cell is

$$\eta_{\max} = \left(\frac{1 - \sqrt{1 - q^2}}{q} \right)^2 , \tag{13}$$

where the degree of coupling between diffusion and migration is defined as

$$q = \frac{L_{\mathrm{we}}}{\sqrt{L_{\mathrm{ww}} L_{\mathrm{ee}}}}. \tag{14}$$

Since the majority of experimentators presents their results in terms of conductivity, diffusion coefficient and electro-osmotic drag coefficient instead of the phenomenological coefficients L_{ww}, L_{we}, L_{ee}, one has to convert their values by using the formulas (11), (12) and (7) in order to calculate q and $\eta_{\max}$. The only difficulty in doing this is to calculate $c_{\mathrm{H_2O}}$.

1.2 Water concentration in the membrane and membrane water content

Most commonly, PEM fuel cell membranes are made of a perfluorosulfonic acid known as Nafion. From now on, we shall consider exclusively membranes made of Nafion 117. Let us recall that, generally, the transport parameters do not depend on driving forces (i. e., $\nabla \mu_{\mathrm{H_2O}}$ and $\nabla \mu_{\mathrm{H_3O^+}}$), but they do depend on state variables (i. e., the *values* of $\mu_{\mathrm{H_2O}}$ and $\mu_{\mathrm{H_3O^+}}$). In practice, however, transport parameters of Nafion fuel cell membranes are considered to be functions of the so-called membrane water content λ, which is the average number of water molecules per sulfonic acid site. In other words,

$$\lambda = \frac{c_{\mathrm{H_2O}}}{c_{\mathrm{SO_3^-}}}. \tag{15}$$

For membranes in the dry state, the concentration of sulfonic acid sites can be calculated as

$$c^{\mathrm{dry}}_{\mathrm{SO_3^-}} = \frac{\rho^{\mathrm{dry}}_{\mathrm{m}}}{M_{\mathrm{m}}}, \tag{16}$$

where $\rho^{\mathrm{dry}}_{\mathrm{m}}$ is density of the membrane in its dry state, and M_{m} is the effective molar mass (or equivalent weight) of the membrane. For Nafion 117, $\rho^{\mathrm{dry}}_{\mathrm{m}} = 1980$ kg/m^3 and $M_{\mathrm{m}} = 1100$ g/mol, thus $c^{\mathrm{dry}}_{\mathrm{SO_3^-}} = 1800$ mol/m^3.

The simplest option is to assume that $c_{\mathrm{SO_3^-}} = c^{\mathrm{dry}}_{\mathrm{SO_3^-}}$ which results in the formula

$$c_{\mathrm{H_2O}} = \lambda c^{\mathrm{dry}}_{\mathrm{SO_3^-}}. \tag{17}$$

This simplified relation was used e. g. in [3].[1] Its problem lies in the assumption that the membrane retains a constant volume regardless of the amount of water it contains

[1] Note that an incorrect value of $c^{\mathrm{dry}}_{\mathrm{SO_3^-}} = 200$ mol/m^3 was reported in [3].

(i. e., $\rho_{\mathrm{m}} = \rho_{\mathrm{m}}^{\mathrm{dry}}$). A real membrane, however, will increase its volume with rising water content. This phenomenon is known as membrane swelling. In [24], the following linear formula was proposed for its description:

$$V = V^{\mathrm{dry}}(1 + s\lambda), \tag{18}$$

where V is volume of the swollen membrane, V^{dry} is volume of the same membrane when it contains no water, and s is a proportional constant. The corresponding relation between $c_{\mathrm{H_2O}}$ and λ can be derived as follows: From (18), (21), and (16) we obtain

$$1 + s\lambda = \frac{V}{V^{\mathrm{dry}}} = \frac{c_{\mathrm{SO_3^-}}^{\mathrm{dry}}}{c_{\mathrm{SO_3^-}}} = \frac{\frac{\rho_{\mathrm{m}}^{\mathrm{dry}}}{M_{\mathrm{m}}}}{\frac{c_{\mathrm{H_2O}}}{\lambda}} = \frac{\rho_{\mathrm{m}}^{\mathrm{dry}}}{M_{\mathrm{m}}} \frac{\lambda}{c_{\mathrm{H_2O}}}, \tag{19}$$

which can easily be rearanged into the following form that was presented e. g. in [16]:

$$c_{\mathrm{H_2O}} = \frac{\rho_{\mathrm{m}}^{\mathrm{dry}}}{M_{\mathrm{m}}} \frac{\lambda}{1 + s\lambda}. \tag{20}$$

Clearly, (17) can be obtained from (20) by putting $s = 0$. Thus, the only question that remains to be solved is to determine the correct value of s. Let us briefly review the existing results on s. From the measured thickness of dry and fully hydrated Nafion 117 membranes, the authors of [24] determined s to have a value of 0.0126. Correctness of this value was called into question in [16]. Experimental results encourage this scepticism. For Nafion 117, the following values were reported [2]: $c_{\mathrm{SO_3^-}} = 1290$ mol/m^3 for $\lambda = 13$, and $c_{\mathrm{SO_3^-}} = 1050$ mol/m^3 for $\lambda = 21$. Since clearly

$$\frac{V}{V^{\mathrm{dry}}} = \frac{c_{\mathrm{SO_3^-}}^{\mathrm{dry}}}{c_{\mathrm{SO_3^-}}}, \tag{21}$$

we can easily calculate the corresponding s from (18), which gives $s = 0.0304$ for $\lambda = 13$, and $s = 0.0340$ for $\lambda = 21$. A similar value of $s = 0.0324$ was derived theoretically in [27].

2 Results and discussion

All in all, we have at least three different values of s to choose from (i. e., 0, 0.0126, and ~ 0.03). In order to examine which one is the most plausible one as well as to what extent would the results obtained for different values of s differ, we decided to calculate the dependency of the degree of coupling q and the maximum efficiency $\eta_{\max}$ for various experimental data sets by using the three values 0, 0.0126, and 0.0324. We tried to gather as much experimental data as possible, but it turned out that it was rather difficult to find suitable experimental data sets. More precisely, we encountered the following two problems considering experimental data on transport parameters:

1. There are not many research groups that measured all three transport parameters.

2. Even those who measured all three transport parameters did not do so at the same conditions and for the same values of λ.

To mitigate the first issue and obtain more variety of data sets, we used besides data of Kreuer et al. (Refs. [11], [6]) and Zawodzinski et al. (Refs. [31], [29]) also those presented by van Bussel et al. (Ref. [3]), although in their work no original data on the electro-osmotic drag coefficient were published and data from [29] were used instead. The second problem was addressed as follows: For those values of λ where only two transport coefficients had been measured, the missing coefficient was calculated by means of inter- and extrapolation. The possible difference between other conditions under which the individual transport coefficients were measured was neglected.

The obtained dependencies of $\eta_{\max}$ on λ are shown in Figs. 1, 2, 3. We can see that the choice of s has a significant impact on the resulting dependency. Since reported efficiencies of real fuel cells are certainly higher than 50 percent, we can conclude that the choice of $s = 0.0324$ gives results that correspond with the physical reality the most. Quite surprising is the dramatical difference among the shapes of the dependencies obtained from the individual data sets. Interestingly, the data of van Bussel et al. (Fig. 3) give the closest results to what one would expect: For low membrane water contents the efficiency is poor, while good membrane hydration results in better fuel cell performance. The data of Zawodzinski et al. (Fig. 2) confirm the fact that the membrane must be well-hydrated in order to obtain reasonable performance, but the corresponding dependency is not monotonously increasing. A cause might lie in the possible inaccuracy of the electro-osmotic drag coefficient values for lower water contents (these values come from another experiment [29]). The most surprising result — a concave dependency — was obtained from the data of Kreuer et al. (Fig. 1). However, their data were rather sparse and extrapolation had to be used to a greater extent in this case, so reliability of this dependency is questionable.

3 An application: In-situ measurement of the electro-osmotic drag coefficient

As indicated above, there is a lack of experimental data sets of all three transport parameters in existing literature. Furthermore, the majority of them comes from measurements of the membrane properties under artificial conditions. Thus, it is desirable to find a method that would be able to measure all three transport parameters simultaneously and in-situ, i. e., in a working fuel cell. We give a description of such experiment based on the irreversible-thermodynamical approach developed above.

First, let us notice that in-situ measurement of conductivity is possible and has been performed [28]. Therefore, we can restrict ourselves on measurement of diffusion and electro-osmotic drag coefficients. Second, a method for in-situ measurement of the so-called effective drag coefficient is also available [8]. The effective (or net) drag coefficient is the ratio of water and proton fluxes through the membrane. It immediately follows from this definition that the effective drag coefficient coincides with the "classical" electro-osmotic drag coefficient n^{d} if and only if the concentration gradient of water in the membrane is zero — cf. (9). Using the relation (9), we are able to determine $-D_{\mathrm{H_2O}}$

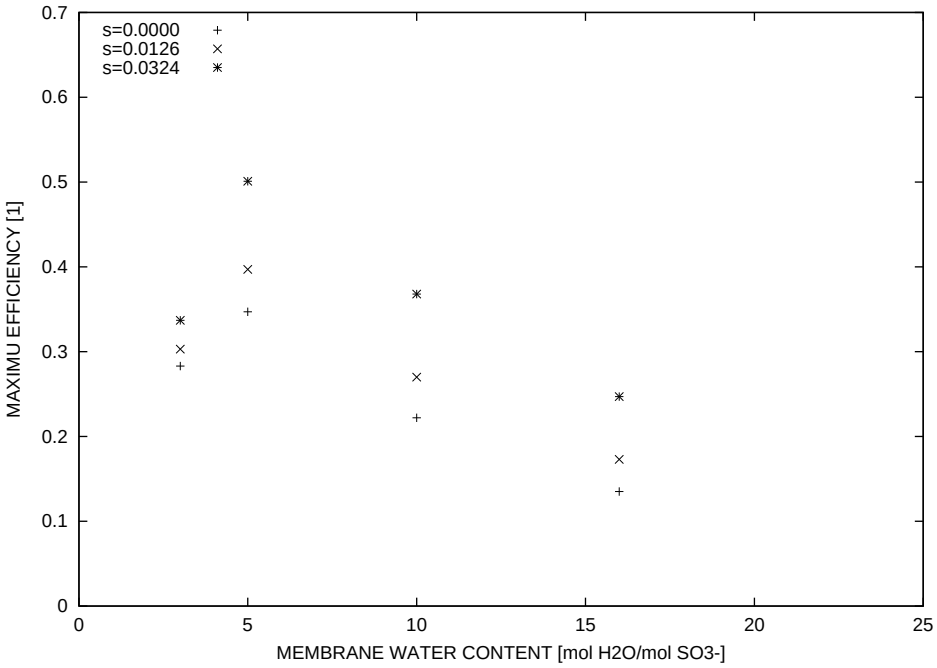


Figure 1: The maximum efficiency $\eta_{\max}$ versus membrane water content λ according to Nafion 117 data of Kreuer et al. [11], [6].

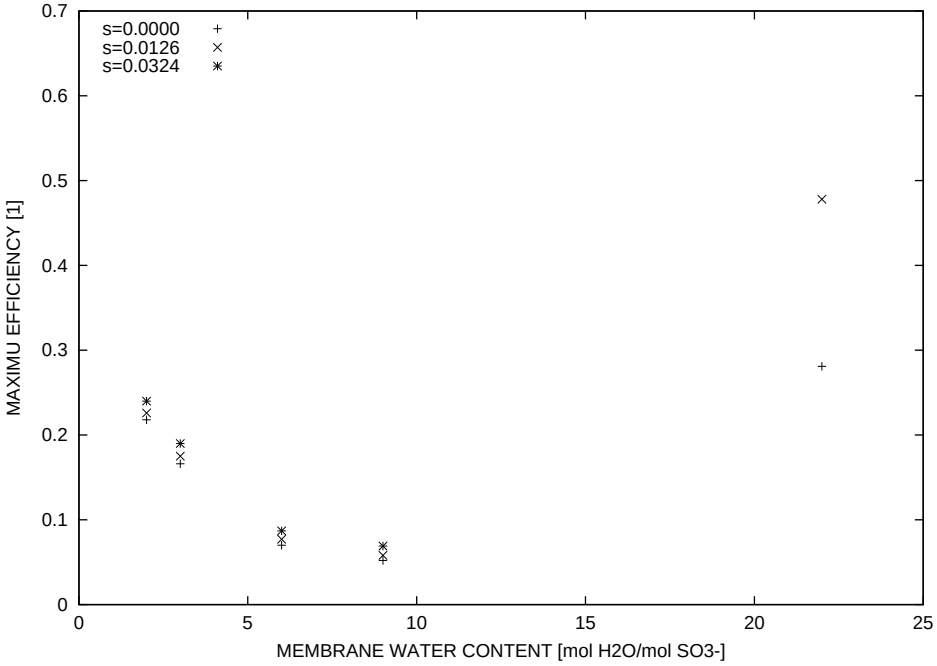


Figure 2: The maximum efficiency $\eta_{\max}$ versus membrane water content λ according to Nafion 117 data of Zawodzinski et al. [31], [29].

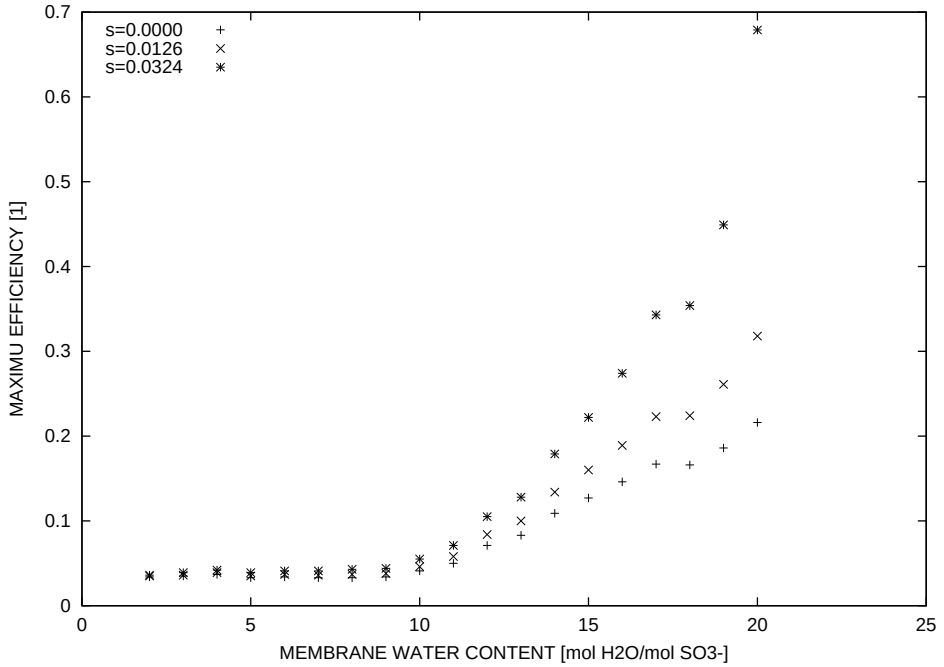


Figure 3: The maximum efficiency $\eta_{\max}$ versus membrane water content λ according to Nafion 117 data of van Bussel et al. [3].

and $\frac{n^{\mathrm{d}}}{F}$ by measuring water flux through the cell in response to known values of water concentration gradient and current density. The only issue to solve is how to measure the water flux $\boldsymbol{j}_{\mathrm{H_2O}}$ and the concentration gradient $\nabla c_{\mathrm{H_2O}}$.

In [8], $\boldsymbol{j}_{\mathrm{H_2O}}$ was determined by means of condensing the water at the exit of the cell gas channels in a cold trap and weighing. For further details, the reader is referred to Ref. [8].

To determine the water concentration gradient, the composition of gas samples obtained from cathode and anode channels of the fuel cell should be analyzed. One could also think of the following simplification: If we assume that all gas components within the fuel cell obey the ideal gas law, we can write

$$\frac{RT}{c_{\mathrm{H_2O}}}\nabla c_{\mathrm{H_2O}} = \frac{RT}{c_{\mathrm{H_2O}}}\nabla\frac{p_{\mathrm{H_2O}}}{RT} = \frac{\nabla p_{\mathrm{H_2O}}}{c_{\mathrm{H_2O}}}, \tag{22}$$

where $p_{\mathrm{H_2O}}$ is the partial pressure of water, which could be approximed with saturation pressure corresponding to cell temperature. The applicability of this approximation should be examined, however, in case that actual partial pressure could not be measured, it might be used as the first attempt.

Once the values of the transport parameters σ, n^{d}, and $D_{\mathrm{H_2O}}$ were determined, we can also determine the corresponding water content λ by means of measuring the cell open circuit voltage. Since the efficiency of a real fuel cell is the ratio of the actual open circuit voltage U^{OC} and the reversible open circuit voltage

$$U^{\mathrm{OC}}_{\mathrm{rev}} = -\frac{\Delta_{\mathrm{r}}G}{2F},$$

$\Delta_r G$ being the Gibbs energy of the electrochemical reaction occuring within the cell [15], we can equate this ratio to the maximum efficiency of energy conversion (13), which yields

$$\frac{U^{\mathrm{OC}}}{U^{\mathrm{OC}}_{\mathrm{rev}}} = \left(\frac{1 - \sqrt{1 - q^2}}{q} \right)^2 . \tag{23}$$

From this formula we can calculate q, and then express L_{ww} from the definition of the degree of coupling (14). Finally, we obtain $c_{\mathrm{H_2O}}$ from (11) and the corresponding λ from (20).

4 Conclusion

Available experimental data on transport coefficients together with irreversible thermodynamics were used to examine plausibility of choice of the value of a proportional parameter in a linear model describing swelling of Nafion 117 membranes. The result confirms previously reported empirical data of Beattie et al. [2] as well as theoretical considerations of Weber and Newman [27]. Furthermore, the dependency of maximum possible efficiency of energy conversion on membrane water content was investigated. Surprisingly, qualitatively different results were obtained for different experimental data sets. Finally, an experiment for simultaneous in-situ measurement of all transport parameters was described. Since data on transport coefficients available in the literature are rather sparse and contradictory, this experiment could help to deepen the corresponding knowledge.

5 Acknowledgments

This work has been supported by IAPWS Young Scientist Fellowship from the International Association for the Properties of Water and Steam. The author would like to express many thanks and gratefulness to his advisors, Prof. F. Maršík at Charles University and Prof. S. Lvov at Pennsylvania State University (USA).

References

[1] P. W. Atkins, *Physical Chemistry (6th Edition)*. Oxford University Press, New York (1998).

[2] P. D. Beattie, F. P. Orfino, V. I. Basura, K. Zychowska, J. Ding, C. Chuy, J. Schmeisser, S. Holdcroft, *Ionic conductivity of proton exchange membranes*. Journal of Electroanalytical Chemistry **503** (2001), 45–56.

[3] H. P. L. H. van Bussel, F. G. H. Koene, R. K. A. M. Mallant, *Dynamic model of solid polymer fuel cell water management*. Journal of Power Sources **71** (1998), 218–222.

[4] K. S. Førland, T. Førland, S. K. Ratkje, *Irreversible Thermodynamics. Theory And Applications*. Wiley – Interscience, London (1988).

[5] P. Glansdorff, I. Prigogine, *Thermodynamic Theory of Structure, Stability and Fluctuations*. Wiley – Interscience, London (1971).

[6] M. Ise, K. D. Kreuer, J. Maier, *Electroosmotic drag in polymer electrolyte membranes: an electrophoretic NMR study*. Solid State Ionics **125** (1999), 213–223.

[7] G. J. M. Janssen, *A Phenomenological Model of Water Transport in a Proton Exchange Membrane Fuel Cell*. J. Electrochem. Soc. **148** (2001), A1313–A1323.

[8] G. J. M. Janssen, M. L. J. Overvelde, *Water transport in the proton-exchange-membrane fuel cell: measurements of the effective drag coefficient*. Journal of Power Sources **101** (2001), 117–125.

[9] D. Kondepudi, I. Prigogine, *Modern Thermodynamics: From Heat Engines to Dissipative Structures*. Wiley, Chichester (1998).

[10] S. Koter, C. H. Hamann, *Characteristics of Ion-Exchange Membranes for Electrodialysis on the Basis of Irreversible Thermodynamics*. J. Non-Equilib. Thermodyn. **15** (1990), 315–333.

[11] K. D. Kreuer, *On the development of proton conducting materials for technological applications*. Solid State Ionics **97** (1997), 1–15.

[12] K. D. Kreuer, *On the development of proton conducting polymer membranes for hydrogen and methanol fuel cells*. Journal of Membrane Science **185** (2001), 29–39.

[13] R. Krishna, J. A. Wesselingh, *The Maxwell-Stefan approach to mass transfer*. Chemical Engineering Science **52** (1997), 861–911.

[14] A. A. Kulikovsky, *Quasi-3D Modeling of Water Transport in Polymer Electrolyte Fuel Cells*. J. Electrochem. Soc. **150** (2003), A1432–A1439.

[15] J. Larminie, A. Dicks, *Fuel Cell Systems Explained (Second Edition)*. Wiley, London (2003).

[16] S. Mazumder, *A Generalized Phenomenological Model and Database for the Transport of Water and Current in Polymer Electrolyte Membranes*. J. Electrochem. Soc. **152** (2005), A1633–A1644.

[17] O. Mičan, *Thermodynamics of Fuel Cell Membrane Transport*. In 'Doktorandské dny 2006 (Proceedings)', Czech Technical University, Prague (2006).

[18] D. R. Morris, X. Sun, *Water-Sorption and Transport Properties of Nafion 117 H*. Journal of Applied Polymer Science **50** (1993), 1445–1452.

[19] T. Okada, N. Nakamura, M. Yuasa, I. Sekine, *Ion and Water Transport Characteristics in Membranes for Polymer Electrolyte Fuel Cells Containing H^+ and Ca^{2+} Cations*. J. Electrochem. Soc. **144** (1997), 2744–2750.

[20] T. Okada, G. Xie, O. Gorseth, S. Kjelstrup, N. Nakamura, T. Arimura, *Ion and water transport characteristics of Nafion membranes as electrolytes.* Electrochimica Acta **43** (1998), 3741–3747.

[21] X. Ren, S. Gottesfeld, *Electro-osmotic Drag of Water in Poly(perfluorosulfonic acid) Membranes.* J. Electrochem. Soc. **148** (2001), A87–A93.

[22] T. Schultz, *Experimental and Model-based Analysis of the Steady-state and Dynamic Operating Behaviour of the Direct Methanol Fuel Cell (DMFC).* Dissertation Thesis, Otto von Guericke University, Magdeburg (2004).

[23] Y. Sone, P. Ekdunge, D. Simonsson, *Proton Conductivity of Nafion 117 as Measured by a Four-Electrode AC Impedance Method.* J. Electrochem. Soc. **143** (1996), 1254–1259.

[24] T. E. Springer, T. A. Zawodzinski, S. Gottesfeld, *Polymer Electrolyte Fuel Cell Model.* J. Electrochem. Soc. **138** (1991), 2334–2342.

[25] J. J. Sumner, S. E. Creager, J. J. Ma, D. D. DesMarteau, *Proton Conductivity in Nafion® 117 and in a Novel Bis[(perfluoroalkyl)sulfonyl]imide Ionomer Membrane.* J. Electrochem. Soc. **145** (1998), 107–110.

[26] A. Z. Weber, J. Newman, *Modeling Transport in Polymer-Electrolyte Fuel Cells.* Chem. Rev. **104** (2004), 4679–4726.

[27] A. Z. Weber, J. Newman, *Transport in Polymer-Electrolyte Membranes II. Mathematical Model.* J. Electrochem. Soc. **151** (2004), A311–A325.

[28] J. A. Weston, *Theoretical And Experimental Investigations of The Transport Processes within Polymer Electrolyte Fuel Cell Membranes.* Diploma Thesis, The Pennsylvania State University, University Park (2002).

[29] T. A. Zawodzinski, J. Davey, J. Valerio, S. Gottesfeld, *The Water Content Dependence of Electro-osmotic Drag in Proton-conducting Polymer Electrolytes.* Electrochimica Acta **40** (1995), 297–302.

[30] T. A. Zawodzinski, C. Derouin, S. Radzinski, R. J. Sherman, V. T. Smith, T. E. Springer, S. Gottesfeld, *Water Uptake by and Transport Through Nafion® 117 Membranes.* J. Electrochem. Soc. **140** (1993), 1041–1047.

[31] T. A. Zawodzinski, T. E. Springer, J. Davey, R. Jestel, C. Lopez, J. Valerio, S. Gottesfeld, *A Comparative Study of Water Uptake By and Transport Through Ionomeric Fuel Cell Membranes.* J. Electrochem. Soc. **140** (1993), 1981–1985.

Algebraic Optimization of Database Queries with Preferences*

Radim Nedbal

4th year of PGS, email: radned@seznam.cz
Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU
advisor: Július Štuller, Institute of Computer Science, AS CR

Abstract. The paper resumes a logical framework for formulating preferences and proposes their embedding into relational algebra through a single *preference operator* parameterized by a set of user preferences of sixteen various kinds, inclusive of *ceteris paribus* preferences, and returning only the most preferred subsets of its argument relation. Most importantly, conflicting set of preferences is permitted and preferences between sets of elements can be expressed.

Formal foundation for algebraic optimization, applying heuristics like *push preference*, also is provided: abstract properties of the preference operator and a variety of algebraic laws describing its interaction with other relational algebra operators are presented.

Abstrakt. Příspěvek shrnuje logické přístupy k vyjadřování preferencí a navrhuje jejich začlenění do relační algebry pomocí jediného *preferenčního operátoru* parametrizovaného množinou až šestnácti různých druhů preferencí, včetně preferencí *ceteris paribus*, a vracejícího nejpreferovanější podmnožiny relace, která je v jeho argumentu. Podstatné je, že koncept zahrnuje preference, které mohou být navzájem v konfliktu a umožňuje reprezentovat i preference mezi množinami.

Navrženy jsou také základní principy algebraické optimalizace jako je např. propagování preferenčního operátoru výrazem relační algebry směrem ke vstupním relacím. Podobné heuristické metody vycházejí z algebraických vztahů operací relační algebry – v tomto případě preferenčního operátoru, které jsou také prezentovány.

1 Introduction

If users have requirements that are to be satisfied completely, their database queries are characterized by *hard constraints*, delivering exactly the required objects if they exist and otherwise empty result. This is how traditional database query languages treat all the requirements on the data. However, requirements can be understood also in the sense of wishes: in case they are not satisfied, database users are usually prepared to accept worse alternatives and their database query is characterized by *soft constraints*. Requirements of the latter type are called preferences.

Building on a logical framework for formulating preferences and their embedding into relational algebra (RA) through a single *preference operator*, introduced in [10] to combat the *empty result* and the *flooding effects*, this paper presents an approach to algebraic

*This has been supported by the project 1ET100300419 of the Program Information Society (of the Thematic Program II of the National Research Program of the Czech Republic) "Intelligent Models, Algorithms, Methods and Tools for the Semantic Web Realization" and by the Institutional Research Plan AV0Z10300504 "Computer Science for the Information Society: Models, Algorithms, Applications".

optimization of relational queries with various kinds of preferences. The preference operator selects from its argument relation the *best-matching alternatives* with regard to user preferences, but *nothing worse*.[1] Preferences are specified using a propositional logic notation and their semantics is related to that of a disjunctive logic program. The language for expressing preferences i) is declarative, ii) includes various kinds of preferences, iii) is rich enough to express preferences between sets of elements, iv) and has an intuitive, well defined semantics allowing for conflicting preferences.

In Sect. 2, the above mentioned framework for formulating preferences and in Sect. 3 an approach to their embedding into RA are revisited. Presenting a variety of algebraic laws that describe interaction with other RA operators to provide a formal foundation for algebraic optimization, Sect. 4 provides the main contribution of this paper. A brief overview of related work in Sect. 5 and conclusions in Sect. 6 end this paper. All the nontrivial proofs are given.

To improve the readability, $\succeq (x,y) \wedge \neg \succeq (y,x)$ and $\succeq (x,y) \wedge \succeq (y,x)$ is substituted by $\succ (x,y)$ and $= (x,y)$, respectively.

2 User Preferences

A user preference is expressed by a preference statement, e.g. "a is preferred to b", or symbolically by an appropriate preference formula. Preference formulas comprise a simple declarative language for expressing preferences. To capture its declarative aspects, model-theoretic semantics is defined: considering a set of states of affaires S and a set $W = 2^S$ of all its subsets – worlds, if $\mathcal{M} = \langle W, \succeq \rangle$ is an order $\succeq$ on W such that $w \succeq w'$ holds for some words w, w' from W, then $\mathcal{M}$ is termed a *preference model* of $w > w'$ – a preference of the world w over the world w', which we express symbolically as $\mathcal{M} \models w > w'$.

The basic differentiation between preferences is based on notions of optimism and pessimism. Defining a-world as a world in which a occurs, if we are optimistic about a and pessimistic about b for example, we expect some a-world to precede at least one b-world in each preference model of a preference statement "a is preferred to b". This kind of preference is called *opportunistic*. By contrast, if we are pessimistic about a and optimistic about b, we expect every a-world to precede each b-world in each preference model of a preference statement "a is preferred to b". This kind of preference is called *careful*. Alternatively, we might be optimistic or pessimistic about both a and b. Then we expect some a world to precede each b-world or each a-world to precede some b-world in each preference model of a preference statement "a is preferred to b". This kind of preference is called *locally optimistic* or *locally pessimistic*, respectively. Locally optimistic, locally pessimistic, opportunistic and careful preferences are symbolically expressed by preference formulas of the form: $a \ {}^{M}{>}^{M} \ b$, $a \ {}^{m}{>}^{m} \ b$, $a \ {}^{M}{>}^{m} \ b$, and $a \ {}^{m}{>}^{M} \ b$, respectively.

Also, we distinguish between strict and non-strict preferences. For example, if w precedes w' strictly in a preference model, then we strictly prefer w to w'.

In addition, we distinguish between preferences with and without *ceteris paribus* proviso – a notion introduced by von Wright [11] and generalized by Doyle and Wellman

[1] A similar concept was proposed independently by Kießling et al. [6, 7] and Chomicki et al. [2] and, in a more restricted form, by Börzsönyi et al. [1] (for more detail refer to Sect. 5).

[3] by means of contextual equivalence relation – an equivalence relation on W.[2] For example, a preference model of a preference statement "a is *carefully* preferred to b *ceteris paribus*" is such an order on W that a-worlds precede b-worlds in the same contextual equivalence class. Specifically, the preference statement "I prefer playing tenis to playing golf ceteris paribus" might express by means of an contextual equivalence that I prefer playing tenis to playing golf only if the context of weather is the same, i.e., it is not true that I prefer playing tenis in strong winds to playing golf during a sunny day.

Next, we revisit the basic definitions introducing syntax and model-theoretic semantics of the language for expressing user preferences:

Definition 1 (Language)**.** Given a finite set of propositional variables $p, q, \ldots$, the set L_0 of *propositional formulas* and the set L of *preference formulas* is defined as the smallest set satisfying the following:

$L_0 \ni \varphi, \psi$: $p \mid (\varphi \wedge \psi) \mid \neg\varphi$
$L \ni \Phi, \Psi$: $\varphi \; {}^{x}{>}^{y} \; \psi \mid \varphi \; {}^{x}{\geq}^{y} \; \psi \mid \neg\Phi \mid (\Phi \wedge \Psi) \quad$ for $x, y \in \{m, M\}$

If we identify propositional variables with tuples over a relation schema R, then the elements of L are termed *preference formulas over* R. A relation instance $I(R)$, i.e., a set of tuples over R, creates a *world* w, an element of a set W.

The preference model is defined so that any set of (possibly conflicting) preferences is consistent: the partial pre-order, i.e., a binary relation which is reflexive and transitive, in the definition of the preference model, enables to express some kind of conflict by incomparability:

Definition 2 (Preference model)**.** A preference model $\mathcal{M} = \langle W, \succeq \rangle$ over a relation schema R is a couple in which W is a set of worlds, relation instances of R, and $\succeq$ is a *partial pre-order* over W, the *preference relation* over R.

A set of user preferences of various kinds can by represented symbolically by a *preference specification*, which corresponds to an appropriate complex preference formula in the above defined language.

Definition 3 (Preference specification)**.** Let R be a relation schema and $\mathcal{P}_\rhd$ a set of preference formulas over R of the form $\{\varphi_i \rhd \psi_i : i = 1, \ldots, n\}$. A preference specification $\mathcal{P}$ over R is a tuple $\langle \mathcal{P}_\rhd | \rhd \in \{ \; {}^{x}{>}^{y}, \; {}^{x}{\geq}^{y} \; | x, y \in \{m, M\}\}\rangle$, and $\mathcal{M}$ is its model, i.e., a *preference specification model*, iff it models all elements $\mathcal{P}_\rhd$ of the tuple:

$$\mathcal{M} \models \mathcal{P}_\rhd \iff \forall(\varphi_i \rhd \psi_i) \in \mathcal{P}_\rhd : \mathcal{M} \models \varphi_i \rhd \psi_i \; .$$

3 Preference Operator

To embed preferences into RQL, the *preference operator* $\omega_\mathcal{P}$ returning only the best sets of tuples in the sense of user preferences $\mathcal{P}$ is defined:

[2] As it has been shown [5] that any preference with contextual equivalence specification can be expressed by a set of preferences without contextual specification, we can restrict ourselves only to preferences without ceteris paribus proviso.

Definition 4 (Preference operator)**.** If R is a relation schema, $\mathcal{P}$ a preference specification over R, and $\mathscr{M}$ the set of its models; then the preference operator $\omega_{\mathcal{P}}$ is defined for all instances $I(R)$ of R as follows:

$$\omega_{\mathcal{P}}(I(R)) = \{w \in W \mid w \subseteq I(R) \wedge \exists \mathcal{M}_k = \langle W, \succeq_k \rangle \in \mathscr{M} \text{ s.t. } \forall w' \in W : \\ w' \subseteq I(R) \wedge \succeq_k (w', w) \Rightarrow \succeq_k (w, w')\} \ .$$

Remark 5 (Preference operator notation)**.** To be precise, we should write $\omega_{\mathcal{P}}(2^{I(R)})$ instead of $\omega_{\mathcal{P}}(I(R))$. Thus it makes sense to write $\omega_{\mathcal{P}}(\{a, b, c\})$, where the argument of preference operator is a set of elements a, b, and c.

3.1 Basic Properties.

The following propositions are essential for investigation of algebraic properties describing interaction of the preference operator with other RA operations:

Proposition 6. *Given a relation schema R and a preference specification $\mathcal{P}$ over R, for all instances $I(R)$ of R the following properties hold:*

$$\begin{aligned} \omega_{\mathcal{P}}(I(R)) &\subseteq 2^{I(R)} \ , \\ \omega_{\mathcal{P}}\left(\omega_{\mathcal{P}}(I(R))\right) &= \omega_{\mathcal{P}}(I(R)) \ , \\ \omega_{\mathcal{P}_{\text{empty}}}(I(R)) &= 2^{I(R)} \ , \end{aligned}$$

where $\mathcal{P}_{\text{empty}}$ is the empty preference specification, i.e., containing no preference.

Preference operator is not *monotone* or *antimonotone* with respect to its relation argument. However, partial antimonotonicity holds:

Proposition 7 (Partial antimonotonicity)**.** *Given a relation schema R and a preference specification $\mathcal{P}$ over R, for all instances $I(R), I'(R)$ of R the following property holds:*

$$I(R) \subseteq I'(R) \Rightarrow 2^{I(R)} \cap \omega_{\mathcal{P}}(I'(R)) \subseteq \omega_{\mathcal{P}}(I(R)) \ .$$

Proof. Assume $w \in 2^{I(R)} \cap \omega_{\mathcal{P}}(I'(R))$. It follows that $w \subseteq I(R)$ and from the definition (Def. 4) of preference operator $w \subseteq I'(R) \wedge \exists \mathcal{M}_k \in \mathscr{M}$ s.t. $\forall w' \in W : w' \subseteq I'(R) \wedge \succeq_k (w', w) \Rightarrow \succeq_k (w, w')$. As $I(R) \subseteq I'(R)$, we can conclude that $\exists \mathcal{M}_k \in \mathscr{M}$ s.t. $\forall w' \in W : w' \subseteq I(R) \wedge \succeq_k (w', w) \Rightarrow \succeq_k (w, w')$, which together with $w \subseteq I(R)$ implies $w \in \omega_{\mathcal{P}}(I(R))$. □

The following theorem enables to reduce cardinality of an argument relation of the preference operator without changing the return value:

Theorem 8 (Reduction)**.** *Given a relation schema R, a preference specification $\mathcal{P}$ over R, for all instances $I(R), I'(R)$ of R the following property holds:*

$$I(R) \subseteq I'(R) \wedge \omega_{\mathcal{P}}(I'(R)) \subseteq 2^{I(R)} \Rightarrow \omega_{\mathcal{P}}(I(R)) = \omega_{\mathcal{P}}(I'(R)) \ .$$

Proof. $\subseteq$: Assume $w \in \omega_{\mathcal{P}}(I(R))$. Then, it follows from the definition of the preference operator $w \subseteq I(R) \wedge \exists \mathcal{M}_k \in \mathscr{M}$ s.t. $\forall w' \in W : w' \subseteq I(R) \wedge \succeq_k (w', w) \Rightarrow \succeq_k (w, w')$. The assumption $\omega_{\mathcal{P}}(I'(R)) \subseteq 2^{I(R)}$ implies $\forall w' \in 2^{I'(R)} - 2^{I(R)} : \neg \succeq_k (w', w)$, and we can conclude $\exists \mathcal{M}_k \in \mathscr{M}$ s.t. $\forall w' \in W : w' \subseteq I'(R) \wedge \succeq_k (w', w) \Rightarrow \succeq_k (w, w')$, which together with the assumption $I(R) \subseteq I'(R)$ implies $w \in \omega_{\mathcal{P}}(I'(R))$.

$\supseteq$: Immediately follows from Prop. 7. □

The following theorem ensures that the empty query result effect is successfully eliminated:

Theorem 9 (Non-emptiness)**.** *Given a relation schema R, a preference specification $\mathcal{P}$ over R, then for every finite, nonempty instance $I(R)$ of R, $\omega_{\mathcal{P}}(I(R))$ is nonempty.*

3.2 Multidimensional Composition.

In multidimensional composition, we have a number of preference specifications defined over several relation schemas, and we define preference specification over the Cartesian product of those relations: the most common ways are Pareto and lexicographic composition.

Definition 10 (Pareto and lexicographic composition)**.** Given two relation schemas R_1 and R_2, preference specifications $\mathcal{P}_1$ over R_1 and $\mathcal{P}_2$ over R_2, and their sets of models $\mathscr{M}_1$ and $\mathscr{M}_2$, the *Pareto composition* $P(\mathcal{P}_1, \mathcal{P}_2)$ and the *lexicographic composition* $L(\mathcal{P}_1, \mathcal{P}_2)$ of $\mathcal{P}_1$ and $\mathcal{P}_2$ is a preference specification $\mathcal{P}_0$ over the Cartesian product $R_1 \times R_2$, whose set of models $\mathscr{M}_0$ is defined as:

$$\forall \mathcal{M}_m = \langle W_1 \times W_2, \succeq_m \rangle \in \mathscr{M}_0, \exists \mathcal{M}_k = \langle W_1, \succeq_k \rangle \in \mathscr{M}_1, \exists \mathcal{M}_l = \langle W_2, \succeq_l \rangle \in \mathscr{M}_2 \text{ s.t.}$$

$$\forall w_1, w_1' \in W_1, \forall w_2, w_2' \in W_2 : \succeq_m (w_1 \times w_2, w_1' \times w_2') \quad \equiv \quad \succeq_k (w_1, w_1') \wedge \succeq_l (w_2, w_2')$$

and

$$\forall w_1, w_1' \in W_1, \forall w_2, w_2' \in W_2 : \succeq_m (w_1 \times w_2, w_1' \times w_2') \quad \equiv \quad \succ_k (w_1, w_1') \vee (=_k (w_1, w_1') \wedge \succeq_l (w_2, w_2')) \ ,$$

respectively.

4 Algebraic Optimization

As the preference operator extends RA, the optimization of queries with preferences can be realized as an extension of a classical relational query optimization. Most importantly, we can inherit all well known laws from RA, which, together with algebraic laws governing the commutativity and distributivity of the preference operator with respect to RA operations, constitute a formal foundation for rewriting queries with preferences using the standard strategies (*push selection*, *push projection*) aiming at reducing the sizes of intermediate relations.

4.1 Commuting with Selection

The following theorem identifies a sufficient condition under which the preference operator commutes with RA selection:

Theorem 11 (Commuting with selection)**.** *Given a relation schema R, a preference specification $\mathcal{P}$ over R, the set of its preference models $\mathscr{M}$, and a selection condition φ over R, if the formula*

$$\forall \mathcal{M}_k = \langle W, \succeq_k \rangle \in \mathscr{M}, \forall w, w' \in W : \quad \succ_k (w', w) \wedge w = \sigma_\varphi(w) \Rightarrow w' = \sigma_\varphi(w')$$

is valid, then for any relation instance $I(R)$ of R:

$$\omega_\mathcal{P}(\sigma_\varphi(I(R))) = \sigma_\varphi(\omega_\mathcal{P}(I(R))) \stackrel{def}{=} \{w \in \omega_\mathcal{P}(I(R)) | \sigma_\varphi(w) = w\} \ .$$

Proof. Observe that:

$$w \in \omega_\mathcal{P}(\sigma_\varphi(I(R))) \equiv w \subseteq I(R) \wedge \sigma_\varphi(w) = w \wedge \\ \neg(\forall \mathcal{M}_k \in \mathscr{M}, \exists w' : (w' \subseteq I(R) \wedge \sigma_\varphi(w') = w' \wedge \succ_k (w', w)) \ .$$

$$w \in \sigma_\varphi(\omega_\mathcal{P}(I(R))) \equiv w \subseteq I(R) \wedge \sigma_\varphi(w) = w \wedge \\ \neg(\forall \mathcal{M}_k \in \mathscr{M}, \exists w' : (w' \subseteq I(R) \wedge \succ_k (w', w)) \ ,$$

Obviously, the second formula implies the first. To see that the opposite implication also holds, we assume $w \notin \sigma_\varphi(\omega_\mathcal{P}(I(R)))$ and prove that than also $w \notin \omega_\mathcal{P}(\sigma_\varphi(I(R)))$. There are three cases when $w \notin \sigma_\varphi(\omega_\mathcal{P}(I(R)))$. If $w \nsubseteq I(R)$ or $\sigma_\varphi(w) \neq w$, it is immediately clear that $w \notin \omega_\mathcal{P}(\sigma_\varphi(I(R)))$. In the third case, $\forall \mathcal{M}_k \in \mathscr{M}, \exists w' : (w' \subseteq I(R) \wedge \succ_k (w', w)$. However, due to assumption of the theorem, $\forall \mathcal{M}_k \in \mathscr{M}, \exists w' : (w' \subseteq I(R) \wedge \sigma_\varphi(w') = w' \wedge \succ_k (w', w)$, which completes the proof. □

4.2 Commuting with Projection

The following theorem identifies sufficient conditions under which the preference operator commutes with RA projection. To prepare the ground for the theorem, some definitions have to be introduced:

Definition 12 (Restriction of a preference relation)**.** Given a relation schema R, a set of attributes X of R, and a preference relation $\succeq$ over R, the restriction $\theta_X(\succeq)$ of $\succeq$ to X is a preference relation $\succeq_X$ over $\pi_X(R)$ defined using the following formula:

$$\succeq_X (w_X, w'_X) \quad \equiv \quad \forall w, w' \in W : \pi_X(w) = w_X \wedge \ \pi_X(w') = w'_X \Rightarrow \ \succeq (w, w') \ .$$

Definition 13 (Restriction of the preference model)**.** Given a relation schema R, a set of relation attributes X of R, and a preference model $\mathcal{M} = \langle W, \succeq \rangle$ over R, the restriction $\theta_X(\mathcal{M})$ of $\mathcal{M}$ to X is a preference model $\mathcal{M}_X = \langle W_X, \succeq_X \rangle$ over $\pi_X(R)$ where $W_X = \{\pi_X(w) \mid w \in W\}$.

Definition 14 (Restriction of the preference operator)**.** Given a relation schema R, a set of attributes X of R, a preference specification $\mathcal{P}$ over R, and the set $\mathscr{M}_X$ of its models restricted to X, the restriction $\theta_X(\omega_\mathcal{P})$ of the preference operator $\omega_\mathcal{P}$ to X is the preference operator $\omega_\mathcal{P}^X$ defined as follows:

$$\omega_\mathcal{P}^X(\pi_X(I(R))) = \{w_X \in W_X \mid w_X \subseteq \pi_X(I(R)) \wedge \exists \mathcal{M}_X \in \mathscr{M}_X \text{ s.t.} \\ \forall w'_X \in W_X : w'_X \subseteq \pi_X(I(R)) \wedge \succeq_X (w'_X, w_X) \Rightarrow \succeq_X (w_X, w'_X)\} \ .$$

Theorem 15 (Commuting with projection)**.** *Given a relation schema R, a set of attributes X of R, a preference specification $\mathcal{P}$ over R, and the set of its preference models $\mathscr{M}$, if the following formulae*

$$\forall \mathcal{M}_k \in \mathscr{M}, \forall w_1, w_2, w_3 \in W : \\ \pi_X(w_1) = \pi_X(w_2) \wedge \pi_X(w_1) \neq \pi_X(w_3) \wedge \succeq_k (w_1, w_3) \Rightarrow \succeq_k (w_2, w_3) \ ,$$

$$\forall \mathcal{M}_k \in \mathscr{M}, \forall w_1, w_3, w_4 \in W : \\ \pi_X(w_3) = \pi_X(w_4) \wedge \pi_X(w_1) \neq \pi_X(w_3) \wedge \succeq_k (w_1, w_3) \Rightarrow \succeq_k (w_1, w_4)$$

are valid, then for any relation instance $I(R)$ of R:

$$\omega_\mathcal{P}^X(\pi_X(I(R))) = \pi_X(\omega_\mathcal{P}(I(R))) \stackrel{def}{=} \{\pi_X(w) \mid w \in \omega_\mathcal{P}(I(R))\} \ .$$

Proof. We prove: $\pi_X(w) \notin \omega_\mathcal{P}^X(\pi_X(I(R))) \Longleftrightarrow \pi_X(w) \notin \pi_X(\omega_\mathcal{P}(I(R)))$.

$\Rightarrow$: Assume $\pi_X(w_3) \notin \omega_\mathcal{P}^X(\pi_X(I(R)))$. The case $\pi_X(w_3) \nsubseteq \pi_X(I(R))$ is trivial. Otherwise, it must be the case that $\forall \mathcal{M}_X \in \mathscr{M}_X, \exists w_X$ s.t. $w_X \subseteq \pi_X(I(R))$ and $\succ_X (w_X, \pi_X(w_3))$, which implies $\forall \mathcal{M}_k \in \mathscr{M}, \forall w_1, w_4 \in W : \pi_X(w_1) = w_X \wedge \pi_X(w_4) = \pi_X(w_3) \Rightarrow \succ_k (w_1, w_4)$ and thus $\pi_X(w_3) \notin \pi_X(\omega_\mathcal{P}(I(R)))$.

$\Leftarrow$: Assume $\pi_X(w_3) \notin \pi_X(\omega_\mathcal{P}(I(R)))$. Then $\forall \mathcal{M}_k \in \mathscr{M}$ and $\forall w_4 \subseteq I(R)$ s.t. $\pi_X(w_4) = \pi_X(w_3)$, there is $w_1 \subseteq I(R)$ s.t. $\succ_k (w_1, w_4)$ and $\pi_X(w_1) \neq \pi_X(w_4)$. From the assumption of the theorem, it follows that $\forall w_2, w_4 \subseteq I(R) : \pi_X(w_2) = \pi_X(w_1) \wedge \pi_X(w_4) = \pi_X(w_3) \Rightarrow \succ_k (w_2, w_4)$, which implies $\theta_X(\succ_k)(\pi_X(w_1), \pi_X(w_3))$ and thus $\pi_X(w_3) \notin \omega_\mathcal{P}^X(\pi_X(I(R)))$. □

4.3 Distributing over Cartesian Product

For preference operator to distribute over the Cartesian product of two relations, the preference specification, which is the parametr of the preference operator, needs to be decomposed into the preference specifications that will distribute into the argument relations:

Theorem 16 (Distributing over Cartesian product)**.** *Given two relation schemas R_1 and R_2, and preference specifications $\mathcal{P}_1$ over R_1 and $\mathcal{P}_2$ over R_2, for any two relation instances $I(R_1)$ and $I(R_2)$ of R_1 and R_2, the following property holds:*

$$\omega_{\mathcal{P}_0}(I(R_1) \times I(R_2)) = \omega_{\mathcal{P}_1}(I(R_1)) \times \omega_{\mathcal{P}_2}(I(R_2)) \stackrel{def}{=} \\ \{w_1 \times w_2 \mid w_1 \in \omega_{\mathcal{P}_1}(I(R_1)) \wedge w_2 \in \omega_{\mathcal{P}_2}(I(R_2))\} \ ,$$

where $\mathcal{P}_0 = P(\mathcal{P}_1, \mathcal{P}_2)$ is a Pareto composition of $\mathcal{P}_1$ and $\mathcal{P}_2$.

Proof. We prove:

$$w_1 \times w_2 \notin \omega_{\mathcal{P}_0}(I(R_1) \times I(R_2)) \Longleftrightarrow w_1 \times w_2 \notin \omega_{\mathcal{P}_1}(I(R_1)) \times \omega_{\mathcal{P}_2}(I(R_2)) \ .$$

$\Rightarrow$: Assume $w_1 \times w_2 \notin \omega_{\mathcal{P}_0}(I(R_1) \times I(R_2))$. Then $\forall \mathcal{M}_m \in \mathscr{M}_0$, models of $\mathcal{P}_0$, there are $w_1' \subseteq I(R_1), w_2' \subseteq I(R_2)$ s.t. $\succ_m (w_1' \times w_2', w_1 \times w_2)$. Consequently, $\forall \mathcal{M}_k \in \mathscr{M}_1, \forall \mathcal{M}_l \in \mathscr{M}_2$, models of $\mathcal{P}_1$ and $\mathcal{P}_2$, there are $w_1' \subseteq I(R_1), w_2' \subseteq I(R_2)$ s.t. $\succ_k (w_1', w_1)$ or $\succ_l (w_2', w_2)$, which implies $w_1 \notin \omega_{\mathcal{P}_1}(I(R_1))$ or $w_2 \notin \omega_{\mathcal{P}_2}(I(R_2))$ and thus $w_1 \times w_2 \notin \omega_{\mathcal{P}_1}(I(R_1)) \times \omega_{\mathcal{P}_2}(I(R_2))$.

$\Leftarrow$: Assume $w_1 \times w_2 \notin \omega_{\mathcal{P}_1}(I(R_1)) \times \omega_{\mathcal{P}_2}(I(R_2))$. Then $w_1 \notin \omega_{\mathcal{P}_1}(I(R_1))$ or $w_2 \notin \omega_{\mathcal{P}_2}(I(R_2))$. Assume the first. Then $\forall \mathcal{M}_k \in \mathscr{M}_1$, models of $\mathcal{P}_1$, there must be $w_1' \subseteq I(R_1)$ s.t. $\succ_k (w_1', w_1)$. Consequently, $\forall \mathcal{M}_m \in \mathscr{M}_0$, models of $\mathcal{P}_0$, $\exists w_1' \subseteq I(R_1) : \succ_m (w_1' \times w_2, w_1 \times w_2)$, which implies $w_1 \times w_2 \notin \omega_{\mathcal{P}_0}(I(R_1) \times I(R_2))$. The second case is symmetric. □

For lexicographic composition, we obtain the same property as for Pareto composition:

Theorem 17 (Distributing over Cartesian product)**.** *Given two relation schemas R_1 and R_2, and preference specifications $\mathcal{P}_1$ over R_1 and $\mathcal{P}_2$ over R_2, for any two relation instances $I(R_1)$ and $I(R_2)$ of R_1 and R_2, the following property holds:*

$$\omega_{\mathcal{P}_0}(I(R_1) \times I(R_2)) = \omega_{\mathcal{P}_1}(I(R_1)) \times \omega_{\mathcal{P}_2}(I(R_2)) \stackrel{def}{=} \{w_1 \times w_2 \mid w_1 \in \omega_{\mathcal{P}_1}(I(R_1)) \wedge w_2 \in \omega_{\mathcal{P}_2}(I(R_2))\} \ ,$$

where $\mathcal{P}_0 = L(\mathcal{P}_1, \mathcal{P}_2)$ is a lexicographic composition of $\mathcal{P}_1$ and $\mathcal{P}_2$.

Proof. We prove:

$$w_1 \times w_2 \notin \omega_{\mathcal{P}_0}(I(R_1) \times I(R_2)) \Longleftrightarrow w_1 \times w_2 \notin \omega_{\mathcal{P}_1}(I(R_1)) \times \omega_{\mathcal{P}_2}(I(R_2)) \ .$$

$\Rightarrow$: Assume $w_1 \times w_2 \notin \omega_{\mathcal{P}_0}(I(R_1) \times I(R_2))$. Then $\forall \mathcal{M}_m \in \mathscr{M}_0$, models of $\mathcal{P}_0$, there are $w_1' \subseteq I(R_1), w_2' \subseteq I(R_2)$ s.t. $\succ_m (w_1' \times w_2', w_1 \times w_2)$. Consequently, $\forall \mathcal{M}_k \in \mathscr{M}_1, \forall \mathcal{M}_l \in \mathscr{M}_2$, models of $\mathcal{P}_1$ and $\mathcal{P}_2$, there are $w_1' \subseteq I(R_1), w_2' \subseteq I(R_2)$ s.t. $\succ_k (w_1', w_1)$ or $=_k (w_1', w_1) \wedge \succ_l (w_2', w_2)$, which implies $w_1 \notin \omega_{\mathcal{P}_1}(I(R_1))$ or $w_2 \notin \omega_{\mathcal{P}_2}(I(R_2))$ and thus $w_1 \times w_2 \notin \omega_{\mathcal{P}_1}(I(R_1)) \times \omega_{\mathcal{P}_2}(I(R_2))$.

$\Leftarrow$: Assume $w_1 \times w_2 \notin \omega_{\mathcal{P}_1}(I(R_1)) \times \omega_{\mathcal{P}_2}(I(R_2))$. Then $w_1 \notin \omega_{\mathcal{P}_1}(I(R_1))$ or $w_2 \notin \omega_{\mathcal{P}_2}(I(R_2))$. Assume the first. Then $\forall \mathcal{M}_k \in \mathscr{M}_1$, models of $\mathcal{P}_1$, there must be $w_1' \subseteq I(R_1)$ s.t. $\succ_k (w_1', w_1)$. Consequently, $\forall \mathcal{M}_m \in \mathscr{M}_0$, models of $\mathcal{P}_0$, there must be w_1' s.t. $\succ_m (w_1' \times w_2, w_1 \times w_2)$, which implies $w_1 \times w_2 \notin \omega_{\mathcal{P}_0}(I(R_1) \times I(R_2))$. The second case is symmetric. □

Both Theorem 16 and Theorem 17 make it possible to derive the transformation rule that pushes preference operator with a one-dimensional preference specification down the appropriate argument of the Cartesian product:

Corollary 18. *Given two relation schemas R_1 and R_2, a preference specifications $\mathcal{P}_1$ over R_1, and an empty preference specification $\mathcal{P}_2$ over R_2, for any two relation instances $I(R_1)$ and $I(R_2)$ of R_1 and R_2, the following property holds:*

$$\omega_{\mathcal{P}_0}(I(R_1) \times I(R_2)) = \omega_{\mathcal{P}_1}(I(R_1)) \times 2^{I(R_2)} \stackrel{def}{=} \{w_1 \times w_2 \mid w_1 \in \omega_{\mathcal{P}_1}(I(R_1)) \wedge w_2 \subseteq I(R_2)\} \ ,$$

where $\mathcal{P}_0 = P(\mathcal{P}_1, \mathcal{P}_2)$ is a Pareto of lexicographic composition of $\mathcal{P}_1$ and $\mathcal{P}_2$.

Proof. Follows from previous theorems and from the equality $\omega_{\mathcal{P}_{\text{empty}}}(I(R)) = 2^{I(R)}$. □

4.4 Distributing over Union

The following theorem shows how the preference operator distributes over the union of two relations:

Theorem 19 (Distributing over union). *Given two compatible relation schemas*[3] *R and S, and a preference specification $\mathcal{P}$ over R (and S), if the following formula*

$$\omega_{\mathcal{P}}(I(R) \cup I(S)) \subseteq 2^{I(R)} \cup 2^{I(S)}$$

is valid for relation instances $I(R)$ and $I(S)$ of R and S, then the following property holds:

$$\omega_{\mathcal{P}}(I(R) \cup I(S)) = \omega_{\mathcal{P}}(\omega_{\mathcal{P}}(I(R)) \cup \omega_{\mathcal{P}}(I(S))) \ .$$

Proof. Obviously, $\omega_{\mathcal{P}}(I(R)) \cup \omega_{\mathcal{P}}(I(S)) \subseteq 2^{I(R) \cup I(S)}$. If we show that $\omega_{\mathcal{P}}(I(R) \cup I(S)) \subseteq \omega_{\mathcal{P}}(I(R)) \cup \omega_{\mathcal{P}}(I(S))$, the theorem immediately follows from Theorem 8.

Indeed, if $w \in \omega_{\mathcal{P}}(I(R) \cup I(S))$, then it follows from the definition of the preference operator $w \subseteq I(R) \cup I(S) \wedge \exists \mathcal{M}_k \in \mathscr{M}$ s.t. $\forall w' \in W : w' \subseteq I(R) \cup I(S) \wedge \succeq_k (w', w) \Rightarrow \succeq_k (w, w')$. As we know that $w \subseteq I(R) \vee w \subseteq I(S)$ from the assumption of the theorem, we can conclude $w \in \omega_{\mathcal{P}}(I(R)) \cup \omega_{\mathcal{P}}(I(S))$. □

4.5 Distributing over Difference

Only in the trivial case, the preference operator can be distributed over difference:

Theorem 20 (Distributing over difference). *Given two compatible relation schemas R and S, and a preference specification $\mathcal{P}$ over R (and S), for any two relation instances $I(R)$ and $I(S)$ of R and S, the following property holds:*

$$\omega_{\mathcal{P}}(I(R) - I(S)) = \omega_{\mathcal{P}}(I(R)) - \omega_{\mathcal{P}}(I(S))$$

iff the preference specification $\mathcal{P}$ is empty.

4.6 Push Preference

The question arises how to integrate the above algebraic laws into the classical, well-known hill-climbing algorithm. In particular, we want to add heuristic strategy of *push preference*, which is based on the assumption that early application of the preference operator reduces intermediate results. Indeed, the Theorem 8 provides a formal evidence that it is correct to pass exactly all the tuples that have been included in any world returned by the preference operator to the next operator in the operator tree. This leads to a better performance in subsequent operators.

[3] We call two relation schemas *compatible* if they have the same number of attributes and the corresponding attributes have identical domains.

5 Related Work

The study of preferences in the context of database queries has been originated by Lacroix and Lavency [8]. They, however, haven't addressed the issue of algebraic optimization.

Nevertheless, only at the turn of the millennium this area attracted broader interest again. Kießling [6] and Chomicki et al. [2] have pursued independently a similar, *qualitative* approach within which preferences between tuples are specified directly, using binary *preference relations*. They have defined an operator returning only the best preference matches. However, they, by contrast to the approach presented in this paper, don't consider preferences between *sets* of elements and are concerned only with one type of preference. Moreover, the relation to a preference logic unfortunately is unclear. On the other hand, both Chomicki et. al. [2] and Kießling [7, 4] have laid the foundation for preference query optimization that extends established query optimization techniques.

A special case of the same embedding represents *skyline operator* introduced by Börzsönyi et al. [1]. Some examples of possible rewritings for skyline queries are given but no general rewriting rules are formulated.

In [9], actual values of an arbitrary attribute were allowed to be partially ordered according to user preferences. Accordingly, RA operations, aggregation functions and arithmetic were redefined. However, some of their properties were lost, and the the query optimization issues were not discussed.

6 Conclusions

We build on the framework of embedding preferences into RQL through the preference operator that is parameterized by user preferences expressed in a declarative, logical language containing sixteen kinds of preferences and that returns the most preferred sets of tuples of its argument relation. Most importantly, the language is suitable for expressing preferences between sets of elements and its semantics allows for conflicting preferences.

The main contribution of the paper consists in presenting basic properties of the preference operator and a number of algebraic laws describing its interaction with other RA operators. Particularly, sufficient conditions for commuting the preference operator with RA selection or projection and for distributing over Cartesian product, set union, and set difference have been identified. Thus key rules for rewriting the preference queries using the standard algebraic optimization strategies like *push preference* or *push projection* have been established. Moreover, a new optimization strategy of *push preference* has been suggested.

Future work directions include identifying further algebraic properties and finding the best possible ordering of transformations for optimization of RA statements with the preference operator. Also, expressiveness and complexity issues have to be addressed in detail.

References

[1] S. Börzsönyi, D. Kossmann, and K. Stocker. The skyline operator. In 'Proceedings of the 17th International Conference on Data Engineering', 421–430, Washington, DC, USA, (2001). IEEE Computer Society.

[2] J. Chomicki. *Preference Formulas in Relational Queries.* ACM Trans. Database Syst. **28** (2003), 427–466.

[3] J. Doyle and M. P. Wellman. Representing preferences as ceteris paribus comparatives. In 'Decision-Theoretic Planning: Papers from the 1994 Spring AAAI Symposium', 69–75. AAAI Press, Menlo Park, California, (1994).

[4] B. Hafenrichter and W. Kießling. Optimization of relational preference queries. In 'CRPIT '39: Proceedings of the sixteenth Australasian conference on Database technologies', 175–184, Darlinghurst, Australia, Australia, (2005). Australian Computer Society, Inc.

[5] S. Kaci and L. W. N. van der Torre. Non-monotonic reasoning with various kinds of preferences. In 'IJCAI-05 Multidisciplinary Workshop on Advances in Preference Handling', Ronen I.Brafman and U. Junker, (eds.), 112–117, (August 2005).

[6] W. Kießling. Foundations of Preferences in Database Systems. In 'Proceedings of the 28th VLDB Conference', 311–322, Hong Kong, China, (2002).

[7] W. Kießling and B. Hafenrichter. Algebraic optimization of relational preference queries. Technical Report 2003-01, Institute of Computer Science, University of Augsburg, (February 2003).

[8] M. Lacroix and P. Lavency. Preferences; Putting More Knowledge into Queries. In 'VLDB', P. M. Stocker, W. Kent, and P. Hammersley, (eds.), 217–225. Morgan Kaufmann, (1987).

[9] R. Nedbal. *Relational Databases with Ordered Relations.* Logic Journal of the IGPL **13** (2005), 587–597.

[10] R. Nedbal. Non-monotonic reasoning with various kinds of preferences in the relational data model framework. In 'ITAT 2007, Information Technologies – Applications and Theory', P. Vojtáš, (ed.), 15–20, Poľana, (September 2007). PONT.

[11] G. von Wright. *The logic of preference.* Edinburgh University Press, Edinburgh, (1963).

Numerická simulace dislokační dynamiky*

Petr Pauš

2. ročník PGS, email: pauspetr@fjfi.cvut.cz
Katedra matematiky, Fakulta jaderná a fyzikálně inženýrská, ČVUT
školitel: Michal Beneš, Katedra matematiky, Fakulta jaderná a fyzikálně inženýrská, ČVUT

Abstract. This paper deals with the numerical simulation of dislocation dynamics. Dislocations are described by means of the evolution of a family of closed and open smooth curves $\Gamma^t : S^1 \to \mathbb{R}^2$, $t \geqq 0$. The curves are driven by the normal velocity ν which is the function of curvature κ and the position vector $x \in \Gamma^t$. In this case the equation is defined this way: $\nu = -\kappa + F$. The equation is solved using direct approach by two numerical schemes, ie. semi-implicit and semi-discrete, both are compared with analytical solution. Results of the dislocation dynamics simulation are presented.

Abstrakt. Tento článek se zabývá numerickou simulací dislokační dynamiky. Dislokace jsou popsány pomocí časového vývoje množiny uzavřených a otevřených hladkých křivek $\Gamma^t : S^1 \to \mathbb{R}^2$, $t \geqq 0$. Vývoj křivek je ovlivňován normálovou rychlostí ν, jenž je funkcí křivosti κ a polohového vektoru $x \in \Gamma^t$. V tomto případě má rovnice tvar $\nu = -\kappa + F$. Rovnice je řešena přímou metodou pomocí dvou různých numerických schémat, semi-implicitním a semi-diskrétním. Obě tato schémata jsou porovnána s analytickým řešením. Výsledky simulace dislokační dynamiky jsou také uvedeny.

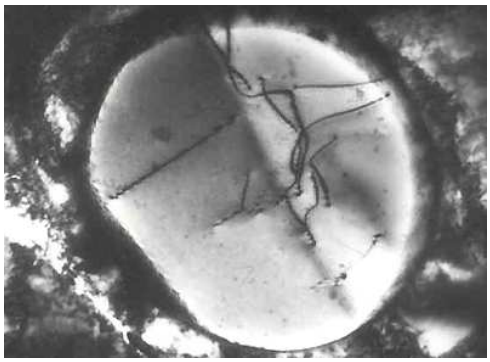
1 Úvod

V odvětví výzkumu materiálů a pevných látek se dislokace definují jako porucha či nepravidelnost v krystalové mřížce materiálu. Přítomnost dislokací v materiálu výrazně ovlivňuje mnohé z jeho vlastností, a proto je velmi důležité vyvinout vhodný matematický model. Z matematického hlediska jsou dislokace definovány jako hladké uzavřené nebo otevřené rovinné křivky, které se vyvíjejí v čase. Ukázka dislokace v materiálu je znázorněna na obrázku 1.

2 Matematický model

Křivky, které se v čase vyvíjejí, je možné matematicky popsat několika způsoby. Jednou z možností je přístup metodou *Level Set* [1, 2, 3], kdy výsledná křivka je nulovou hladinou plochy. Další možností je metoda fázového pole (Phase Field) [4] a nakonec i přímá (parametrická) metoda [5, 6], kdy je křivka parametrizována běžným způsobem. V tomto článku se budeme zabývat právě touto metodou.

*Tato práce byla podpořena grantem č. MSM 6840770010, projekt č. LC06052 Nečasova centra pro mathematické modelování.



Obrázek 1: Dislokace v nerezové oceli, zdroj: Wikipedia.org

2.1 Rovnice

Při použití parametrického přístupu popíšeme dislokační křivku $\Gamma(t)$ pomocí hladké časově závislé vektorové funkce $X(S, I)$

$$\mathbf{X} : S \times I \to \mathbb{R}^2,$$

kde $S = \langle 0, 1\rangle$ je pevný interval pro parametrizaci křivky a $I = \langle 0, T\rangle$ je časový interval. Dislokační křivka $\Gamma(t)$ je pak dána jako

$$\Gamma(t) = \{\mathbf{X}(u, t), u \in S\},$$

kde u je parametr a $t \in \langle 0, T\rangle$ je čas.

Množina těchto křivek musí splňovat rovnici pro časový vývoj, která je obecně zadána jako $\nu = \beta(\kappa, \mathbf{x})$, kde ν je normálová rychlost vývoje křivky. Normálová rychlost je obecně funkcí křivosti κ a polohového vektoru $\mathbf{x}$. V našem případě má tato rovnice jednoduchou podobu

$$\nu = -\kappa + F, \tag{1}$$

kde F vyjadřuje externí sílu aplikovanou ve směru normály ke křivce.

2.2 Odvození diferenciální rovnice

Pro odvození diferenciální rovnice je třeba definovat několik základních pojmů. Mějme hladkou křivku $x : S^1 \to \mathbb{R}^2$. Jednotkový tečný vektor $\vec{T}$ lze definovat jako $\vec{T} = \partial_u x/|\partial_u x|$. Jednotkový normálový vektor $\vec{N}$ je kolmý na tečný vektor a platí $\vec{N}\vec{T} = 0$. Křivost křivky je dána vzorcem

$$\kappa = \frac{\partial_u x^\perp}{|\partial_u x|} \cdot \frac{\partial_u^2 x}{|\partial_u^2 x|} = \vec{N} \cdot \frac{\partial_u^2 x}{|\partial_u^2 x|}.$$

Normálová rychlost je definována jako derivace x podle času t ve směru normály.

Nyní lze zapsat rovnici (1) jako

$$\frac{\partial x}{\partial t} = -\frac{\partial_u^2 x}{|\partial_u^2 x|} + F\frac{\partial_u x}{|\partial_u x|}. \tag{2}$$

2.3 Numerické schéma

Pro řešení diferenciální rovnice (2) jsou použita dvě numerická schémata: semi-implicitní a semi-diskrétní. Použití dvou schémat je z důvodu možnosti porovnání výsledků a zjištění jejich přesností.

Pro řešení rovnice pomocí semi-diskrétního schématu je použita Runge-Kuttova metoda čtvrtého řádu s automatickou volbou časového kroku. Pro diskretizaci derivací v prostoru jsou použity středové diference rovněž čtvrtého řádu. První derivace je diskretizována takto:

$$\frac{\partial x}{\partial u} \approx \left[\frac{x^1_{j-2} - 8x^1_{j-1} + 8x^1_{j+1} - x^1_{j+2}}{12h}, \frac{x^2_{j-2} - 8x^2_{j-1} + 8x^2_{j+1} - x^2_{j+2}}{12h}\right],$$

a druhá derivace takto:

$$\frac{\partial^2 x}{\partial u^2} \approx \left[\frac{-x^1_{j-2} + 16x^1_{j-1} - 30x^1_j + 16x^1_{j+1} - x^1_{j+2}}{12h^2}, \frac{-x^2_{j-2} + 16x^2_{j-1} - 30x^2_j + 16x^2_{j+1} - x^2_{j+2}}{12h^2}\right].$$

Náhrady derivací označíme x_u pro první derivaci a x_{uu} pro druhou.

Rovnice (2) pro semi-diskrétní schéma má tvar

$$\frac{dx_j}{dt} = \frac{x_{uuj}}{Q^2(x_{uj})} + F\frac{x^\perp_{uj}}{Q(x_{uj})}, \quad j = 1, \cdots, m-1, t \in (0, T), \tag{3}$$

kde $Q(x, y) = \sqrt{x^2 + y^2 + \varepsilon^2}$ a $x^\perp_{uj}$ je kolmý vektor k x_{uj}. Člen ε je do rovnice přidán proto, aby při řešení nedocházelo k ukončení výpočtu při dosažení singularity.

Používá se i semi-implicitní schéma řešené přímým řešičem. V tomto případě jsou použity jednodušší prostorové diference. První derivace je diskretizována pomocí zpětné diference

$$\frac{\partial x}{\partial u} \approx \left[\frac{x^1_j - x^1_{j-1}}{h}, \frac{x^2_j - x^2_{j-1}}{h}\right]$$

a druhá derivace je diskretizována takto:

$$\frac{\partial^2 x}{\partial u^2} \approx \left[\frac{x^1_{j+1} - 2x^1_j + x^1_{j-1}}{h^2}, \frac{x^2_{j+1} - 2x^2_j + x^2_{j-1}}{h^2}\right].$$

Rovnice (2) pro semi-implicitní schéma má tvar

$$x^{k+1}_j - \tau\frac{x^{k+1}_{uuj}}{Q^2(x^k_{uj})} = x^k_j + \tau F\frac{x^{\perp k}_{uj}}{Q(x^k_{uj})}, \quad j = 1, \cdots, m-1, k = 0, \cdots, N_T - 1, \tag{4}$$

kde $Q(x, y)$ a $x^\perp_{uj}$ mají stejný význam jako pro semi-diskrétní schéma. Struktura matice pro jednu komponentu x^{k+1} má tvar

$$\begin{pmatrix} 1 - \frac{2t}{h^2Q^2} & \frac{-t}{h^2Q^2} & 0 & \cdots \\ \frac{-t}{h^2Q^2} & \ddots & \ddots & \ddots \\ 0 & \ddots & & \\ \vdots & \ddots & & \end{pmatrix}$$

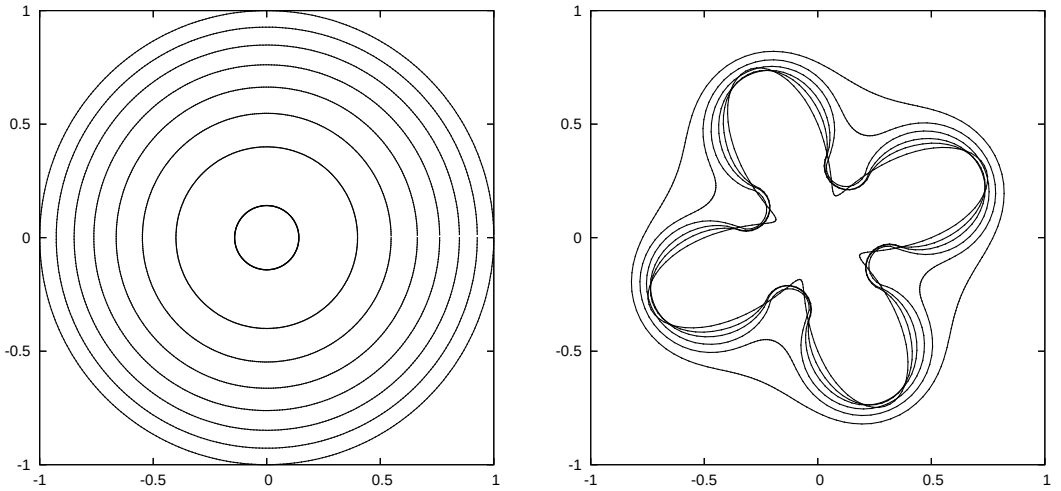
.

3 Výsledky numerické simulace

V této kapitole budou prezentovány výsledky numerické simulace pomocí výše uvedených numerických schémat. Lety prověřený řešič pro semi-diskrétní schéma používá Runge-Kuttovu metodu čtvrtého řádu a je napsán a pro rychlost odladěn v jazyce Fortran. Řešič pro semi-implicitní schéma je napsán v jazyce C, zatím však není tak sofistikovaný. Na jeho vývoji se stále pracuje.

3.1 Testování na uzavřených křivkách

K ověření, zda je zvolený přístup k řešení rovnice vhodný, bylo nutno provést testování s různými počátečními podmínkami. Výsledky byly poté porovnány buď s analytickým řešením nebo s výsledky dostupnými v literatuře [6, 7, 8].



(a) $F = 0$, $t \in (0, 0.495)$, $h = 1/200$

(b) $t \in (0, 0.184)$, $h = 1/200$, $F = 10$ pro $|x| < 0.35$, $F = -5$ pro $|x| > 0.35$

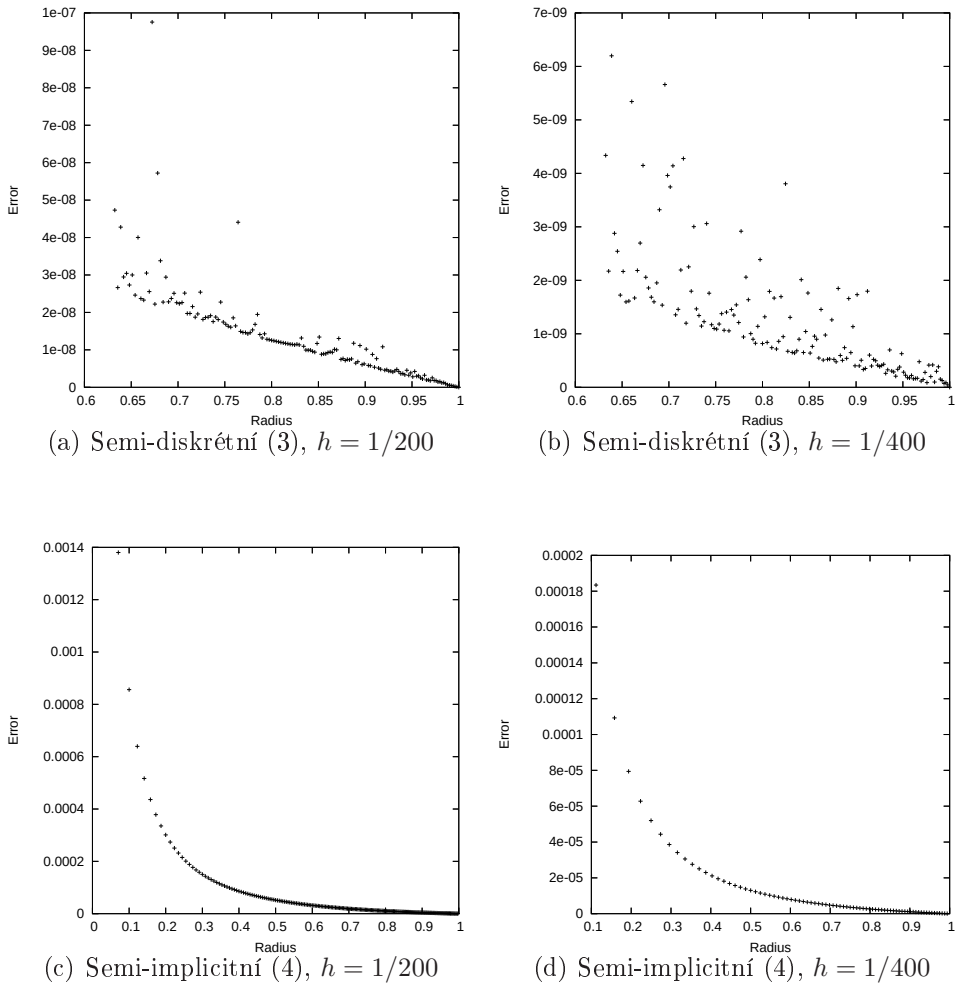
Obrázek 2: Časový vývoj uzavřených křivek

Nejjednodušší případ pro ověření řešení rovnice (2) je zvolit počáteční podmínku jako jednotkový kruh se středem v počátku souřadnic a jako externí sílu zvolit 0, tj. žádnou externí sílu. Simulace tohoto případu je vidět na obrázku 2(a). Bez externí síly se kruh postupně zmenšuje a nakonec zůstane pouze jeden bod. Pro tento případ je také možné zjistit analytické řešení, které je dané vzorcem

$$r = \sqrt{1 - 2t}, \tag{5}$$

kde r je poloměr zmenšujícího se kruhu a t je čas. Porovnání numerického výpočtu a analytického řešení je uvedeno níže. Při zvolení externí síly rovné křivosti si kruh zachová svou velikost a v čase se pak nebude vyvíjet.

Na obrázku 2(b) je zobrazena komplikovanější situace, kdy externí síla F není v prostoru konstantní, ale pro body $|x| < 0.35$ je $F = 10$, pro ostatní je $F = -5$. Vzhledem k vyšší křivosti v záhybech je ale síla $F = 10$ překonána a tvar se postupně transformuje na kruh, který se dále rozpíná.



(a) Semi-diskrétní (3), $h = 1/200$ (b) Semi-diskrétní (3), $h = 1/400$

(c) Semi-implicitní (4), $h = 1/200$ (d) Semi-implicitní (4), $h = 1/400$

Obrázek 3: Chyba schémat vůči analytickému řešení

3.2 Ověření numerického řešení

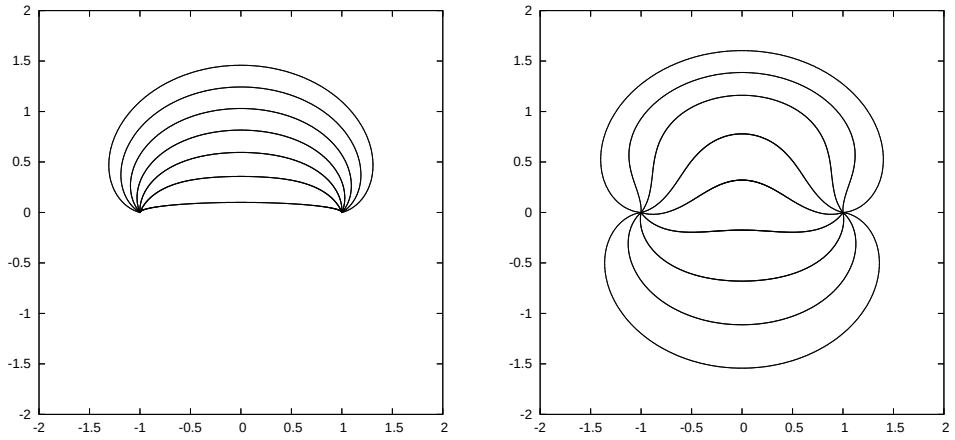
Jak již bylo zmíněno, rovnici (2) je možné pro jisté případy počátečních podmínek řešit i analyticky. Jedním z těchto případů je počáteční podmínka ve tvaru kruhu s jednotkovým poloměrem. Řešení je pak dáno vzorcem (5). Obě použitá numerická schémata (3) (4) byla porovnána s tímto řešením. Chyba je daná rozdílem analytického řešení a numerického, tj. $r_{analyt} - r_{numer}$.

Výsledky pro semi-diskrétní schéma zachycují grafy 3(a) a 3(b). Je vidět, že se zmenšením velikosti h se také sníží velikosti chyby. Díky použití metody čtvrtého řádu a také diferencí téhož řádu se dosáhlo dobré přesnosti v řádu 10^{-7} pro $h = 1/200$, resp. 10^{-9} pro $h = 1/400$.

Semi-implicitní řešič zatím nedosahuje takových přesností jako semi-diskrétní řešič. V grafech 3(c) a 3(d) je zobrazen rozdíl analytického a numerického řešení pro stejnou situaci jako v předchozím případě. Opět zde dochází k růstu přesnosti při snižování velikosti h, tj. přesnost je v řádu 10^{-3} pro $h = 1/200$, resp. 10^{-4} pro $h = 1/400$.

3.3 Dislokační dynamika

Hlavním cílem této metody bude simulace dislokační dynamiky. Dislokační křivky v materiálu se různě vyvíjejí, kmitají mění svou topologii a podobně. Následující simulace mají ověřit, zda metoda může být použita pro tento účel.

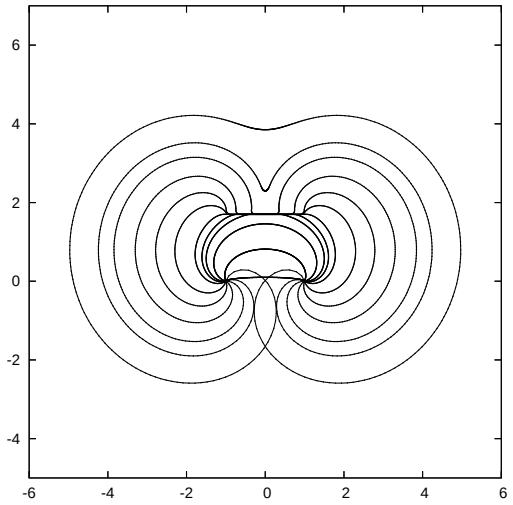


(a) Dislokační křivka expanduje ($F = -3$) (b) Křivka se stáhne a expanduje na druhou stranu.

Obrázek 4: Vývoj dislokační křivky s proměnlivou externí silou F

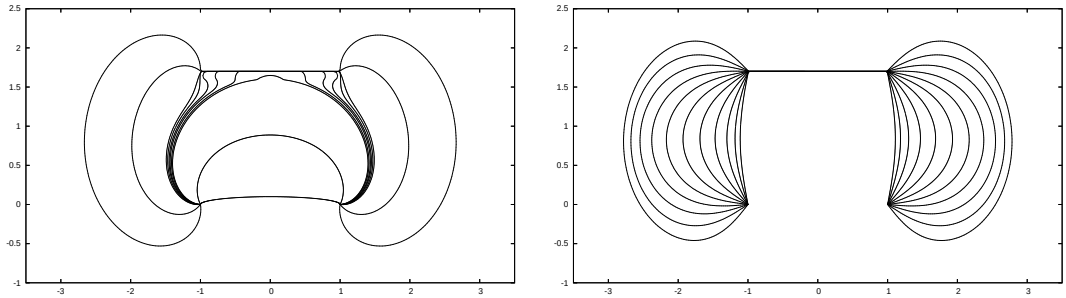
Obrázek 4 ilustruje průběh dislokační křivky v čase. Na křivku působí externí síla $F = -3$, což způsobí expanzi nahoru. V čase $t = 0,54$ se změní směr síly $F = 3$. Křivka se vrací do původního stavu a prokmitne na druhou stranu. V materiálu se objevuje právě podobné chování, ale je ovlivněno více faktory.

Při evoluci dislokační křivky se může stát, že se objeví bariéra, která brání v jejím vývoji. Podle velikosti síly se může stát, že křivka buď bariéru překoná nebo zůstane v této bariéře uzamčena. Obrázek 5 ilustruje právě tuto situaci. Dislokační křivka expanduje za působení externí síly $F = -3$, dokud nedosáhne bariéry tvořené prostorovou změnou externí síly F v $y = 1.7$. Tato bariéra ovšem není dost silná ($|F| = 9$), aby křivku udržela, protože na koncích je velmi vysoká křivost, a tím pádem i vysoká síla působící proti externí



Obrázek 5: Dislokační křivka překoná bariéru tvořenou prostorově proměnnou externí silou.

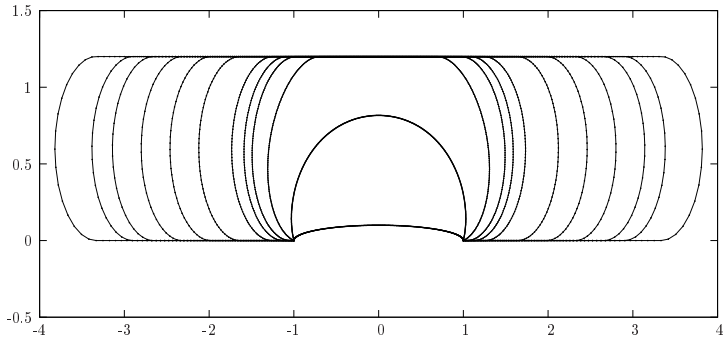
síle. Křivka se uvolní a pokračuje dál v expanzi. Tato simulace se odehrává v časovém intervalu $t \in (0, 2.1)$. V reálném případě by se po dotknutí různých částí křivky měla křivka rozdělit na více částí, které se pak vyvíjejí dále. Toto zatím současný matematický model nedovoluje.



(a) Expanze křivky (b) Návrat křivky zpět

Obrázek 6: Prostorově proměnná externí síla F s vysokou hodnotou

Když je externí síla příliš silná, tak nedovolí křivce pokračovat ve vývoji. Křivka je tedy držena mezi dvěma body a expanduje do stran. Bariéra je opět v $y = 1.7$ a hodnota externí síly v ní je $|F| = 35$. Obrázek 6(a) zobrazuje expanzi křivky a její zachycení při $F = -3$ a v čase $t \in (0, 1.5)$. Obrázek 6(b) zobrazuje návrat křivky zpět při $F = 3$ v čase $t \in (1.5, 3)$. Křivka je ovšem zadržena v bariéře a nemůže se vrátit do původního stavu. Tato simulace má demonstrovat vývoj dislokační křivky v materiálu, kdy zachycena v tzv. kanálu.



Obrázek 7: Vývoj křivky v kanálu

Vývoj křivky v kanálu je znázorněn na obrázku 7. Pomocí externí síly jsou zde vytvořeny nekonečně dlouhé bariéry, přes které se dislokace nemůže dostat. Vyvíjí se tedy pouze do stran. V reálném případě by pak měla dislokace kmitat právě v tomto kanálu.

4 Závěr

Modelování dislokační dynamiky je pro praxi velmi důležité, protože dislokace mají velký vliv na většinu vlastností materiálu. Dislokace je možné modelovat pomocí v čase se vyvíjejících hladkých křivek. V tomto článku je uveden matematický model pro řešení pomocí parametrického přístupu a navržena dvě numerická schémata pro řešení rovnice na počítači. Výsledky jsou porovnány buď s analytickým řešením nebo s výsledky v literatuře. Uvedeny jsou také simulace demonstrující vývoj křivek podobný realným dislokacím, ale ty zatím plně neodpovídají, protože do modelu není implementována vhodná fyzika.

Literatura

[1] J. A. Sethian. *Level set methods.* Cambridge University Press, 1996.

[2] J. A. Sethian. *Level set methods and fast marching methods.* Cambridge University Press, 1999.

[3] G. Dziuk, A. Schmidt, A. Brillard, and C. Bandle. *Course on mean curvature flow.* Freiburg, 1994.

[4] M. Beneš *Phase field model of microstructure growth in solidification of pure substances.* Praha, 1997.

[5] K. Deckelnick, G. Dziuk. *Mean curvature flow and related topics.*

[6] K. Mikula, D. Ševčovič. *Computational and qualitative aspects of evolution of curves driven by curvature and external force.* Computing and Visualization in Science, 2004.

[7] K. Mikula, D. Ševčovič. *Evolution of plane curves driven by a nonlinear function of curvature and anisotropy.* SIAM Vol. 61, No. 5, pp. 1473-1501, 2001.

[8] K. Mikula, D. Ševčovič. *A direct method for solving an anisotropic mean curvature flow of plane curves with and external force.* Mathematical methods in the applied sciences, 2004.

Optimisation of Tree-Structured Self-Organising Maps

Philip Prentis*

4th year of PGS, email: prentisp@km1.fjfi.cvut.cz
Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU
advisor: Ladislav Andrej, Institute of Computer Science, AS CR

Abstract. This paper describes two variants of the tree-structured self-organising map (TS-SOM) algorithm that are used in two different image-browsing and retrieval systems, GalSOM and PicSOM. It studies how methods used to optimise GalSOM work in PicSOM and makes some suggestions on parameter optimisation based on the results of heuristic analysis.

Abstrakt. Tento článek popisuje dvě varianty algoritmu samoorganizující se mapy se stromovou strukturou, které jsou použity ve dvou různých systémech prohlížení a vyhledávání obrazu, GalSOM a PicSOM. Studuje jak metody použitívané k optimalizaci GalSOM pracují v systému PicSOM a navrhuje několik možností jak optimalizovat jeho parametry na základě výsledků heuristické analýzy.

1 Introduction

In our previous workshop paper [10], we described the tree-structured variant of the well-known self-organising map (SOM) algorithm, and some of the problems that can occur when using it for multi-resolution visualisation tasks. In this paper we analyse a variant of the tree-structured self-organising map and explore some aspects of the task of optimising its parameters.

Self-organising maps are a type of artificial neural network that uses unsupervised learning. They were developed by Kohonen [2] in the eighties and have since been employed successfully in a number of applications; in particular they have been used for cluster analysis and visualising high-dimensional data.

In an effort to speed up large SOM and acquire mappings at different resolutions, hierarchical variants were developed. These include the Multi-Layer SOM [1], the Evolving Tree [7] and most importantly, the Tree-Structured Self-Organising Map (TS-SOM) [3, 4], which this paper describes in detail. One of the best-known applications to use TS-SOM is PicSOM [5, 6], which uses multiple TS-SOM to facilitate content-based image retrieval.

Incorporating TS-SOM into the GalSOM image browser led to an improvement of the algorithm with *multi-resolution correction* as described in [8]. In this paper we apply the results from experiments with GalSOM to the PicSOM in an attempt to optimise it. More general information about using TS-SOM for image browsing and data analysis with GalSOM is discussed in [9, 11].

*In cooperation with J. Laaksonen, M. Koskela and M. Sjöberg of *The Laboratory of Computer and Information Science, Helsinki University of Technology.*

2 Algorithms

In this section we describe the SOM and TS-SOM algorithms. Because the two primary applications of the TS-SOM, GalSOM and PicSOM, implement it differently, they are described separately in sections 2.2 and 2.3.

2.1 SOM

The self-organisation process is achieved as follows:

1. Initialise the codebook vectors $n_{ij}(0)$ at random (usually by setting them to randomly chosen input vectors).

2. Select a random input $i(t)$ and find the best matching neuron (BMN) $n_{best}(t)$ (i.e. the neuron with the closest codebook vector). Every input sample has the same probability of being selected.

3. Move the BMN and its topological neighbours within a certain neighbourhood distance towards the selected input vector. Units located topologically further from BMN are moved less.

$$n_{ij}(t+1) = n_{ij}(t) + \eta(t) \cdot \phi(i,j,t) \cdot [i(t) - n(t)], \tag{1}$$

   where

$$\eta(t): \ N_0 \rightarrow <0;1> \qquad \text{monotonously decreasing}, \tag{2}$$
$$\phi(i,j,t): \ N_0 \times N_0 \times N_0 \rightarrow <0;1>,$$

   ϕ decreases monotonously with the topological distance of n_{ij} from n_{best} and with t. The topological distance is the length of the shortest path from one neuron to the other in the graph (grid) that represents the network's topology.

4. Proceed to iteration $t+1$. Repeat 2 and 3 iteratively, reducing the proportion of the distance moved η and the neighbourhood distance ϕ each iteration, until they reach a certain predetermined threshold.

As a result the codebook vectors will be attracted to large clusters of input vectors as these will have a higher probability of being selected than sparsely populated areas of input space. η and ϕ must be selected with care if the algorithm is to achieve good results [12].

2.1.1 Neighbourhood function

Function ϕ in (2) is called the *neighbourhood function*. It determines how much and how distant (from the BMN) neurons will be affected at a given moment during the adaptation process. Typically, it will apply a *neighbourhood kernel* to the topology of the network surrounding the BMN to determine how much neurons within a the current topological radius should be affected.

By default, GalSOM uses a triangular neighbourhood kernel, which is defined as follows.

$$\phi_{\mathrm{Tri}}(t) = \begin{cases} 1 - d/r(t) & d \leq r(t) \\ 0 & d \geq r(t) \end{cases}, \tag{3}$$

where d is the topological distance of a neuron from the BMN and $r(t)$ is the neighbourhood radius at time t.

2.2 TS-SOM (GalSOM)

One way we can simultaneously analyse input space with high and low input ratios is to use a tree-structured self-organising map (TS-SOM), [3, 4]. This is a hierarchical structure of SOMs of exponentially increasing size. Each level of the TS-SOM adapts separately, but in the lower levels, the search for the best-matching neuron is limited to those hierarchically connected to the BMN of the previous layer. See figure 1.

2.2.1 The basic algorithm

The algorithm works as follows:

1. Perform one iteration of the SOM algorithm on the top layer.
2. Perform one iteration of the SOM algorithm on the next layer, but limit the search for the BMN to the neurons located under the winning neuron of the previous layer.
3. Repeat 2 until all layers have been updated.
4. Repeat 1 to 3 until the SOM thresholds have been met.

The advantages of such a structure are obvious. Instead of performing a full-search for the BMN at the lower layers, we restrict ourselves to a constant number of neurons per given layer, thus greatly increasing the adaptation speed. The complexity of the algorithm is $O(logN)$, where N is the number of neurons on the bottom layer [4]. Also, due to the hierarchical structuring, all the SOM will be orientated similarly in input space and the TS-SOM as a whole may be considered a multi-resolution mapping of the given data set. See figure 2.

Unfortunately, reducing the scope when searching for the BMN will often return suboptimal results, i.e. finding neurons that are further from the input than the closest one. As shown in my detailed analysis [8], this effect increases with each subsequent layer causing the lower high NI-ratio layers to return poor results.

2.2.2 Wide-search TS-SOM

The unfortunate property of the TS-SOM to propagate errors to the lower layers is caused by inputs bordering between two neurons on a higher layer, which gradually become more and more poorly quantified as the search for the BMN becomes more and more restricted. As noted in [6], better results may be achieved by allowing searching for the BMN in a wider scope, which includes neurons adjacent to those directly under a higher layer (figure 3). This is further corroborated by my experiments in [8], where I show that wide-searching is superior to the standard TS-SOM in almost all respects.

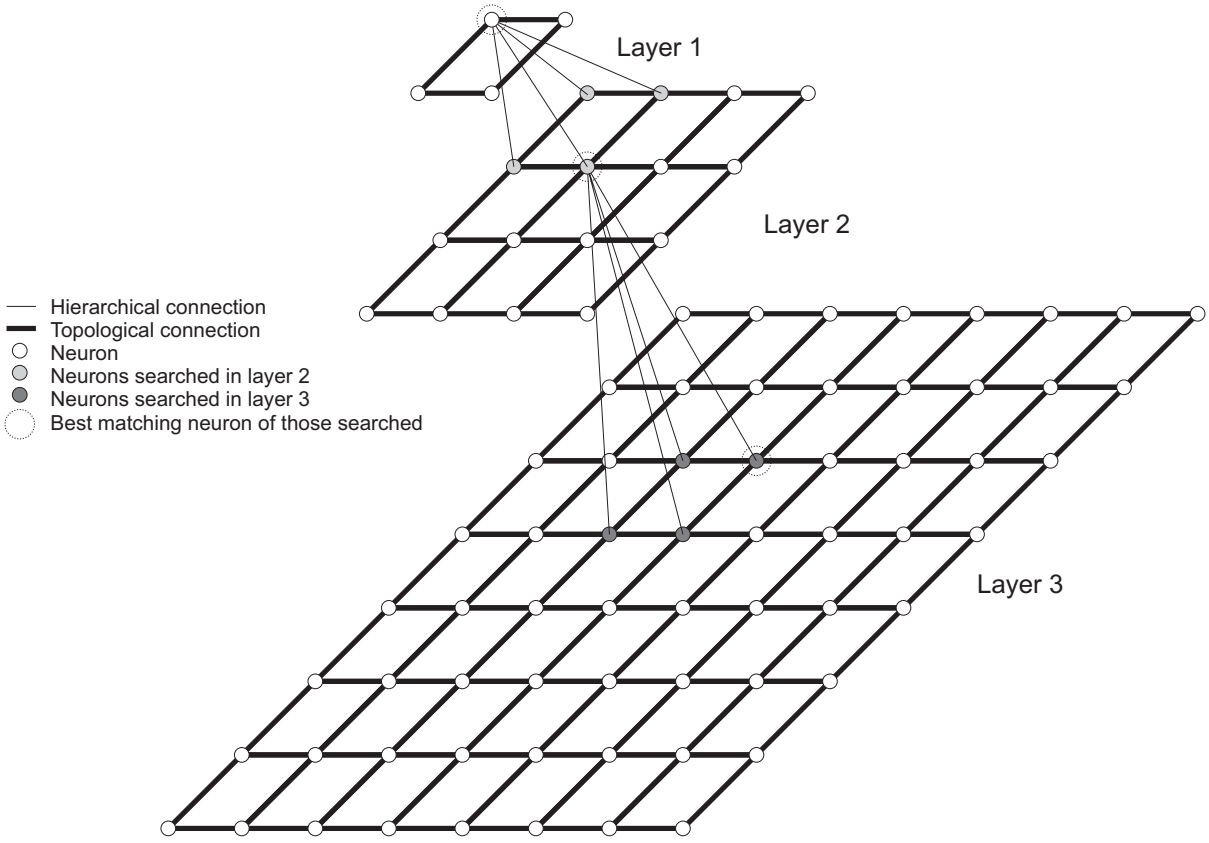


Figure 1: A 3-layer TS-SOM with 4 neurons at the top layer and 64 at the bottom.

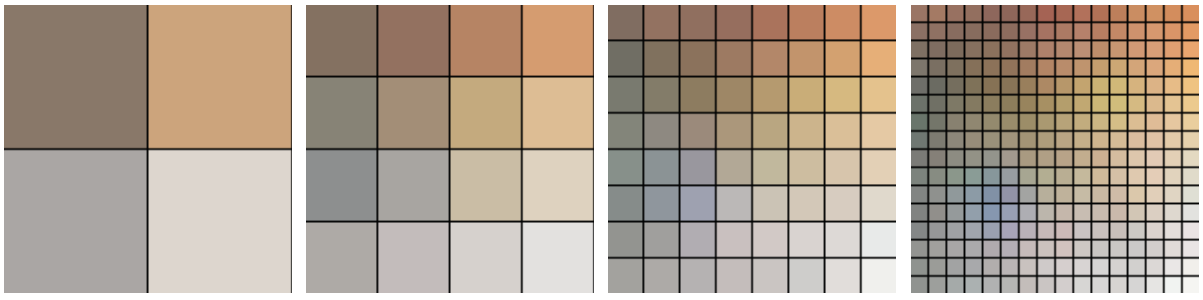


Figure 2: 4 layers of a TS-SOM show colour distribution of an input space of images at different resolutions.

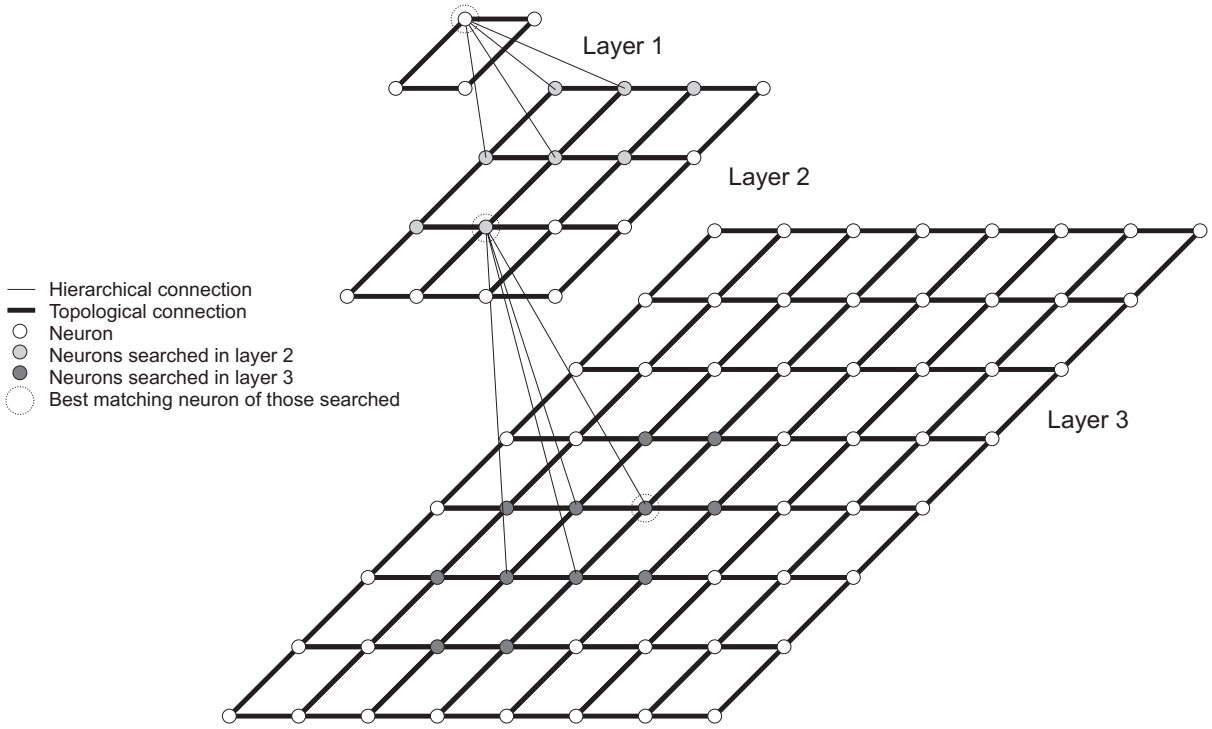


Figure 3: Wide-search for the BMN in neighbouring neurons.

2.2.3 Multi-resolution correction

One unfortunate side effect of the wide-search improvement of the TS-SOM algorithm is that the separate layers become unsynchronised. This degrades the quality of the TS-SOM as a multi-resolution mapping. In [8] we describe a simple and effective method for rectifying this problem and removing the desynchronising effect. It should be noted that good synchronisation also improves the quality of the search for the BMN and brings the TS-SOM closer to the quality of the standard full-search methods. However, as we will show later on, *this property only applies to the GalSOM variant of the TS-SOM.*

2.3 TS-SOM (PicSOM)

The PicSOM system [5] was originally developed for the task of content-based image retrieval using query by pictorial example [6]. The data is mapped to multiple SOM, each of which maps a different representation of it using different feature vectors.

The user is presented with a number of images from the database, which he labels either positive (similar to the target image) or negative. This feedback is plotted to the BMN of each TS-SOM and a "value map" of each is updated — the unit on the map that corresponds to the BMN for positively-labelled images has its value increased, while that corresponding to the negatively-labelled ones is decreased. A convolution filter and normalisation is then applied to the value-maps to generalise the results. A new set of images located at the nodes that score highest on the value-maps is then presented to the user, and the process is repeated until the target image has been found.

The user feedback can be replaced with annotated training samples and used for

learning to detect high-level features and classes using the multiple mappings of low-level features.

Only the bottom levels of the TS-SOM are used, so the algorithm is optimised towards getting the best result on them. The only purpose of the other levels is to speed up the search process. For this reason, the TS-SOM structure used by PicSOM is slightly different than the ones that GalSOM uses, as were described in section 2.2.

In a PicSOM TS-SOM, each node is hierarchically connected to an area of 4×4 nodes on the subsequent level. Thus, each map level contains 16 times as many nodes as the previous one, causing the size to increase with great rapidity. PicSOM uses wide-searching similar to that described in section 2.2.2, but the search is restricted to the 10×10 area beneath a node rather than just the immediate neighbours of the descendants.

The second major difference is that the levels of the TS-SOM are adapted one at a time, starting from the top level. Each subsequent level has its neurons initiated by extrapolating the position of those on the previous level according to the hierarchical connections. This is to improve the performance on the lower levels per given number of iterations by reusing information gained from the iterations made on the higher ones.

3 Experiments

We have conducted a number of experiments studying and comparing the TS-SOM methods used in GalSOM and PicSOM.

3.1 MRC with PicSOM

In this experiment we compared the basic PicSOM algorithm with variants using the multi-resolution correction as described in section 2.2.3. Four different variants were used.

1. 100 runs[1] using the default PicSOM configuration.
2. As in 1, but with MRC performed after all runs have been completed.
3. MRC is performed twice, once every 50 runs.
4. MRC is performed after every run.

In 3 and 4 the runs following each MRC begin at the top level again.

Table 1 shows the average quantisation error (AQE) and tree search quality (TSQ) of the 4 variants on each level of 4-level TS-SOM. As can be seen, variants 2-4 have a higher AQE on the more important lower levels, but an improved result on the higher ones.

TSQ is defined as the percentage of BMN located using tree search that are the same as BMN located using a full search on a given level of the map. As can be seen, MRC reduces the TSQ on all levels except the top ones, where tree search is equivalent to full search.

[1] A run consists of one iteration performed for each input.

MRC test								
Measure	AQE				TSQ			
Level	1.	2.	3.	4.	1.	2.	3.	4.
1	197,213	189,362	188,764	190,869	1	1	1	1
2	159,151	153,726	153,878	159,857	0,951545	0,900166	0,901104	0,928091
3	128,144	113,595	113,772	126,069	0,914735	0,811203	0,813411	0,709161
4	72,9938	85,7821	85,6351	101,604	0,941446	0,812583	0,812804	0,707726

Table 1: MRC used in variants 2-4 reduces AQE on the lower levels of the TS-SOM, but increases it on the higher ones. It reduces TSQ on all levels except the top ones.

MRC test - HLF				
Measure	1.	2.	3.	4.
Avgprec	0,1472	0,1444	0,1325	0,1409
W. avgprec	0,0112	0,0112	0,0102	0,0109
A priori avgprec	16,3233	16,2401	14,2412	16,3332

Table 2: MRC decreases the average precision regardless of whether it is weighted in favour of small or large classes.

Table 2 shows the average, the weighted average, and the *a priori* average *average precision* (avgprec) for locating objects containing queried high-level features (HLF) of 34 classes. These classes includes such HLF as *people*, *water*, *buildings* and *crowds*.

Because each class had a different number of objects for training and testing, it was necessary to examine whether small classes with few objects were not skewing the results. This was done with the *weighted average*, where the avgprec of each class was weighted by the number of objects it contained.

However, because objects in small classes have a much lower *a priori* probability of being located in a large database, a higher avgprec may be considered a greater achievement. The *a priori avgprec* of class i is calculated by dividing its avgprec by its *a priori* probability P_i, which is defined as follows.

$$P_i = O_i/N, \tag{4}$$

where O_i is the number of objects in class i and N is the total number of objects in the entire database. Unfortunately, this measure is more susceptible to noise than the other two as it magnifies the results of small classes.

As can be seen from the results in table 2, variant 1 scored the best in all three different measures.

MRC essentially damages the lower-level mappings, which were produced using a static positioning of the higher ones. Because the lower levels were produced recycling the iterations of the higher levels as well as their own, they are much more finely tuned. MRC recycles this fine-tuning when it corrects the positions of the higher levels. However, in non-visualisation tasks that only use the lowest level of the PicSOM TS-SOM (such as HLF detection), MRC provides no benefit at all.

Tri test - improvement %		
Level	AQE	TSQ
1	4,36	0
2	2,50	0,0035
3	2,52	0,0176
4	8,48	0,0375

Table 3: Using a triangular neighbourhood kernel improves AQE and TSQ on all levels of the TS-SOM.

MRC test - HLF		
Measure	Rect	Tri
Avgprec	0,1190	0,1281
W. avgprec	0,0092	0,0098
A priori avgprec	14,5450	15,1471

Table 4: Using a triangular neighbourhood kernel improves HLF detection.

3.2 Triangular neighbourhood kernel

The default configuration of PicSOM uses a rectangular neighbourhood kernel (see section 2.1.1). In this experiment we replaced it with a triangular one and compared the results.

Table 3 shows the percentage of improvement to average AQE and TSQ of 7 separate tests using different features, which this modification achieved. As can be seen, using a triangular kernel improved the map quality of each level of the TS-SOM. This also applies to each separate feature as well as the average score.

Table 4 compares the HLF test results of rectangular and triangular neighbourhood kernels using the average of the same 7 features. The avgprec (using the same three measures as in the previous experiment) was on average better in each case, although for one or two features the *a priori* avgprec decreased slightly.

Overall it may be concluded that the triangular outperforms the rectangular neighbourhood kernel.

3.3 Convolution radius

During the convolution phase of the PicSOM algorithm, a triangular kernel of radius C (not to be confused with the neighbourhood kernel in the previous section) is used to generalise the input from training data or user feedback (see section 2.3). The default value of C is 9 (i.e. the values on the value map of neurons within a topological range of 9 from the BMN of positive or negative hits will be modified in proportion to their distance).

The graph in figure 4 shows the avgprec of each of the 34 classes examined at values of C ranging from 1 (the convolution does not affect the value map) to 20. In most cases, the avgprec starts very low and rapidly increases to a maximum value, after which it slowly begins to decrease.

There are a few exceptions to this general rule. The plotted values that decrease

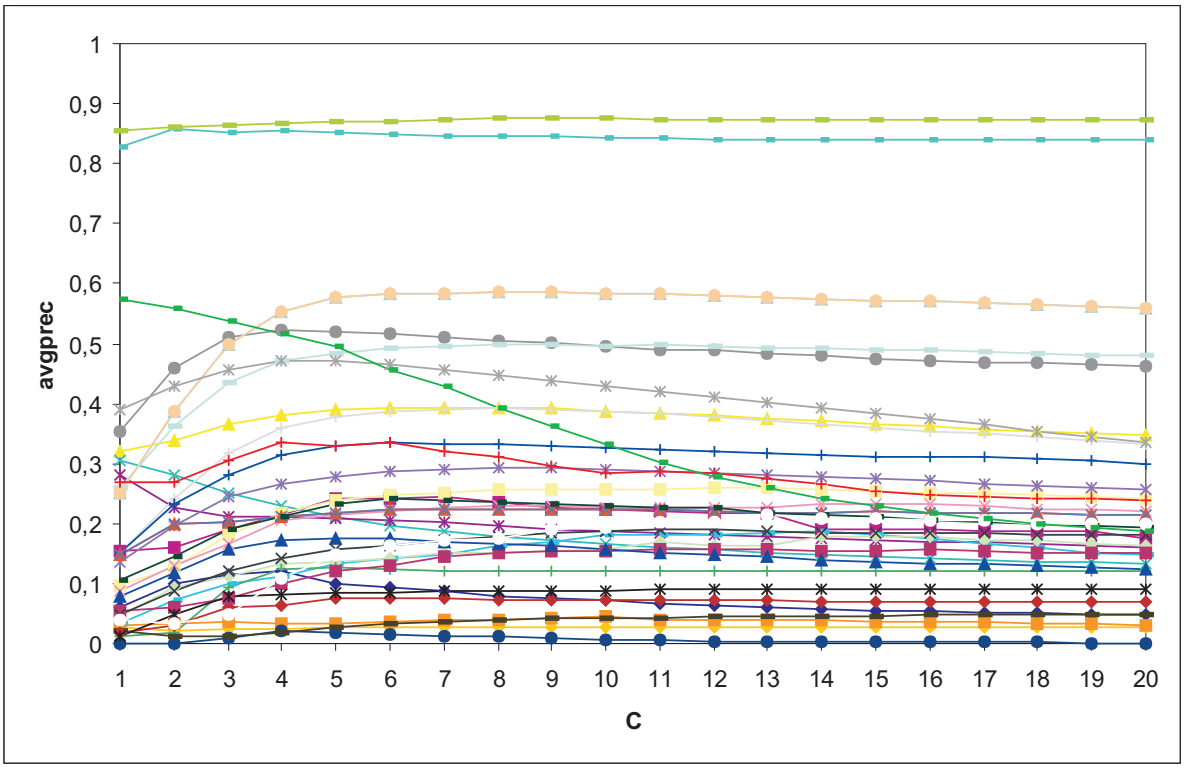


Figure 4: Dependence of average precision on convolution radius for 34 different classes.

monotonically with the increasing value of C belong to very small classes with very similar objects. These classes do not benefit from the generalisation that convolution produces. Also, there are two high-scoring classes that remain more or less constant. These are the classes, *face* and *person*, which are primarily dependent on a feature that registers successful face-detection. This produces a high degree of separability, which diminishes the influence of convolution.

When the optimal value of C is selected for each class, instead of the default value of 9, the avgprec was improved by 4.7%. However, when these optimised values of C were employed with subsets and supersets of the training data, the resulting avgprec was lower than with the default value. The optimisation of the convolution parameter was effectually over-learning. It was observed that when using a constant value of C for all classes, the subset experiment scored highest for higher values, while the superset did better with lower ones. It may be concluded from this that sparser data requires a greater degree of generalisation. Also, based on the general trend of most classes, it can be noted that higher values of C offer less risk, as they increase the chances of avoiding the steep rise prior to the peak.

4 Conclusions

In this paper we described in detail the difference between two variants of the TS-SOM algorithm used in the projects GalSOM and PicSOM. Methods used for optimising Gal-

SOM were applied to PicSOM with mixed success. MRC, which had previously been used successfully to improve the quality of GalSOM mappings was found to be unsuitable for PicSOM. Conversely, triangular neighbourhood kernels were found to improve PicSOM's results significantly. Finally, experiments with the convolution parameter of PicSOM showed that optimisation on a class-by-class basis for high-level feature detection results in over-learning. The analysis also suggests that higher values of the parameter will have more stable, albeit suboptimal results.

References

[1] J. Koh, M. Suk and S. M. Bhandarkar, *A Multilayer Self-Organizing Feature Map for Range Image Segmentation*, *Neural Networks*, 8:67-86, 1995.

[2] T. Kohonen, *Self-Organizing Maps*, Third Edition, Springer-Verlag Berlin 2001.

[3] P. Koikkalainen, E. Oja, *Self-organizing hierarchical feature maps*, Proceedings of *International Joint Conference on Neural Networks*, San Diego, CA, 1990; II:279-284.

[4] P. Koikkalainen, *Progress with the tree-structured self-organizing map*, *11th European Conference on Artificial Intelligence*, August 1994.

[5] J. Laaksonen, M. Koskela and E. Oja, *Application of Self-Organizing Maps in Content Based Image Retrieval*, in Proceedings of *ICANN'99*, Edinburgh, 1999.

[6] J. Laaksonen, M. Koskela, S. Laakso and E. Oja, *Self-Organizing Maps as a Relevance Feedback Technique in Content Based Image Retrieval*, *Pattern analysis & Applications*, 4(2-3): 140-152, June 2001.

[7] J. Pakkanen, J. Iivarinen and E. Oja, *The Evolving Tree - A Novel Self-Organizing Network for Data Analysis*, *Neural Processing Letters 20*, 2004.

[8] P. Prentis *Multi-Resolution Visualisation of Data with Self-Organizing Maps*, submitted to *Neural Network World* for publication, 2006.

[9] P. Prentis *Multi-Resolution Visualisation of Image Sets with Self-Organizing Maps*, in proceedings of *Doktorandský den Ústavu informatiky, Akademie věd České republiky*, 2006.

[10] P. Prentis *Problems with Multi-Resolution Visualisation of Data using Self-Organising Maps*, in proceedings of *Doktorandský den Katedry matematiky, Fakulta jaderná a fyzikálně inženýrská, ČVUT*, 2006.

[11] P. Prentis *GalSOM - Colour-Based Image Browsing and Retrieval with Tree-Structured Self-Organising Maps*, in Proceedings of the *6th Workshop on Self-Organizing Maps*, Bielefeld 2007.

[12] R. Rojas, *Neural Networks - A Systematic Introduction*, Springer-Verlag Berlin 1996.

Model spalování práškového uhlí v kotli*

Robert Straka†

3. ročník PGS, email: straka@kmlinux.fjfi.cvut.cz
Katedra matematiky, Fakulta jaderná a fyzikálně inženýrská, ČVUT
školitel: Michal Beneš, Katedra Matematiky, Fakulta Jaderná a Fyzikálně Inženýrská, ČVUT

Abstract. We describe behavior of the burning air-coal mixture in power plant furnace, using the Navier-Stokes equations for gas and particle phases, accompanied by a turbulence model. The undergoing chemical reactions are described by the Arrhenian kinetics (reaction rate proportional to $\exp\left(-\frac{E}{RT}\right)$, where T is temperature). We also consider the heat transfer via conduction and radiation. The system of PDEs is discretized using the finite volume method (FVM) and an advection upstream splitting method as the Riemann solver. The resulting ODEs are solved using the 4th-order Runge-Kutta method. Sample simulation results for typical power production levels are presented.

Abstrakt. Popisujeme chování hořící směsi práškovéhoo uhlí a vzduchu v elektrárenském kotli pomocí Navier-Stokesových rovnic pro plynnou a pevnou fázi spolu s modelem turbulence. Modelovaní chemických reakcí se popisuje tzv. Arrheniovskou kinetikou (rychlost reakce je uměrná $\exp\left(-\frac{E}{RT}\right)$, kde T je teplota). Dále uvažujeme přenos tepla vedením a radiací. Systém parciálních diferenciálních rovnic je diskretizovám metodami konečných objemů a "advection upstream splitting". Výsledný systém obyčejných diferenciálních rovnic je řešen metodou Runge-Kutty-Merson 4. řádu. Jsou prezentovány výsledky simulací pro typický výkon elektrárny.

1 Úvod

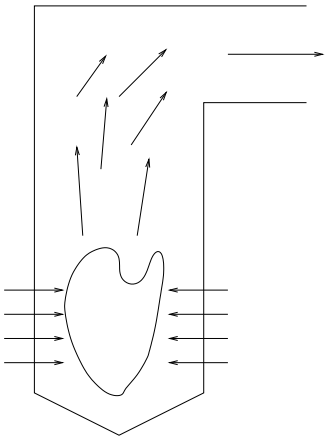
Hlavním cílem zkoumání hoření a vytvoření modelu hoření je jeho použití k vývoji řídících systémů pro industriální aplikace (stávající model používá parametry z kotle K5 v Otrokovicích). Dalším důvodem je optimalizace produkce oxidů dusíku, jejichž tvorba je silně ovlivněna teplotním profilem, přítomností spalovacího vzduchu a paliva, může být tedy kontrolována vhodnou distribucí uhelného prášku a sekundárního vzduchu na hořácích.

Spalovací komora je obvykle čtvercového průřezu s několika patry hořáků (viz Obrázek 1), které mohou být umístěny na stěnách nebo v rozích. V našem případě je komora

*Tato práce je částečně podpořena grantem "Applied Mathematics in Technical and Physical Sciences" MSM 6840770010 Ministerstva školství, mládeže a tělovýchovy České Republiky a grantem "Advanced Control and Optimization for Power Generation" číslo 1H-PK/22 Ministerstva průmyslu a obchodu České Republiky.

†Spolupracovali Jindřich Makovička, Michal Beneš and Vladimír Havlena

30m vysoká a 7 metrů široká, hořáky jsou ve čtyřech patrech. Rozměry spalovací komory a počet hořáků je dán energetickými požadavky . V našem případě výkon kotle činí 90 MW, spolu s parogenerátorem produkuje až 100 tun stlačené přehřáté páry za hodinu. Hořáky se obvykle nacházejí ve spodní části komory. Do nich je přiváděna směs primárního vzduchu a uhelného prášku, předehřátého na určitou teplotu, dále pak sekundárního vzduchu a nakonec nad posledním patrem je přiveden vzduch dohořívací. Jak směs hoří přenáší část tepla do pláště komory, který obsahuje trubky z vodou, zbytek tepla odchází spolu se spalinami do horní části kotle, jenž obsahuje různé kovekční plochy (ohříváky vzduchu a vody, výparník, přehřívák) a kde dochází k dalšímu přenosu tepla (viz například [4]). Největší snahou bylo namodelování procesů v oblasti, kde dochází k hoření a k tvorbě oxidů dusíku spolu s přesnějším fyzikálním modelem samotného hoření a jevů které ho doprovázejí.



Obrázek 1: Schéma kotle

2 Matematický model

Matematický model hoření je založen na Navier-Stokes rovnicích pro reakční směs, kde uhelné částečky jsou považovány za jednu z komponent (jinačí přístup je založen na vlastních rovnicích pro uhelné částice a spaliny [1]). Tento přístup byl zvolen neboť zjednodušuje model v případě uvážení turbulence a odstraňuje několik empirických vztahů a konstant.

Stávající model zahrnuje nasledující komponenty směsi:

- chemické látky učastnící se při tvorbě oxidů dusíku: dusík (N_2), kyslík (O_2), oxid dusnatý (NO), kyanovodík (HCN), amoniak (NH_3) a voda (H_2O)

- pevná (char) a těkavá (volatile) část uhelné částice

Plynná fáze je popsána následujícími rovnicemi.
Zákon zachování hmoty pro každou komponentu (je použito Einsteinovo sumační pravidlo):

$$\frac{\partial}{\partial t}(\rho Y_i) + \frac{\partial}{\partial x_j}(\rho Y_i u_j) = \nabla \vec{J_i} + R_i, \tag{1}$$

kde ρ je hustota spalin, Y_i hmotnostní koncentrace komponenty a u_j jsou složky rychlosti spalin. Členy na pravé straně popisují laminarní a turbulentní difuzi a tvorbu či zánik komponenty v chemických reakcích.

Výše uvedené rovnice jsou doplněny rovnicí kontiniuty:

$$\frac{\partial \rho}{\partial t} + \frac{\partial(\rho u_j)}{\partial x_j} = 0. \tag{2}$$

Rovnicí zachování momentu hybnosti:

$$\frac{\partial}{\partial t}(\rho u_i) + \frac{\partial}{\partial x_j}(\rho u_i u_j) = -\frac{\partial p}{\partial x_i} + \frac{\partial}{\partial x_j}\left[\mu_{\text{eff}}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} - \frac{2}{3}\delta_{ij}\frac{\partial u_l}{x_l}\right)\right] + g_i, \tag{3}$$

kde $\vec{g} = [g_1, g_2, g_3]$ je působení vnější síly na tekutinu, v našem případě gravitace. Efektivní koeficient dynamické viskozity μ_{eff} je pomocí modelu turbulence vyjádřen jako:

$$\mu_{\text{eff}} = \mu + \mu_t = \mu + \rho C_\mu \frac{k^2}{\epsilon},$$

kde μ je laminarní dynamická viskozita, k turbulentní kinetická energie a ϵ rychlost disipace k. Konstanty jako C_μ, a další aditivní konstanty zmíněné později v textu, musí být určeny empiricky pro daný problém, v našem případě užíváme $C_\mu = 0.09$, což dává uspokojivé výsledky. Všechny empirické konstanty pro model turbulence jsou převzaty z [13].

Poslední rovnice popisuje zachování energie:

$$\frac{\partial}{\partial t}(\rho h) + \frac{\partial}{\partial x_j}(\rho u_j h) = -n_{\text{coal}}\frac{\mathrm{d}m_{\text{coal}}}{\mathrm{d}t}h_{\text{comb}} + q_r + q_c + q_s, \tag{4}$$

kde členy na pravé straně jsou po řadě, spalné teplo, přenos tepla radiací, vedením a člen popisující vznik či zánik tepla. Členy jsou modelovány následovně:

$$-q_c = \nabla \cdot (\lambda \nabla T)\,,$$

přenos tepla vedením, který je popsán Fourierovým zákonem vedením tepla a

$$-q_r = \nabla \cdot \left(cT^3 \nabla T\right),$$

pro přenos tepla sáláním (radiací). Přenos tepla radiací je plně popsán integro-diferenciálními rovnicemi, které je velmi výpočetně nákladné řešit. Nicméně, spaliny lze považovat za opticky hustý materiál a lze aplikovat přibližný model radiace, tzv. Rosselandův model [13].

Člen popisující zánik tepla je nenulový jen na hranici komory a popisuje výměnu tepla se zdmi komory vedením a sáláním

$$q_s = A(T_{\text{gas}} - T_{\text{wall}}) + B(T_{\text{gas}}^4 - T_{\text{wall}}^4),$$

kde A a B jsou konstanty zavisející na vlastnostech rozhraní mezi modelovanou oblastí a jejím okolí.

Hmotnostní změna uhelných částic je popsána jednokrokovou Arheniovskou kinetikou, zvlášť pro pevnou a těkavou část – těkavá složka hoří mnohem rychleji nez pevná.

$$\frac{\mathrm{d}m_p}{\mathrm{d}t} = -A_v m_p^\alpha [\mathrm{O}_2]^\beta \exp\left(-\frac{E_v}{RT_p}\right),$$

kde m_p je hmotnost hořlaviny, A_v, E_v jsou empirické konstanty, $[O_2]$ koncentrace kyslíku a T_p je teplota částice.

Výše uvedené rovnice jsou doplněny stavovou rovnicí

$$p = (\kappa - 1)\rho_{\text{gas}} \left(e_{\text{gas}} - \frac{1}{2} v_{\text{gas}}^2 \right).$$

Zde, κ je Poissonova konstatna a e_{gas} je celková energie na jednotku hmotnosti.

Model turbulence je standartní k-ϵ model, který popisuje vývoj turbulence pomocí dvou rovnic — rovnice turbulentní kinetické energie

$$\frac{\partial}{\partial t}(\rho k) + \frac{\partial}{\partial x_j}(\rho k u_j) = \frac{\partial}{\partial x_j}\left[\left(\mu + \frac{\mu_t}{\sigma_k} \right) \frac{\partial k}{\partial x_j} \right] + G_k - \rho\epsilon, \tag{5}$$

a rovnice rychlosti disipace turbulentní kinetické energie

$$\frac{\partial}{\partial t}(\rho \epsilon) + \frac{\partial}{\partial x_j}(\rho \epsilon u_j) = \frac{\partial}{\partial x_j}\left[\left(\mu + \frac{\mu_t}{\sigma_\epsilon} \right) \frac{\partial \epsilon}{\partial x_j} \right] + C_{1\epsilon} \frac{\epsilon}{k} G_k - C_{2\epsilon} \rho \frac{\epsilon^2}{k}. \tag{6}$$

Konstanty musí byt zjištěny empiricky a v našem případě užíváme: $C_{1\epsilon} = 1.44$, $C_{2\epsilon} = 1.92$, $\sigma_k = 1.0$, $\sigma_\epsilon = 1.3$.

Produkci turbulence G_k, lze odvodit z Reynoldsova procesu zprůměrování a zapsán ve tvaru fluktujících částí rychlostí nabývá tvaru

$$G_k = \tau_{jl} \frac{\partial u_j}{\partial x_l} = -\rho \overline{u_j' u_l'} \frac{\partial u_j}{\partial x_l},$$

kde τ_{jl} je Reynoldsův tenzor napětí. Avšak fluktuace u_j' a u_l' jsou během výpočtu neznámé. Po použití hypotézy Boussinesqa, že Reynoldsovo napětí je úměrné střední rychlosti deformace

$$S_{ij} = \frac{1}{2} \left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right),$$

lze zapsat produkci v uzavřené formě

$$G_k = \mu_t S^2, \quad S = (2 S_{jl} S_{jl})^{1/2}.$$

Difuze látek sestává ze dvou procesů — laminárních a turbulentních a difuzní člen v rovnici (1) můžeme napsat jako

$$\vec{J}_i = -\left(\rho D_{i,m} + \frac{\mu_t}{\text{Sc}_t} \right) \nabla Y_i.$$

První člen je laminární difuze, druhý turbulentní. Jelikož turbulentní člen obecně převažuje nad laminárním a $D_{i,m}$ je obtížné zjistit, je laminární člen vypuštěn. Sc_t je turbulentní Schmidtovo číslo a pokládáme $\text{Sc}_t = 0.7$.

K popisu pevné fáze částic (zvláště pak celkovému povrchu částic) potřebujeme ještě popsat početní hustotu částic:

$$\frac{\partial n_{\text{coal}}}{\partial t} + \frac{\partial (n_{\text{coal}} u_{\text{coal}})}{\partial x_1} + \frac{\partial (n_{\text{coal}} v_{\text{coal}})}{\partial x_2} = 0. \tag{7}$$

3 Zjednodušený model tvorby NOx

Tento model byl vypracován tak aby přibližně popisoval množství oxidů dusíku při spalování uhlí v kotli. Přesný mechanismus hoření uhlí je velikce komplikovaný a obsahuje spousty chemických látek a rovnic, proto je použit zjednodušený model, který postihuje jen nejdůležitejší reakce a látky jenž mají vliv na konečnou koncentraci oxdiů dusíku ve spalinách.

V naprosté většině případů značením NOx rozumíme skupinu oxidu dusnatého (NO) a oxidu dusičitého (NO_2). Tyto plyny silně znečišťují životní prostředí a podílejí se na vzniku kyselých dešťů. Jsou známy dvě hlavní cesty jak mohou při spalování NOxy vznikat. První jsou známy pod názvem *Termální NOxy* nebo *Zeldovičovy NOxy* a vznikají oxidací atmosférického dusíku při vysokých teplotách. Druhé jsou nazývány *Palivové NOxy*, povstávající při spalování paliva, které samo o sobě obsahuje dusík vázaný v palivu. Pokud se daří držet teploty v kotli na hladinách nižších než jsou potřebné pro zahájení reakcí, při kterých vznikají Termální NOxy, jsou Palivové NOxy hlavním zdrojem imisí oxidu dusíků.

Jsou známy i další mechanismy vzniku NOxů (*Promptní NOx (Fenimore)* nebo mechanismus s oxidem dusným jako meziproduktem). Příspevek těchto mechanizmů je v případě spalování uhlí za normálních podmínek zanedbatelný a nebyl v modelu uvažován.

3.1 Termální NO

Mechanismus vzniku termálních NOxů funguje pouze při vysokých teplotách (okolo 1800K) a je popsán třemi reakčními rovnicemi poprvé zveřejněnými Zeldovičem [5] a rozšířených Bowmanem [6]

$$\begin{aligned} \mathrm{O} + \mathrm{N_2} &\overset{k_1}{\longleftrightarrow} \mathrm{N} + \mathrm{NO} \\ \mathrm{N} + \mathrm{O_2} &\overset{k_2}{\longleftrightarrow} \mathrm{O} + \mathrm{NO} \\ \mathrm{N} + \mathrm{OH} &\overset{k_3}{\longleftrightarrow} \mathrm{H} + \mathrm{NO} \end{aligned}$$

Tyto rovnice mohou běžet oběma směry a jejich rychlostní konstanty jsou z databáze [7].

K výpočtu koncentrace NO musíme znát koncentraci radikálů O, OH a N o kterém pro zjednodušení předpokládáme, že je ve kvazistabilním stavu. Ve skutečnosti je formace N limitujícím faktorem produkce termálních NOxů, díky velmi vysoké aktivační energii molekuly dusíku (za vše může trojná vazba mezi atomy dusíku). Tedy změna koncentrace NO je popsána rovnicí

$$\frac{\mathrm{d}[\mathrm{NO}]}{\mathrm{dt}} = 2k_1^+ \cdot [\mathrm{O}] \cdot [\mathrm{N_2}] \cdot \frac{1 - \frac{k_1^- k_2^- [\mathrm{NO}]^2}{k_1^+ [\mathrm{N_2}] k_2^+ [\mathrm{O_2}]}}{1 + \frac{k_1^- \cdot [\mathrm{NO}]}{k_2^+ [\mathrm{O_2}] + k_3^+ [\mathrm{OH}]}}.$$

Za jistých podmínek se molekula kyslíku štěpí a rekombinuje cyklicky.

$$\mathrm{O_2} \overset{K_1}{\longleftrightarrow} \mathrm{O} + \mathrm{O},$$

což vyjadřuje následující stav částečné rovnováhy

$$[\mathrm{O}] \;=\; K_1 \cdot [\mathrm{O_2}]^{1/2} \cdot T^{1/2}.$$

Pro radikál OH máme podobnou částečnou rovnováhu

$$\mathrm{O} + \mathrm{H_2O} \;\overset{K_2}{\longleftrightarrow}\; \mathrm{OH} + \mathrm{OH}$$

s přiblížením

$$[\mathrm{OH}] \;=\; K_2 \cdot [\mathrm{O}]^{1/2} \cdot [\mathrm{H_2O}]^{1/2} \cdot T^{-0.57}.$$

Rovnovážné konstanty K_1 a K_2 jsou

$$\begin{aligned} K_1 &= 36.64 \cdot \exp\left(\frac{-27123}{T}\right), \\ K_2 &= 2.129 \cdot 10^2 \cdot \exp\left(\frac{-4595}{T}\right). \end{aligned}$$

3.2 Palivový NO

Analýza složení uhlí ukazuje, že dusíkaté látky jsou v něm více či méně zastoupeny, obvykle až do desítkek hmotnostních procent. Pokud je uhlí zahříváno, tyto látky přecházejí v jisté meziprodukty a následně na NO. Samo uhlí je tedy významným zdrojem oxidů dusíku. Pokud je částice uhlí zahřívána, předpokládáme, že dusíkaté sloučeniny se rozdělí mezi pevnou a těkavou část. Mnoho prací bez opodstatnění uvádí, že polovina vázaného dusíku přejde do pevné části a polovina do těkavé. Jelikož k tomu není žádný důvod, zavádíme parametr α, který popisuje distribuci vázaného dusíku mezi pevnou a těkavou část.

$$\begin{aligned} m^{\mathrm{N}}_{\mathrm{vol}} &= \alpha \cdot m^{\mathrm{N}}_{\mathrm{tot}}, \\ m^{\mathrm{N}}_{\mathrm{char}} &= (1-\alpha) \cdot m^{\mathrm{N}}_{\mathrm{tot}}, \end{aligned}$$

kde $\alpha \in < 0, 1 >$, $m^{\mathrm{N}}_{\mathrm{tot}}$ je celkové množství vázaného dusíku, $m^{\mathrm{N}}_{\mathrm{vol}}$ je množství dusíku v těkavé části a $m^{\mathrm{N}}_{\mathrm{char}}$ v pevné části.

Jak již bylo zmíněno, dusík přechází v imise skrze meziprodukty, kterými jsou obvykle amoniak $\mathrm{NH_3}$ a kyanovodík HCN.

Dále je třeba rozlišit čtyři různé reakční cesty (viz [9, 10]). Aby proces tvorby imisí mohl být co nejvíce komplexní je nutné zavést tři další parametry (podobné α):

- β je distribuce $m^{\mathrm{N}}_{\mathrm{tot}}$ mezi HCN a $\mathrm{NH_3}$.
- γ je distribuce $m^{\mathrm{N}}_{\mathrm{HCN}}$ mezi první a druhou reakční cestou.
- δ je distribuce $m^{\mathrm{N}}_{\mathrm{NH_3}}$ mezi třetí a čtvrtou reakční cestou.
- $\beta,\ \gamma,\ \delta \in \langle 0, 1 \rangle$.

Například, množství dusíku v pevné části vstupující do druhé reakční cesty, můžeme napsat jako

$$m^{\mathrm{N}}_{\mathrm{P2,char}} \;=\; m^{\mathrm{N}}_{\mathrm{tot}} \cdot \beta \cdot (1-\gamma) \cdot (1-\alpha).$$

Pět obecných reakcí popisujících vznik anebo zánik NO bylo použito při modelování spalování.

3.2.1 NO, HCN, NH_3 reakce

Vzhledem k [11], rychlostní konstanty reakcí

$$\begin{aligned} \mathrm{HCN} + \mathrm{O_2} &\xrightarrow{R_1} \mathrm{NO} + \ \ldots \\ \mathrm{NH_3} + \mathrm{O_2} &\xrightarrow{R_2} \mathrm{NO} + \ \ldots \\ \mathrm{HCN} + \mathrm{NO} &\xrightarrow{R_3} \mathrm{N_2} + \ \ldots \\ \mathrm{NH_3} + \mathrm{NO} &\xrightarrow{R_4} \mathrm{N_2} + \ \ldots \end{aligned}$$

jsou dány rovnicemi

$$\begin{aligned} R_1 &= 1.0 \cdot 10^{10} \cdot X_{\mathrm{HCN}} \cdot X_{\mathrm{O_2}}^{a} \cdot \exp\left(\frac{-33732.5}{T}\right), \\ R_2 &= 4.0 \cdot 10^{6} \cdot X_{\mathrm{NH_3}} \cdot X_{\mathrm{O_2}}^{a} \cdot \exp\left(\frac{-16111.0}{T}\right), \\ R_3 &= -3.0 \cdot 10^{12} \cdot X_{\mathrm{HCN}} \cdot X_{\mathrm{NO}} \cdot \exp\left(\frac{-30208.2}{T}\right), \\ R_4 &= -1.8 \cdot 10^{8} \cdot X_{\mathrm{NH_3}} \cdot X_{\mathrm{NO}} \cdot \exp\left(\frac{-13593.7}{T}\right), \end{aligned}$$

kde X je molární zlomek a a řád reakce kyslíku.

3.2.2 Heterogení redukce NO na pevné části

Na pevné části dochází k následujícímu adsorbčnímu procesu

$$\mathrm{Char} + \mathrm{NO} \xrightarrow{R_5} \mathrm{N_2} + \ \ldots$$

Levy [12] užil povrch pórů v pevné části (BET) k vyjádření zániku NO

$$S_{\mathrm{ads}}^{\mathrm{NO}} \ = \ k_5 \cdot c_s \cdot A_{\mathrm{BET}} \cdot M_{\mathrm{NO}} \cdot p_{\mathrm{NO}},$$

kde $k_5 = 2.27 \cdot 10^{-3} \cdot \exp\left(\frac{-17168.33}{T}\right)$ je rychlostní konstanta, $S_{\mathrm{ads}}^{\mathrm{NO}}$ je zdrojový člen NO, c_s je koncentrace částic, A_{BET} je plocha pórů a p_{NO} je parciální tlak NO.

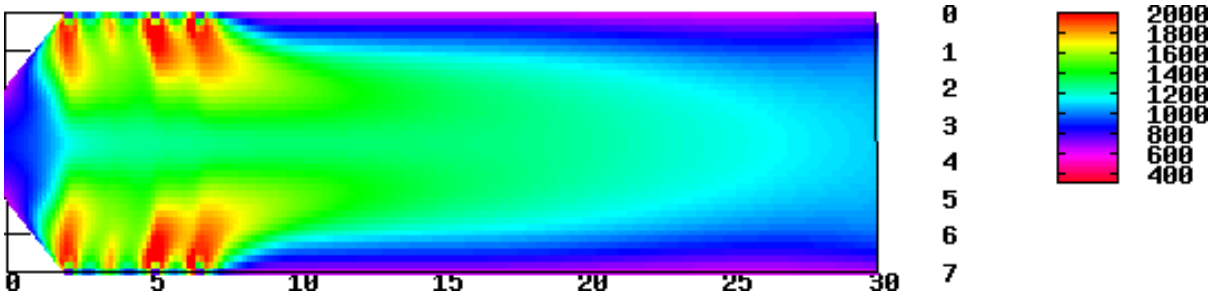
Celkový zdrojový člen pro NO, který vznikne sumací všech zdrojových členů uvedených výše, je posléze použit v transportní rovnici, stejně tak pro ostatní látky. Dále je předpokládáno, že dusík jak z těkavé tak pevné části přechází na meziprodukty rychle a bezezbytku.

4 Numerické řešení

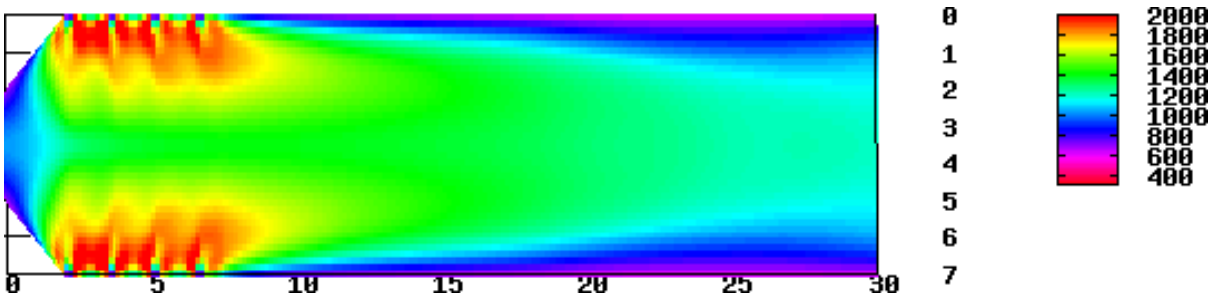
K numerickému řešení rovnic byla použita metoda konečných objemů. Na levé a pravé strany v rovnicích (1), (2), (3), (4), (5), (6), (7), byla aplikována metoda „advection upstream splitting“ (viz [2]) k aproximaci toků a metoda duálních objemů k aproximaci druhých derivací. Detailní popis řešiče lze nalézt v [4].

5 Výsledky simulací

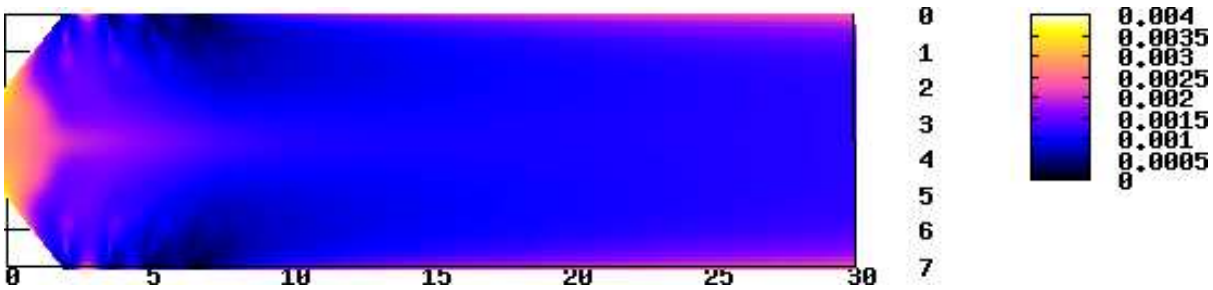
Nyní jsou uvedeny výsledky simulací spalovaní v kotli, spolu s porovnáním dvou modelů. Hlavní rozdíl mezi prvním a druhým modelem je vylepšení fyzikální části(termofyzikální konstanty nahrazeny funkcemi) a důkladnější model turbulence (přidány některé další členy, obvykle zanedbávané). Hlavní rozdíl je ovšem ve výpočtu celé spalovací komory (druhý model) a ne jen její poloviny (první model), což se projeví v symetrii řešení. Teplotní profily a výstupní koncentrace oxidu dusíku se víceméně shodují z hodnotami naměřenými v otrokovické teplárně. Ve všech případech je v provozu 8 hořáků, tj. 4 na každé stěně kotle.



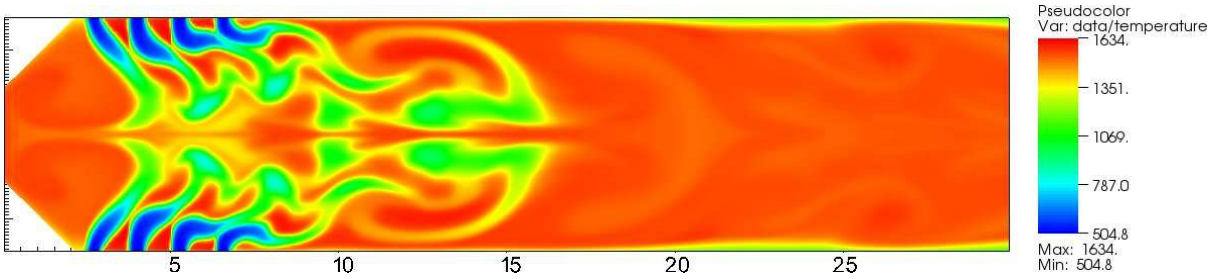
Obrázek 2: Teplotní profil — tok spalin: 18 kg/s, první model



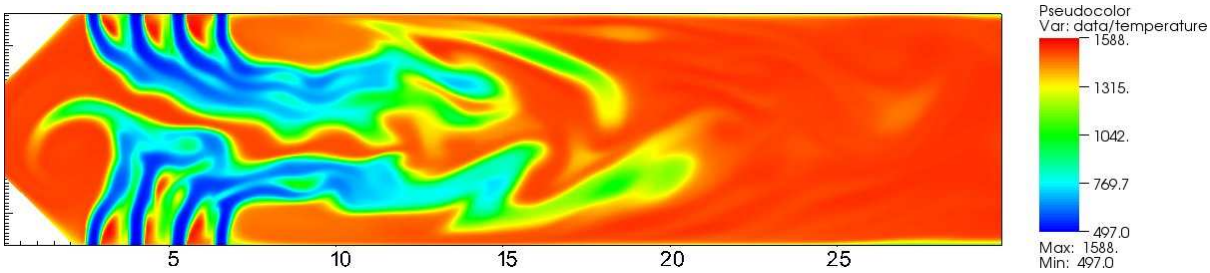
Obrázek 3: Teplotní profil — tok spalin: 28 kg/s, první model



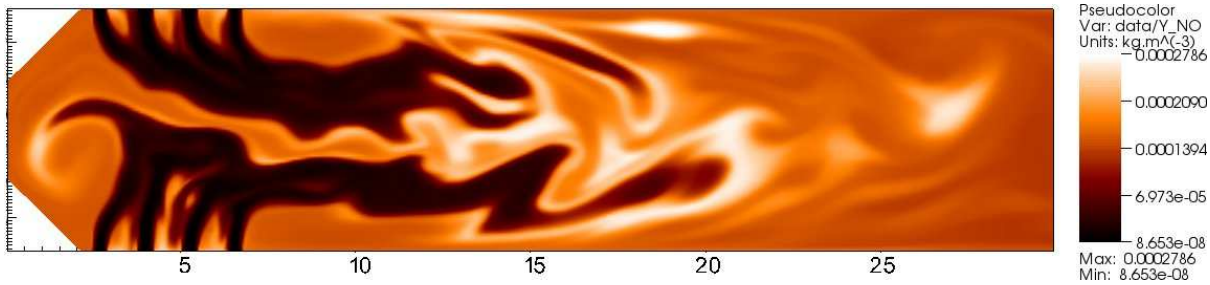
Obrázek 4: NOx profil — tok spalin: 48 kg/s, první model



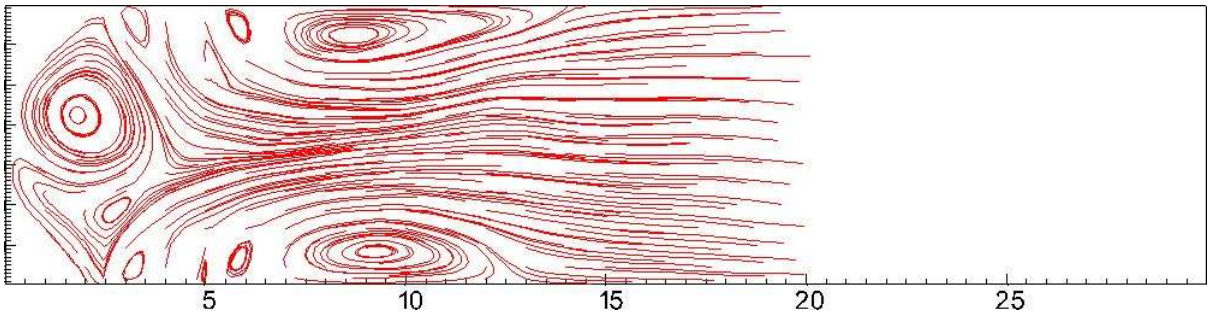
Obrázek 5: Teplotní profil — množství paliva: 5.63 kgf/s, druhý model



Obrázek 6: Teplotní profil — množství paliva : 5.63 kg/s, druhý model



Obrázek 7: NOx profil — množství paliva : 5.63 kg/s, druhý model



Obrázek 8: Proudnice — množství paliva : 5.63 kg/s, druhý model

6 Shrnutí

Byl vytvořen matematický model, který aproximuje složité procesy hoření v průmyslovém kotli. Tento model se ukázal býti přijatelný jak z hlediska výpočetního tak fyzikálního. Jako další, bude snaha použití stávajícího modelu k výpočtu mnohem komplikovanějšího fluidního kotle s dvojím druhem paliva a cirkulací materiálu.

Poděkování

Velmi oceňujeme pomoc od projektu HPC-Europa, a vyjímečnou ochotu pracovníků z superpočítačového centra CINECA z italské Boloni, kde byla dokončena práce na paralelní verzi řešiče. Tato práce na paralelní verzi byla částí projektu HPC-EUROPA (RII3-CT-2003-506079), s podporou European Community — Research Infrastructure Action a programem FP6 "Structuring the European Research Area".

Literatura

[1] Y. C. Guo, C. K. Chan, *Fuel* 79 (12) (2000) 1467–1476.

[2] Meng-Sing Liou, C. Steffen, Jr., *J. Comp. Phys.* 107 (1) (1993) 23–29.

[3] J. Makovička, V. Havlena, in:M. Beneš, J. Mikyška, T. Oberhuber (Eds.), *Proceedings of the Czech-Japanese Seminar in Applied Mathematics 2004*, Faculty of Nuclear Sciences and Physical Engineering, Czech Technical University in Prague, 2005, p. 106.

[4] J. Makovička, V. Havlena, M. Beneš, in: A. Handlovičová, Z. Krivá, K. Mikula, D. Ševčovič (Eds.), *ALGORITMY 2002 Proceedings of contributed papers*, Publ. house of STU, 2002, p. 171.

[5] J. B. Zeldovich, *Acta Physicochimica* 21 (1946) 577–628.

[6] C. T. Bowman, D. J. Seery, *Emissions from Continuous Combustion Systems*, Plenum Press, New York, 1972, p. 123.

[7] NIST, *Chemical Kinetics Database on the Web*, National Institute of Standards and Technology, 2000, `http://www.kinetics.nist.gov`

[8] C. Kim, N. Lior, *Chemical Engineering Journal* 71 (3) (1998) 221–231.

[9] L. D. Smoot, P. J. Smith, *Coal Combustion and Gasification*, Plenum Press, New York, 1985, p. 373.

[10] F. C. Lockwood, C. A. Romo-Millares, *J. Inst. Energy* 65 (1992) 144–152.

[11] G. G. De Soete, *Proc. Combust. Inst.* 15 (1975) 1093–1102.

[12] J. M. Levy, L. K. Chen, A. F. Sarofim, J. M. Beer, *Proc. Combust. Inst.* 18 (1981) 111–120.

[13] FLUENT Inc., *FLUENT user's guide*, 2005.

[14] L. X. Zhou, Y. Zhang, J. Zhang, *Simulation of swirling coal combustion using a full two-fluid model and an AUSM turbulence-chemistry model*, FUEL 82, (2003), 1001–1007.

[15] Makovička, J., *Numerical Model of Turbulent Coal Combustion*, PhD thesis, KM FJFI, in preparation.

Confidence of Classification and its Application to Classifier Aggregation*

David Štefka

2nd year of PGS, email: stefka@cs.cas.cz
Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU
advisor: Martin Holeňa, Institute of Computer Science, AS CR

Abstract. Classifier combining is a succesful method for improving the quality of classification. In this paper, we introduce the concept of confidence of classification and define two confidence measures – the local accuracy and the local diversity. We propose algorithms for classifier aggregation which utilize the concept of confidence. We then compare the performance of these algorithms with several state-of-the-art classifier combining techniques. Results of these benchmark tests show that by incorporating confidence into classifier aggregation algorithms, the state-of-the-art methods can be improved.

Abstrakt. Kombinování klasifikátorů je úspěšná metoda pro zvýšení kvality klasifikace. V tomto článku zavedeme koncept spolehlivosti klasifikace a definujeme dvě míry spolehlivosti klasifikace – lokální přesnost a lokální diverzitu. Vytvoříme algoritmy pro agregaci klasifikátorů, které využívají koncept spolehlivosti klasifikace. Poté porovnáme tyto algoritmy s několika standardně používanými přístupy ke kombinování klasifikátorů. Výsledky těchto testů ukazují, že standardní algoritmy pro agregaci klasifikátorů se dají zavedením spolehlivosti klasifikace vylepšit.

1 Introduction

Classifier combining is a succesful method for improving the quality of classification, based on using many classifiers and combining their outputs, instead of using just one classifier. The literature shows that a team of multiple classifiers can perform the classification task better than any of the individual classifiers. However, to achieve this, the classifier outputs have to be combined wisely. For this purpose, many methods have been introduced in the literature. These can be grouped into classifier selection and classifier aggregation.

In classifier selection, some rule is used to determine which classifier to use for the current pattern; only this "expert" classifier is then used for the final prediction, and the rest of the team is discarded. In classifier aggregation, outputs of all the classifiers are aggregated into the final decision.

Common drawback of classifier aggregation methods is that they are *global*, i.e., they do not adapt themselves to the particular patterns to classify. In other words, the combination is specified during a training phase, prior to classifying a test pattern. A typical example is that if we use the weighted mean aggregation rule, the weights of the individual classifiers are usually based on the classifiers' accuracies. Although it is true that

*The research reported in this paper was partially supported by the Program "Information Society" under project 1ET100300517.

if a classifier has high accuracy, its weight should be higher, still, for the *curent pattern*, some other classifier could be more suitable.

While classifier selection methods use some techniques to determine which classifier is *locally* better than the others, such algorithms select only one classifier, discarding much potentialy useful information, thus reducing the robustness compared to classifier aggregation.

In this paper, we try to incorporate the strong points of classifier selection techniques into classifier aggregation methods. This will enable us to create novel methods for classifier aggregation, which can provide better results than state-of-the-art methods for classifier combining.

We introduce the concept of *confidence* of classification, which can be used both as a criterium for classifier selection, and for improving classifier aggregation. We define two confidence measures, and propose algorithms for classifier aggregation which utilize the concept of confidence. We then show that these algorithms outperform commonly used methods for classifier combining on three benchmark datasets.

The paper is structured as follows: Section 2 deals with basic aspects of classifier combining, namely Section 2.1 contains reference to the literature about ensemble methods, Sections 2.2 and 2.3 describe the differences between classifier selection and classifier aggregation. Section 3 introduces the concept of confidence of classification. Section 4 contains experiments – Section 4.1 describes algorithms used in the experiments, and in Section 4.2, the results of the experiments are discussed. Finally, Section 5 then summarizes the paper.

2 Classifier Combining

Throughout the rest of the paper, we use the following notation. Let $\mathcal{X} \subseteq \mathbf{R}^n$ be a n-dimensional *feature space*, an element $\vec{x} \in \mathcal{X}$ of this space is called *pattern*, and let $C_1, \ldots, C_N \subseteq \mathcal{X}$ be disjoint sets called *classes*. The goal of classification is to determine to which class a given pattern belongs. We call a *classifier* any mapping ϕ from the following:

- *possibilistic classifier* – $\phi : \mathcal{X} \to [0,1]^N$, where $\phi(\vec{x}) = (\mu_1, \ldots, \mu_N)$ are *degrees of classification* to each class.
- *normalized possibilistic classifier* – $\phi : \mathcal{X} \to [0,1]^N$, where $\phi(\vec{x}) = (\mu_1, \ldots, \mu_N)$, $\sum_i \mu_i = 1$.
- *crisp classifier* – $\phi : \mathcal{X} \to \{1, \ldots, N\}$, where $\phi(\vec{x})$ is the predicted class label of pattern $\vec{x}$. Crisp classifier can also be defined as a special case of a normalized possibilistic classifier, such that one degree of membership is equal to 1, and the others are equal to 0.

Normalized possibilistic classifiers are sometimes called *probabilistic* [11]. However, they do not need to be based on probablility theory, so we will call them normalized possibilistic. Other types of classifiers, such as *rank classifier* [13], can be defined, but we deal with crisp and possibilistic classifiers only in the rest of the paper. The conversion

of a possibilistic classifier ϕ_p to a crisp classifier ϕ_c is called *hardening*:

$$\phi_c(\vec{x}) = \arg\max_{i=1,\ldots,N}\{\mu_i\}, \tag{1}$$

where $\phi_p(\vec{x}) = (\mu_1, \ldots, \mu_N)$.

In classifier combining, we create a team of classifiers, let each of the classifiers predict independently, and then combine the classifiers' outputs into one final prediction. This combined classifier can perform its classification task better than any of the individual classifiers in the team. Methods which use more or less this idea can be found under many names in the literature – *classifier combining*, *classifier aggregation*, *classifier fusion*, *classifier selection*, *mixture of experts*, *classifier ensembles*, etc. Basically, there are two main approaches to classifier combining:

- *classifier selection*, where we use some rule to determine which classifier to use for the current pattern; only this "expert" classifier is then used for the final prediction
- *classifier aggregation*, where all the classifiers in the team are used for the final decision

Classifier combining consists of two steps – first, we create a team of classifiers, and then we adopt some strategy to combine the classifiers' outputs into the final decision. The former step is common for both classifier selection and aggregation (algorithms for creating a team of classifiers are descibed in Sec. 2.1), while for the latter step, different algorithms are needed (these are described in Sec. 2.2 and 2.3).

2.1 Ensemble Methods

If the team of classifiers consists only of classifiers of the same type, which differ only in their parameters, dimensionality, or training sets, the team is usually called an *ensemble* of classifiers. That is why the methods which create a team of classifiers are sometimes called *ensemble methods*. The restriction to classifiers of the same type is not essential, but it ensures that the outputs of the classifiers are consistent.

Well-known methods for ensemble creation are *bagging* [4], *boosting* [7], error correction codes [10], or *multiple feature subset* (MFS) methods [3]. These methods try to create an ensemble of classifiers which are both *accurate* and *diverse*.

2.2 Classifier Selection

Crisp classifiers are not much appropriate for classifier combining, because they do not provide information about degree of classification to each class. For these classifiers, only simple techniques like voting or single best selection can be used. That's the reason why we restrict to possibilistic classifiers in this paper. In the rest of the paper, we suppose that we have constructed an ensemble $(\phi_1, \ldots, \phi_r)$ of r possibilistic classifiers using some of the methods described in Sec. 2.1.

Classifier selection algorithms [15, 2, 14] use some criterion to determine which classifier is most suitable for the current pattern, and the output of this classifier is taken as the final result. The criterion for selection can be some global property of the ensemble, as in *single best selection* (SBS), or some local property, as in *dynamic best selection* (DBS).

In SBS, the criterion for selection is usually the validation error rate of the individual classifiers. The classifier with the lowest validation error rate is used for prediction of all the patterns (i.e. the other classifiers are entirely discarded). In DBS, the classifier optimizing some local criterion (for example local accuracy of the classifier in neighborhood of the current pattern) is selected for the prediction.

2.3 Classifier Aggregation

For classifier aggregation, the output of the ensemble $(\phi_1, \ldots, \phi_r)$ for input pattern $\vec{x}$ can be structured to a $r \times N$ matrix, called *decision profile* (DP):

$$DP(\vec{x}) = \begin{pmatrix} \phi_1(\vec{x}) \\ \phi_2(\vec{x}) \\ \vdots \\ \phi_r(\vec{x}) \end{pmatrix} = \begin{pmatrix} \mu_{1,1} & \mu_{1,2} & \cdots & \mu_{1,N} \\ \mu_{2,1} & \mu_{2,2} & \cdots & \mu_{2,N} \\ & & \ddots & \\ \mu_{r,1} & \mu_{r,2} & \cdots & \mu_{r,N} \end{pmatrix} \tag{2}$$

The $i-$th row of $DP(\vec{x})$ is an output of the corresponding classifier ϕ_i, and the $j-$th column contains the degrees of classification of $\vec{x}$ to the corresponding class C_j given by all the classifiers.

Many methods for aggregating the ensemble of classifiers into one final classifier have been reported in the literature. A good overview of the commonly used aggregation methods can be found in [11]. These methods comprise simple arithmetic rules (sum, product, maximum, minimum, average, weighted average, see [11, 8]), fuzzy integral [11, 9], Dempster-Shafer fusion [11, 1], second-level classifiers [11], decision templates [11], and many others.

In this paper, we introduce the concept of *confidence* of classification, which can be used both as a criterion for classifier selection, and for improving classifier aggregation. The concept of confidence is described in the next section.

3 Confidence Classifiers

The classifiers defined in Sec. 2 (both crisp and possibilistic) give us information about the *evidence* of classification (i.e., degrees of classification) of the current pattern $\vec{x}$. This is all we need to know if we are classifying patterns using a single classifier. However, in classifier combining, we have a team of classifiers, and the information about "how can we trust the output of classifier ϕ_i" could be very useful. For this purpose, we introduce a concept of *confidence* of classification.

The concept of confidence is not new to classifier combining – in classifier selection, the criteria for selection can be viewed as some confidence measures. In weighted mean classifier aggregation, the individual classifiers' error rates (which can again be viewed as some confidence measure) are used to adapt the weights of the individual classifiers etc. In this paper, we try to generalize different methods which use this approach, and incorporate all of them into the concept of confidence. This enables us to create general algorithms for classifier aggregation, which use some properties of classifier selection, improving both classifier aggregation and classifier selection. This is what makes the approach novel.

Suppose we have a classifier ϕ, and a pattern $\vec{x}$ to classify. The confidence of classification of the pattern $\vec{x}$ using classifier ϕ is a real number in the unit interval $[0, 1]$, and we model it by a mapping $\kappa_\phi : \mathcal{X} \to [0, 1]$. The mapping κ_ϕ will be called *confidence measure*, and the tuple (ϕ, κ_ϕ) will be called *confidence classifier*.

The confidence of classification $\kappa_\phi(\vec{x})$ can be any property estimating the degree to which we can trust the output of ϕ for current pattern $\vec{x}$. In this paper, we will use the following two confidence measures:

- *local accuracy* with parameter k – LA(k)
  LA(k) is commonly used criterion for classifier selection [14]. The confidence of classification of $\vec{x}$ using ϕ is defined as the estimate of local accuracy of ϕ near $\vec{x}$. Let $N_k(\vec{x})$ denote the set of k nearest neighbors from the training (or validation) set, closest to $\vec{x}$ under Euclidean metric. Then $\kappa_\phi^{LA(k)}(\vec{x})$ is defined as the ratio of the number of patterns from $N_k(\vec{x})$ classified correctly by ϕ, to the number of all patterns from $N_k(\vec{x})$.

- *local diversity* with parameter k – LD(k)
  Diversity of an ensemble is a measure indicating how different are the classifiers in the ensemble. If the diversity of an ensemble is too low, classifier combining fails to improve the classification. Several methods for measuring diversity of an ensemble have been proposed in the literature, see for example [12], but none of these is generally accepted.

  For our experiments, we used the *double-fault* diversity measure [12], computed on neighbors of pattern $\vec{x}$. The double-fault diversity measure expresses the similarity of the classifiers' misclassifications. Let ϕ_i, ϕ_j be two different classifiers from the ensemble $(\phi_1, \ldots, \phi_r)$, and let $N_k(\vec{x})$ be the set of k nearest neighbors from the training (or validation) set, closest to $\vec{x}$ under Euclidean metric. Then we define DF_{ij} as the ratio of the number of patterns from $N_k(\vec{x})$ classified incorrectly by both classifiers ϕ_i and ϕ_j (so-called double-faults), to the number of all patterns from $N_k(\vec{x})$. DF_{ij} varies from 0 (no double-faults) to 1 (both classifiers misclassify all the patterns), and holds information about some degree of similarity of the two classifiers. Let $DF_i = \frac{1}{r-1} \sum_{j=1;j\neq i}^{r} DF_{ij}$ be the average of DF_{ij}. The lower the DF_i, the higher the confidence of classification, therefore we define the confidence of classification of $\vec{x}$ using the $i-$th classifier from the ensemble as $\kappa_{\phi_i}^{LD(k)}(\vec{x}) = 1 - DF_i$.

Of course, instead of the local accuracy, any other measure of quality of classification could be used (precision, sensitivity, etc.), and for local diversity, any other diversity measure could be used (Q-statistics, entropy-based diversity measures, etc.).

State-of-the-art methods for classifier combining do not use both evidence and confidence of classification heavily. In classifier selection, confidence is used to select a classifier, and the evidence of other classifiers is discarded. Simple algorithms for classifier aggregation (mean value, product, maximum, minimum, etc.) use the evidence of classification only, and they disregard the confidence. Advanced classifier aggregation methods (weighted mean, fuzzy integral, etc.) incorporate confidence into aggregation, but only global confidence measures (i.e., measures independent on the current pattern, e.g. based on validation accuracy of the classifiers) are commonly used.

However, by incorporating local confidence measures (like LA or LD) into algorithms for classifier aggregation, performance of the algorithms could be improved. To show this, we modify state-of-the-art methods for classifier aggregation, so that they use the confidence of classification, and study the resulting methods' performances on three datasets – the Phoneme, Balance, and Satimage datasets. The details are given in the next section.

4 Experiments

To show that the concept of confidence of classification can improve state-of-the-art methods for classifier combining, we developed simple algorithms for classifier aggregation (Weighted Mean Value using Confidence, Filtered Mean Value, Filtered Weighted Mean Value), and compared them to other methods (Mean Value, Weighted Mean Value, Dynamic Best Selection), on three datasets from the UCI repository [5] – the Phoneme, Balance, and Satimage datasets.

The algorithms used in the experiments are described in the next section.

4.1 Algorithm Description

Let $(\phi_1, \ldots, \phi_r)$ be a team of classifiers, (2) the output of the team for a pattern $\vec{x}$. For combining the outputs of the individual classifiers, we used the following algorithms:

1. *Dynamic Best Selection* – DBS
   DBS is a classifier selection algorithm. From the team $(\phi_1, \ldots, \phi_r)$, the classifier with the maximal confidence κ_{max} is selected for prediction. If there is more than one classifier with confidence κ_{max}, a random one among them is selected.

2. *Mean Value* – MV
   MV is a classifier aggregation method. MV computes mean value of degree of classification to each class, i.e. the aggregated degree of classification to class C_j is defined as the average of the degrees of classification to class C_j through all the classifiers in the team:
   $$\mu_j = \frac{1}{r} \sum_{i=1}^{r} \mu_{i,j}. \tag{3}$$

3. *Weighted Mean Value* – WMV
   WMV computes weighted mean of the degrees of classification to class C_j through all the classifiers in the team:
   $$\mu_j = \frac{\sum_{i=1}^{r} \omega_i \mu_{i,j}}{\sum_{i=1}^{r} \omega_i}. \tag{4}$$
   The weights $\omega_1, \ldots, \omega_r$ are defined as training accuracies of the classifiers in the team.

4. *Weighted Mean Value using Confidence* – WMVC
   WMVC is a modification of WMV, the difference being that the weights $\omega_1, \ldots, \omega_r$ are not training accuracies of the classifiers, but the confidences of classifications instead, i.e. $\omega_i = \kappa_{\phi_i}(\vec{x})$.

5. *Filtered Mean Value* – FMV
   FMV is a modification of MV, the difference being that prior to computing the mean value, classifiers with confidence of classification of the current pattern lower than some threshold T are discarded. If $T = 0$, FMV coincides with MV. If there are no classifiers with confidence higher than T (i.e., all the classifiers would be discarded), T is lowered to the value of maximal confidence in the team.

6. *Filtered Weighted Mean Value* – FWMV
   FWMV is a modification of WMV, such that prior to computing the weighted mean, classifiers with confidence lower than T are discarded in the same way as in FMV.

In addition to the methods above, the data was classified by the single, non-combined classifier (for comparing the benefits of classifier combining) – this classifier will be denoted NC, and will be used as a reference classifier.

4.2 Experimental Results

For the experiments, we used an ensemble of classifiers $(\phi_1, \ldots, \phi_r)$, constructed using the Multiple Feature Subset method, i.e., we created classifiers with all possible combinations of features (all 1-D classifiers, all 2-D classifers, etc.). The ensemble $(\phi_1, \ldots, \phi_r)$ consisted of Bayesian classifiers [6].

The combination of the ensemble was done using the algorithms described in the previous section. As confidence measures for WMVC, FMV, and FWMV, we used LA(20) and LD(20). The value of the threshold T for FMV and FWMV was set experimentally. All the algorithms were implemented using the Java programming language. The results of the testing are shown in Fig. 1-3. We measured mean error rate and standard deviation of error rate from 10-fold crossvalidation.

The best results were obtained for the Phoneme dataset – for this dataset, the algorithms which used the confidence of classification (WMVC, FMV, FWMV, DBS) shown dramatic improvement to the other combination strategies (MV, WMV). The local accuracy confidence measure shown better performance than the local diversity confidence measure.

For the Balance dataset, the MV, WMV, and WMVC algorithms were worse than the NC classifier – suggesting that the ensemble of classifiers was designed poorly. Still, the FMV and FWMV show slight improvement to the NC classifier, particularly for the local diversity confidence measure – however, this improvement is small, and (due to high standard deviation) probably insignificant.

In the case of the Satimage dataset, the classification was improved best by the FMV and FWMV algorithms using the local accuracy diversity measure. For the local diversity confidence measure, the results were comparable to the NC classifier, except for the DBS algorithm, which performed worse than the NC classifier.

In general, we can summarize the that the FMV and FWMV algorithms shown best performance. The DBS selection algorithm performed well most of the time, but it is quite unstable – for the Phoneme and Satimage datasets, when the local accuracy confidence measure was replaced by the local diversity confidence measure, the performance of DBS fell rapidly, while the results for FMV and FWMV were still comparable. This is due to

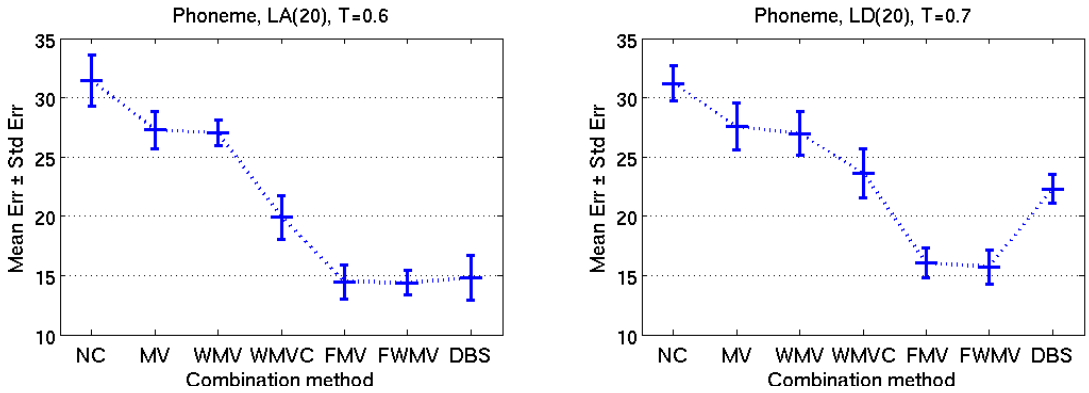


Figure 1: Mean ± standard deviation of the test error rate for the Phoneme dataset.

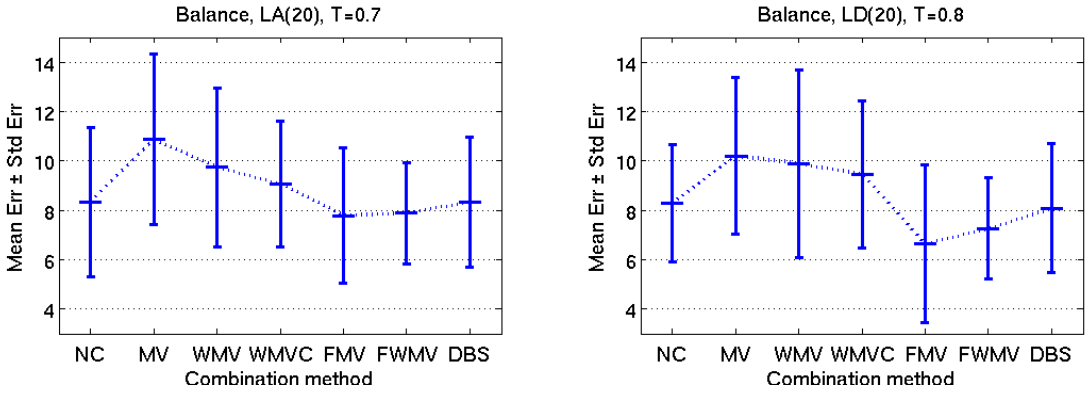


Figure 2: Mean ± standard deviation of the test error for the Balance dataset.

the fact that DBS selects just one classifier, while for the classifier aggregation algorithms, the output is a consensus of several classifiers.

When we compare the two confidence measures (local accuracy and local diversity using the double-fault diversity measure), we can roughly say that the local accuracy measure gives better results. In addition, the time complexity of computing the diversity of an ensemble is higher than computing the accuracy of a single classifier. These facts and figures favorize the local accuracy measure.

5 Summary

In this paper, we introduced the concept of confidence of classification, which can be used both as a criterium for classifier selection, and for modifying classifier aggregation methods. We defined two confidence measures (the local accuracy and the local diversity), and introduced simple algorithms for classifier aggregation which use the concept of confidence of classification – the Filtered Mean Value, Filtered Weighted Mean Value, and Weighted Mean Value using Confidence algorithms. Experimental results showed that even such simple modifications of state-of-the-art classifier aggregation algorithms

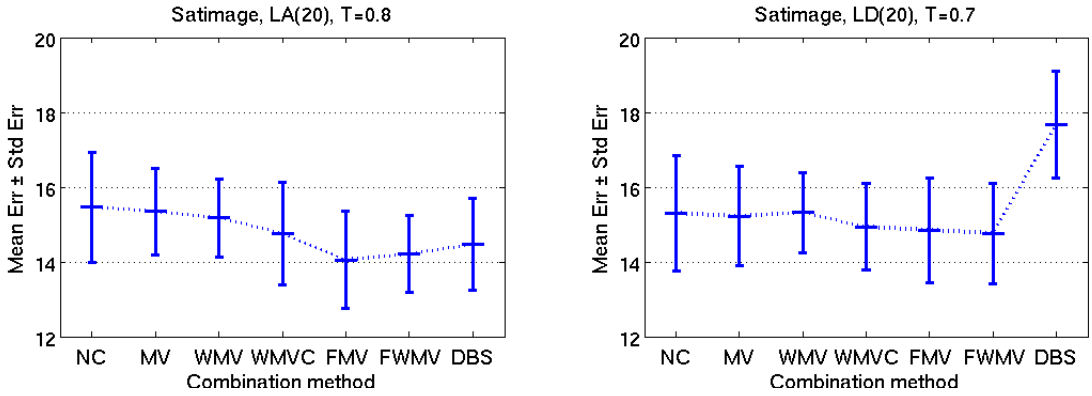


Figure 3: Mean ± standard deviation of the test error rate for the Satimage dataset.

can yield improvements in the classification.

Moreover, the concept of confidence of classification can be incorporated into many classifier combining techniques, possibly resulting in very successful methods. In addition, other confidence measures than those reported in this article can be used to further improve the algorithms. Apart from general confidence measures, based on common attributes of classifiers (like accuracy, diversity, etc.), measures which consider the specific type of the classifier (e.g. confidence based on the sum of distances to neighbors of the current pattern for k-NN classifiers) could be developed. These issues are topics of our future research.

References

[1] M. R. Ahmadzadeh and M. Petrou. *Use of Dempster-Shafer theory to combine classifiers which use different class boundaries.* Pattern Anal. Appl. **6** (2003), 41–46.

[2] M. Aksela. Comparison of classifier selection methods for improving committee performance. In 'Multiple Classifier Systems', 84–93, (2003).

[3] S. D. Bay. *Nearest neighbor classification from multiple feature subsets.* Intelligent Data Analysis **3** (1999), 191–209.

[4] L. Breiman. *Bagging predictors.* Machine Learning **24** (1996), 123–140.

[5] C. B. D.J. Newman, S. Hettich and C. Merz. UCI repository of machine learning databases, (1998). www.ics.uci.edu/∼mlearn/MLRepository.html.

[6] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification (2nd Edition).* Wiley-Interscience, (2000).

[7] Y. Freund and R. E. Schapire. Experiments with a new boosting algorithm. In 'International Conference on Machine Learning', 148–156, (1996).

[8] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas. *On combining classifiers.* IEEE Trans. Pattern Anal. Mach. Intell. **20** (1998), 226–239.

[9] L. I. Kuncheva. *Fuzzy versus nonfuzzy in combining classifiers designed by boosting.* IEEE Transactions on Fuzzy Systems **11** (2003), 729–741.

[10] L. I. Kuncheva. *Using diversity measures for generating error-correcting output codes in classifier ensembles.* Pattern Recogn. Lett. **26** (2005), 83–90.

[11] L. I. Kuncheva, J. C. Bezdek, and R. P. W. Duin. *Decision templates for multiple classifier fusion: an experimental comparison.* Pattern Recognition **34** (2001), 299–314.

[12] L. I. Kuncheva and C. J. Whitaker. *Measures of diversity in classifier ensembles.* Machine Learning **51** (2003), 181–207.

[13] O. Melnik, Y. Vardi, and C.-H. Zhang. *Mixed group ranks: Preference and confidence in classifier combination.* IEEE Trans. Pattern Anal. Mach. Intell. **26** (2004), 973–981.

[14] K. Woods, J. W. Philip Kegelmeyer, and K. Bowyer. *Combination of multiple classifiers using local accuracy estimates.* IEEE Trans. Pattern Anal. Mach. Intell. **19** (1997), 405–410.

[15] X. Zhu, X. Wu, and Y. Yang. Dynamic classifier selection for effective mining from noisy data streams. In 'ICDM '04: Proceedings of the Fourth IEEE International Conference on Data Mining (ICDM'04)', 305–312, Washington, DC, USA, (2004). IEEE Computer Society.

Vznik vlastních hodnot jako následek lokální perturbace periodického kvantového grafu

Ondřej Turek

2. ročník PGS, email: tureko1@km1.fjfi.cvut.cz
Katedra matematiky, Fakulta jaderná a fyzikálně inženýrská, ČVUT
školitel: Pavel Exner, Ústav jaderné fyziky, AV ČR,
Pierre Duclos, Université du Sud, Toulon – Var

Abstract. We will determine the spectrum of the Hamiltonian of an infinite periodic quantum graph formed by joined circles with δ-couplings with a general parameter $\alpha \in \mathbb{R}$ in the points of contact. We will show that the Hamiltonian of such system has a band spectrum. After that, we will consider a bending deformation of the chain and examine its influence on the spectrum. It will be shown that as a result new eigenvalues appear in the spectral gaps. We will describe these eigenvalues and determine their number.

Abstrakt. Obsahem práce je vyšetření spektra hamiltoniánu nekonečného periodického kvantového grafu tvořeného navzájem se dotýkajícími kruhy s δ-vazbami s obecným parametrem $\alpha \in \mathbb{R}$ v místech dotyku. Nejprve ukážeme, že hamiltonián takového systému má pásové spektrum. Poté uvážíme tvarovou deformaci spočívající v ohybu řetízku a vyšetříme její vliv na spektrum. Ukážeme, že důsledkem je vznik vlastních hodnot ve spektrálních mezerách, tyto vlastní hodnoty popíšeme a určíme jejich počet.

1 Úvod

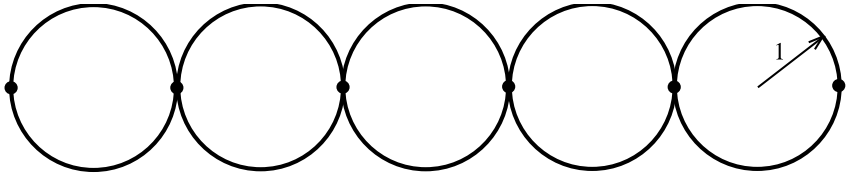
Pod pojmem *kvantový graf* rozumíme uspořádanou dvojici (Γ, H), kde Γ je metrický graf (neorientovaný graf s metrikou) a H je hamiltonián na Γ, tj. samosdružený diferenciální operátor 2.řádu působící na funkce na hranách grafu jako záporně vzatá druhá derivace (viz [3]). Tyto matematické objekty slouží jako přirozené modely pro grafům podobné struktury o rozměrech v řádu nanometrů vytvořené z různých materiálů, často polovodičů. Technologický pokrok v posledních dekádách, jenž umožnil výrobu mikroskopických struktur tohoto typu a tím jejich praktickou využitelnost, otevřel teorii kvantových grafů široké aplikační možnosti. Proto byl koncem osmdesátých let minulého století v této oblasti zahájen intenzivnější výzkum, ve kterém matematická fyzika pokračuje dodnes. Stále se však jedná o relativně novou teorii s mnoha nezodpovězenými otázkami. Jedním z problémů, který dosud není obecně vyřešen, je otázka, jak se obecně projevuje lokální perturbace periodického kvantového grafu na jeho spektru. Panuje přesvědčení, že důsledkem je vždy vznik vlastních hodnot, ale důkaz tohoto tvrzení nebyl dosud předložen, stejně tak zatím nikdo nenalezl protipříklad. K hlubšímu porozumění problému a k objevení cesty, jak jej obecně vyřešit, může napomoci studium konkrétních příkladů. Snahou tedy je prozkoumat vliv lokálních perturbací na spektrum v několika modelech a pokusit se vypozorovat společné rysy změn, ke kterým ve spektru došlo.

Tato práce si klade za cíl přispět k hledání odpovědi konkrétním příkladem. Jedná se o nekonečný řetízek tvořený vzájemně se dotýkajícími kroužky o jednotkovém poloměru

s δ-vazbami v místech dotyku kroužků (viz obr. 1). Připomeňme na tomto místě, že δ-vazbou ve vrcholu kvantového grafu se rozumí vazba vyjádřená následujícími okrajovými podmínkami:

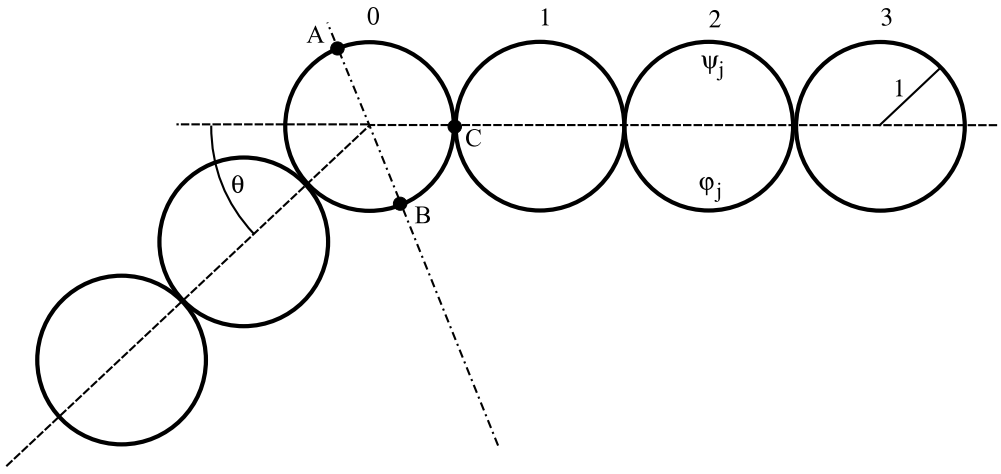
$$\psi_j(0) = \psi_k(0) =: \psi(0)\,, \quad j,k \in \hat{n}\,, \qquad \sum_{j=1}^{n} \psi_j'(0) = \alpha\psi(0)\,,$$

kde $\hat{n} = \{1, 2, \ldots, n\}$ je množina indexů hran vycházejících z uvažovaného vrcholu a $\alpha \in \mathbb{R} \cup \{+\infty\}$ je tzv. parametr vazby.



Obrázek 1: Neperturbovaný graf

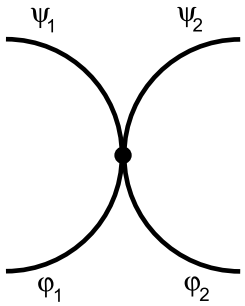
Nejprve nalezneme spektrum hamiltoniánu volné bezspinové částice na uvažovaném nekonečném řetízku, a poté uvážíme perturbaci spočívající v ohybu řetízku v jednom místě (rovinnost grafu zůstane zachována) o úhel ϑ, viz obrázek 2. U perturbovaného grafu vyšetříme spektrum hamiltoniánu a popíšeme jeho vztah ke spektru původního systému.



Obrázek 2: Perturbovaný systém

2 Spektrum nekonečného periodického systému

Původní, neperturbovaný kvantový graf je periodickým systémem, proto k výpočtu jeho spektra využijeme metodu Floquetova rozkladu (viz [4]). Uvažujme jednu elementární buňku (viz obr. 3) s vlnovými funkcemi označenými způsobem naznačeným na obrázku.



Obrázek 3: Elementární buňka periodického systému

Předpokládejme, že částice na grafu má energii E. Protože vlastní funkce musí být v takovém případě lineární kombinací funkcí $\mathrm{e}^{\mathrm{i}\sqrt{E}x}$ a $\mathrm{e}^{-\mathrm{i}\sqrt{E}x}$. Přitom není podstatné, zda je energie nezáporná, anebo záporná: pro $E = -\kappa^2$ ($\kappa > 0$) stačí položit $\sqrt{E} = \mathrm{i}\kappa$. Máme tedy

$$\begin{aligned}
\psi_1(x) &= C_1^+\mathrm{e}^{\mathrm{i}\sqrt{E}x} + C_1^-\mathrm{e}^{-\mathrm{i}\sqrt{E}x}, \quad x \in [-\pi/2, 0] \\
\psi_2(x) &= C_2^+\mathrm{e}^{\mathrm{i}\sqrt{E}x} + C_2^-\mathrm{e}^{-\mathrm{i}\sqrt{E}x}, \quad x \in [0, \pi/2] \\
\varphi_1(x) &= D_1^+\mathrm{e}^{\mathrm{i}\sqrt{E}x} + D_1^-\mathrm{e}^{-\mathrm{i}\sqrt{E}x}, \quad x \in [-\pi/2, 0] \\
\varphi_2(x) &= D_2^+\mathrm{e}^{\mathrm{i}\sqrt{E}x} + D_2^-\mathrm{e}^{-\mathrm{i}\sqrt{E}x}, \quad x \in [0, \pi/2]
\end{aligned} \tag{1}$$

V místě dotyku kroužků je předepsána δ-vazba s parametrem α, tj.

$$\begin{gathered}
\psi_1(0) = \psi_2(0) = \varphi_1(0) = \varphi_2(0) \\
-\psi_1'(0) + \psi_2'(0) - \varphi_1'(0) + \varphi_2'(0) = \alpha \cdot \psi_1(0)
\end{gathered}$$

K použití Floquetova rozkladu předpokládejme, že vlnové funkce splňují následující podmínky:

$$\begin{aligned}
\psi_2(\pi/2) &= \mathrm{e}^{\mathrm{i}k}\psi_1(-\pi/2) \qquad & \psi_2'(\pi/2) &= \mathrm{e}^{\mathrm{i}k}\psi_1'(-\pi/2) \\
\varphi_2(\pi/2) &= \mathrm{e}^{\mathrm{i}k}\varphi_1(-\pi/2) \qquad & \varphi_2'(\pi/2) &= \mathrm{e}^{\mathrm{i}k}\varphi_1'(-\pi/2)
\end{aligned}$$

pro nějaké (libovolné) $k \in [0, 2\pi)$.

Po využití předpisů (1) a úpravách obdržíme rovnosti

$$C_j^+ \cdot \sin\sqrt{E}\pi = D_j^+ \cdot \sin\sqrt{E}\pi\,, \quad C_j^- \cdot \sin\sqrt{E}\pi = D_j^- \cdot \sin\sqrt{E}\pi\,,$$

z nichž plyne, že pro $\sqrt{E} \notin \mathbb{N}_0$ je $C_j^+ = C_j^-$ a $D_j^+ = D_j^-$. Považujme nyní tento předpoklad za splněný s tím, že singulární případ $\sqrt{E} \in \mathbb{N}_0$ vyšetříme nakonec.

Po eliminaci dalších proměnných dojdeme k rovnici druhého stupně pro $\mathrm{e}^{\mathrm{i}k}$,

$$\mathrm{e}^{2\mathrm{i}k} - \mathrm{e}^{\mathrm{i}k}\left(2\cos\sqrt{E}\pi + \frac{\alpha}{2\sqrt{E}}\sin\sqrt{E}\pi\right) + 1 = 0\,, \tag{2}$$

která má reálné koeficienty a jejíž diskriminant je určen výrazem

$$D = \left(2\cos\sqrt{E}\pi + \frac{\alpha}{2\sqrt{E}}\sin\sqrt{E}\pi\right)^2 - 4\,.$$

Úkolem je určit, pro jaké hodnoty E existuje $k \in [0, 2\pi)$ takové, aby byla splněna rovnice (2), neboli pro jaké E má (2) jakožto rovnice o neznámé $\mathrm{e}^{2\mathrm{i}k}$ alespoň jeden kořen o absolutní hodnotě 1. Povšimněme si, že součin každé dvojice kořenů (2), bez ohledu na hodnotu $\sqrt{E}$, je vždy roven 1, neboť jsou si rovny koeficienty u kvadratického a lineárního členu. To ovšem znamená, že buďto jsou oba kořeny komplexní jednotky, anebo má jeden z nich absolutní hodnotu větší než jedna, zatímco druhý menší než jedna. Je zřejmé, že kladný diskriminant odpovídá první situaci a nekladný té druhé. Dostáváme tak závěr:

Věta 1. *Je-li $E \geq 0$ a $\sqrt{E} \notin \mathbb{N}_0$, pak $\sqrt{E} \in \sigma(H)$ právě tehdy, je-li splněna podmínka*

$$\left| \cos \sqrt{E}\pi + \frac{\alpha}{4} \cdot \frac{\sin \sqrt{E}\pi}{\sqrt{E}} \right| \leq 1 \,. \tag{3}$$

Vraťme se ještě k situaci, kdy $\sqrt{E} \in \mathbb{N}_0$. Jak si lze lehce představit, lze zkonstruovat funkci s nosičem na pouhém jednom kroužku tak, aby byla vlastní funkcí hamiltoniánu: stačí zvolit na horní půlkružnici funkci $\sin \sqrt{E}x$ ($x \in [0, \pi]$) a na dolní půlkružnici funkci $-\sin \sqrt{E}x$ ($x \in [0, \pi]$). To ovšem znamená, že hodnoty n^2 pro $n \in \mathbb{N}_0$ patří do bodového spektra.

Vyslovme charakteristiku spektrálních pásů určených podmínkou (3):

Tvrzení 2.

- *Je-li $\alpha > 0$, pak v každém intervalu $[n^2, (n+1)^2]$ ($n \in \mathbb{N}_0$) je právě jeden spektrální pás, jehož levý krajní bod leží uvnitř tohoto intervalu a pravý krajní bod splývá s hodnotou $(n+1)^2$,*
- *je-li $\alpha < 0$, pak v každém intervalu $[n^2, (n+1)^2]$ ($n \in \mathbb{N}_0$) je právě jeden spektrální pás, jehož levý krajní bod splývá s hodnotou n^2 a pravý krajní bod leží uvnitř tohoto intervalu,*
- *je-li $\alpha = 0$, pak podmínku (3) splňuje každé $\sqrt{E} \geq 0$, tedy ve spektru leží všechna nezáporná čísla.*

Vidíme tedy, že právě jeden z krajních bodů každého spektrálního pásu, totiž druhá mocnina celého čísla, odpovídá vlastní hodnotě hamiltoniánu.

Poznamenejme, že podmínka (3), již jsme obdrželi, je podobná odpovídající podmínce z Kronig-Penneyova modelu se vzdáleností interakcí π, jediný rozdíl je v koeficientu u sinu: zde máme $\frac{\alpha}{4}$, zatímco v K-P modelu $\frac{\alpha}{2}$ (viz např. [1]). To znamená, že spektrální pásy uvažovaného "řetízku" s δ-vazbami s parametrem α mají krajní body stejné jako ty u K-P modelu se vzdáleností mezi interakcemi rovnou π a parametrem interakce $\alpha/2$. Podstatný rozdíl mezi oběma modely však spočívá v tom, že K-P model má prázdné bodové spektrum.

Vyřešme ještě otázku, jak vypadá záporná část spektra. Označme $\sqrt{E} = \mathrm{i}\kappa$ pro $\kappa > 0$ a s využitím známých vztahů pro goniometrické funkce komplexních argumentů přepišme nerovnost (3) do podoby

$$\left| \cosh \kappa\pi + \frac{\alpha}{4} \cdot \frac{\sinh \kappa\pi}{\kappa} \right| \leq 1 \,.$$

Pro $E < 0$ je vždy $\sin(\sqrt{E}) \neq 0$, takže zde odpadá problém, na který jsme narazili u nezáporných energií. Jednoduše lze dokázat následující tvrzení:

Tvrzení 3.

- *Je-li* $\alpha \geq 0$, *pak záporné spektrum je prázdné,*
- *je-li* $\alpha \in [-8/\pi, 0)$, *pak záporné spektrum je rovno intervalu* $[-\kappa_1^2, 0]$, *kde* κ_1 *je jediné řešení rovnice* $\left|\cosh \kappa\pi + \frac{\alpha}{4} \cdot \frac{\sinh \kappa\pi}{\kappa}\right| = 1$,
- *Je-li* $\alpha < -8/\pi$, *pak záporné spektrum je rovno intervalu* $[-\kappa_1^2, -\kappa_2^2]$, *kde* $\kappa_{1,2}$ *jsou jediná dvě kladná řešení* $\left|\cosh \kappa\pi + \frac{\alpha}{4} \cdot \frac{\sinh \kappa\pi}{\kappa}\right| = 1$, $\kappa_1 > \kappa_2$.

3 Perturbovaný systém

Předpokládejme nyní, že původně rovný, nekonečný řetízek v jednom místě modifikujeme způsobem dle obrázku 2. Úhel ohybu ϑ budeme uvažovat libovolný v intervalu $(0, \pi)$, byť pro $\vartheta \geq 2\pi/3$ je deformace naznačená na obrázku prakticky neproveditelný. Teoretickým úvahám však nic nebrání, neboť ohyb je možné ekvivalentně nahradit deformací nultého kroužku.

Uvažovanou perturbací sice systém ztrácí původní periodicitu, ale stále si zachovává určitou symetrii, která umožňuje výpočet spektra mírně zjednodušit. Vidíme, že perturbovaný systém je symetrický vůči ose, která je do obrázku 2 zakreslena čerchovanou čarou. Rozložíme vlnovou funkci na součet dvou funkcí: jedné, která je "sudá" vůči této ose, a druhé, která je vzhledem k ní "lichá". Hamiltonián systému pak podobným způsobem rozložíme na direktní součet operátoru H^+, který působí na sudou složku vlnové funkce, a H^-, který působí na lichou složku. Spektrum operátoru $H = H^+ \oplus H^-$ je pak dáno jako sjednocení spekter H^+ a H^-.

Veškeré složky vlnové funkce (ve smyslu funkce na všech hranách grafu) budou dány jako lineární kombinace funkcí $\mathrm{e}^{\mathrm{i}\sqrt{E}x}$ a $\mathrm{e}^{-\mathrm{i}\sqrt{E}x}$, tak jako v případě neperturbovaného systému, je však vhodné zavést nové značení. Kroužky budeme indexovat celými čísly, přičemž zavedeme úmluvu, že kroužek, jímž prochází osa souměrnosti, bude označen číslem 0. Protože se budeme zabývat funkcemi symetrickými vůči ose, stačí studovat situaci na pravé části systému, která je na obrázku zakreslena vodorovně. Vlnovou funkci na každém kroužku rozdělíme na funkci na horní půlkružni-ci a na funkci na dolní půlkružnici a označíme je po řadě ψ_j a φ_j, kde j vyjadřuje index kroužku. Máme tedy

$$\begin{aligned} \psi_j(x) &= C_j^+ \mathrm{e}^{\mathrm{i}\sqrt{E}x} + C_j^- \mathrm{e}^{-\mathrm{i}\sqrt{E}x}, \quad x \in [0, \pi] \\ \varphi_j(x) &= D_j^+ \mathrm{e}^{\mathrm{i}\sqrt{E}x} + D_j^- \mathrm{e}^{-\mathrm{i}\sqrt{E}x}, \quad x \in [0, \pi] \end{aligned} \tag{4}$$

pro $j \in \mathbb{N}$. Je důležité poznamenat, že v případě $j = 0$ sice platí stejný předpis, ale proměnné probíhají jiné intervaly, tj.

$$\begin{aligned} \psi_0(x) &= C_0^+ \mathrm{e}^{\mathrm{i}\sqrt{E}x} + C_0^- \mathrm{e}^{-\mathrm{i}\sqrt{E}x}, \quad x \in \left[\frac{\pi - \vartheta}{2}, \pi\right] \\ \varphi_0(x) &= D_0^+ \mathrm{e}^{\mathrm{i}\sqrt{E}x} + D_0^- \mathrm{e}^{-\mathrm{i}\sqrt{E}x}, \quad x \in \left[\frac{\pi + \vartheta}{2}, \pi\right] \end{aligned} \tag{5}$$

Protože v bodech dotyku jsou předepsány δ-vazby s parametrem α, platí

$$\psi_j(0) = \varphi_j(0) \qquad \psi_j(\pi) = \varphi_j(\pi) \tag{6}$$

a

$$\psi_j(0) = \psi_{j-1}(\pi) \tag{7}$$
$$\psi'_j(0) + \varphi'_j(0) - \psi'_{j-1}(\pi) - \varphi'_{j-1}(\pi) = \alpha \cdot \psi_j(0) \tag{8}$$

Dosazením (4) do (6) a elementární úpravou získáme podmínky

$$C_j^+ \cdot \sin\sqrt{E}\pi = D_j^+ \cdot \sin\sqrt{E}\pi \quad \text{a} \quad C_j^- \cdot \sin\sqrt{E}\pi = D_j^- \cdot \sin\sqrt{E}\pi\,,$$

takže pro $\sqrt{E} \notin \mathbb{N}_0$ musí platit $C_j^+ = D_j^+$ a $C_j^- = D_j^-$. (Případ $\sqrt{E} \in \mathbb{N}_0$ lze okomentovat obdobně jako v případě nekonečného lineárního řetízku a dojít tak k závěru, že druhé mocniny celých čísel jsou vlastními hodnotami.) Využijeme-li dokázané rovnosti k úpravě (7) a (8), dostaneme rovnici

$$\begin{pmatrix} C_j^+ \\ C_j^- \end{pmatrix} = \underbrace{\begin{pmatrix} \left(1 + \frac{\alpha}{4\mathrm{i}\sqrt{E}}\right)\mathrm{e}^{\mathrm{i}\sqrt{E}\pi} & \frac{\alpha}{4\mathrm{i}\sqrt{E}}\mathrm{e}^{-\mathrm{i}\sqrt{E}\pi} \\ -\frac{\alpha}{4\mathrm{i}\sqrt{E}}\mathrm{e}^{\mathrm{i}\sqrt{E}\pi} & \left(1 - \frac{\alpha}{4\mathrm{i}\sqrt{E}}\right)\mathrm{e}^{-\mathrm{i}\sqrt{E}\pi} \end{pmatrix}}_{M} \cdot \begin{pmatrix} C_{j-1}^+ \\ C_{j-1}^- \end{pmatrix} \tag{9}$$

platnou pro všechna $j \geq 2$. Z ní okamžitě plyne, že pro všechna $j \geq 2$ je

$$\begin{pmatrix} C_j^+ \\ C_j^- \end{pmatrix} = M^{j-1} \cdot \begin{pmatrix} C_1^+ \\ C_1^- \end{pmatrix}, \tag{10}$$

odkud vyplývá asymptotické chování posloupnosti absolutních hodnot vektorů $(C_j^+, C_j^-)^T$: Nechť je $(C_1^+, C_1^-)^T$ vlastním vektorem matice M. Pak platí:

- přísluší-li $(C_1^+, C_1^-)^T$ vlastnímu číslu v absolutní hodnotě menšímu než 1, pak $\left\|(C_j^+, C_j^-)^T\right\|$ exponenciálně klesá,
- přísluší-li $(C_1^+, C_1^-)^T$ vlastnímu číslu v absolutní hodnotě menšímu než 1, pak $\left\|(C_j^+, C_j^-)^T\right\|$ exponenciálně roste,
- přísluší-li $(C_1^+, C_1^-)^T$ vlastnímu číslu v absolutní hodnotě menšímu než 1, pak $\left\|(C_j^+, C_j^-)^T\right\|$ nezávisí na j.

V obou případech, jak u operátoru H^+, tak i H^-, je absolutní hodnota vlnové funkce na j-tém i (-j)-tém kroužku přímo určena konstantami C_j^+ a C_j^-. Ty jsou zase, dle vzorce (10), dány vektorem $(C_1^+, C_1^-)^T$. Uvědomíme si, že pokud by rozklad vektoru $(C_1^+, C_1^-)^T$ do vlastních podprostorů M obsahoval nenulovou složku příslušející vlastnímu číslu M s absolutní hodnotou větší než 1, norma vektoru $(C_j^+, C_j^-)^T$ by asymptoticky exponenciálně rostla. Je evidentní, že v takovém případě by konstanty $C_j^\pm$ nemohly určovat vlastní funkci ani zobecněnou vlastní funkci H^+, resp. H^-.

Stejně tak lze nahlédnout, že skládá-li se vektor $(C_1^+, C_1^-)^T$ jen z vlastních vektorů M příslušejících vlastním číslům o absolutní hodnotě menší než (resp. nebo rovné) 1, pak je konstantami $C_j^\pm$ určena vlastní (resp. zobecněná vlastní) funkce, tj. odpovídající hodnota E patří do bodového (resp. spojitého) spektra.

U obou operátorů H^+ i H^- nejprve určíme, jak vypadá vektor $(C_1^+, C_1^-)^T$, a poté rozhodneme, pro jaké hodnoty $\sqrt{E}$ může být vlastním vektorem M příslušejícím vlastnímu

číslu o absolutní hodnotě menší než 1, eventuálně rovné 1. Charakteristickým polynomem matice M je

$$\lambda^2 - \lambda \cdot 2 \cdot \left(\cos\sqrt{E}\pi + \frac{\alpha}{4\sqrt{E}} \sin\sqrt{E}\pi \right) + 1 \,,$$

což je polynom s reálnými koeficienty. Podobnými úvahami jako v kapitole věnované neperturbovanému systému dostáváme, že M má vlastní číslo v absolutní hodnotě menší než jedna tehdy a jen tehdy, je-li diskriminant tohoto polynomu kladný, což je ekvivalentní podmínce

$$\left| \cos\sqrt{E}\pi + \frac{\alpha}{4\sqrt{E}} \sin\sqrt{E}\pi \right| > 1 \,,$$

a vlastní číslo v absolutní hodnotě rovné 1, pokud

$$\left| \cos\sqrt{E}\pi + \frac{\alpha}{4\sqrt{E}} \sin\sqrt{E}\pi \right| \leq 1 \,,$$

Porovnáním s podmínkou (3) vidíme, že spektrální pásy se po perturbaci zachovávají (což ostatně plyne i z toho, že hamiltonián perturbovaného systému má konečné indexy defektu, a tedy nedochází ke změně esenciálního spektra, viz [2]), a dále, že případné nově vzniklé vlastní hodnoty hamiltoniánu mohou ležet jen ve spektrálních mezerách.

Vzhledem k tomu, že v případě kladného diskriminantu jsou vlastní čísla dána předpisem

$$\lambda_{1,2} = \cos\sqrt{E}\pi + \frac{\alpha}{4\sqrt{E}} \sin\sqrt{E}\pi \pm \sqrt{\left(\cos\sqrt{E}\pi + \frac{\alpha}{4\sqrt{E}} \sin\sqrt{E}\pi \right)^2 - 1} \,,$$

platí, že je-li diskriminant kladný, je příslušným vlastním číslem v absolutní hodnotě menším než jedna

- λ_2 pro $\cos\sqrt{E}\pi + \frac{\alpha}{4\sqrt{E}} \sin\sqrt{E}\pi > 1$,
- λ_1 pro $\cos\sqrt{E}\pi + \frac{\alpha}{4\sqrt{E}} \sin\sqrt{E}\pi < -1$.

3.1 Spektrum H^+

Operátor H^+ odpovídá sudé složce vlnové funkce, uvažováno ve smyslu symetrie vůči ose. Sudost je vyjádřena následujícími okrajovými podmínkami v bodech A a B (viz obr. 2):

$$\psi_0'\left(\frac{\pi - \vartheta}{2}\right) = 0 \,, \quad \varphi_0'\left(\frac{\pi + \vartheta}{2}\right) = 0 \,.$$

V místě dotyku nultého a prvního kroužku (ozn. C) je δ-vazba s parametrem α:

$$\psi_0(\pi) = \varphi_0(\pi) = \psi_1(0) \tag{11}$$

$$\psi_1'(0) + \varphi_1'(0) - \psi_0'(\pi) - \varphi_0'(\pi) = \alpha \cdot \psi_0(\pi) \tag{12}$$

Dosadíme-li nyní za jednotlivé funkce z (4) a (5) a použijeme-li už známý vztah $\varphi_1'(0) = \psi_1'(0)$, získáme vektor $(C_1^+, C_1^-)^T$ (až na jeho násobení konstantou):

$$\begin{pmatrix} C_1^+ \\ C_1^- \end{pmatrix} = \begin{pmatrix} 1 + \mathrm{i}\left(\frac{\sin\sqrt{E}\pi}{\cos\sqrt{E}\pi+\cos\sqrt{E}\vartheta} - \frac{\alpha}{2\sqrt{E}}\right) \\ 1 - \mathrm{i}\left(\frac{\sin\sqrt{E}\pi}{\cos\sqrt{E}\pi+\cos\sqrt{E}\vartheta} - \frac{\alpha}{2\sqrt{E}}\right) \end{pmatrix}.$$

Aby $\sqrt{E} \in \sigma_p(H^+)$, musí být vektor $(C_1^+, C_1^-)^T$ násobkem vlastního vektoru M příslušejícího vlastnímu číslu v absolutní hodnotě menšímu než 1. Řešením této podmínky dostaneme rovnost

$$\cos\sqrt{E}\vartheta = -\cos\sqrt{E}\pi + \frac{\sin^2\sqrt{E}\pi}{\frac{\alpha}{4\sqrt{E}}\sin\sqrt{E}\pi \pm \sqrt{\left(\cos\sqrt{E}\pi + \frac{\alpha}{4\sqrt{E}}\sin\sqrt{E}\pi\right)^2 - 1}}, \qquad (13)$$

kde znaménko ve jmenovateli je rovno znaménku $\cos\sqrt{E}\pi + \frac{\alpha}{4\sqrt{E}}\sin\sqrt{E}\pi$.

Označme výraz na pravé straně symbolem $f(\sqrt{E})$. Bez důkazu nyní uvedeme následující důležité tvrzení o funkci f:

Tvrzení 4. *V každém intervalu, v němž* $|\cos x\pi + \frac{\alpha}{4x}\sin x\pi| \geq 1$, *je funkce* f *ostře monotonní a probíhá interval* $[-1, 1]$.

Na levé straně rovnosti (13) stojí výraz $\cos\sqrt{E}\vartheta$, kde ϑ je úhel ohybu, ležící v intervalu $(0, \pi)$. Jelikož $\vartheta < \pi$, je délka intervalu, ve kterém funkce $\cos x\vartheta$ proběhne interval $[-1, 1]$, větší než 1. Na druhou stranu, jak už víme, je délka spektrálních mezer menší než 1. S ohledem na poslední Tvrzení tak s použitím věty o střední hodnotě dostáváme, že v každém uzávěru spektrální mezery existuje právě jeden bod, v němž nastává rovnost. Buď tedy vznikne jedna vlastní hodnota ve spektrální mezeře, anebo rovnost nastane na hranici spektrálního pásu bez vzniku vlastní hodnoty.

Prozkoumejme ještě zápornou část spektra. Pro $\sqrt{E} = \mathrm{i}\kappa$ $(\kappa > 0)$ získává rovnost (13) podobu

$$\cosh\kappa\vartheta = -\cosh\kappa\pi - \frac{\sinh^2\kappa\pi}{\frac{\alpha}{4\kappa}\sinh\kappa\pi \pm \sqrt{\left(\cosh\kappa\pi + \frac{\alpha}{4\kappa}\sinh\kappa\pi\right)^2 - 1}}, \qquad (14)$$

kde opět horní znaménko odpovídá $\cosh\kappa\pi + \frac{\alpha}{4\kappa}\sinh\kappa\pi > 1$, dolní znaménko odpovídá $\cosh\kappa\pi + \frac{\alpha}{4\kappa}\sinh\kappa\pi < -1$.

Označíme-li výraz na pravé straně symbolem $f^-(\kappa)$, můžeme vyslovit následující pomocné tvrzení, které okamžitě implikuje chování spektra operátoru H^+:

Tvrzení 5.

- *Je-li* $\alpha \geq 0$, *pak* $f^-(\kappa) < -\cosh(\kappa\theta)$ *pro všechna* $\kappa > 0$ *a* $\theta \in (0, \pi)$.
- *Je-li* $\alpha < 0$, *pak platí:*
  - *pro* $\cosh\kappa\pi + \frac{\alpha}{4\kappa}\sinh\kappa\pi < -1$ *je* $f^-(\kappa) < -\cosh\kappa\vartheta$ *pro všechna* $\kappa > 0$ *a* $\vartheta \in (0, \pi)$,
  - *pro* $\cosh\kappa\pi + \frac{\alpha}{4\kappa}\sinh\kappa\pi > 1$ *a zároveň* $\kappa \cdot \mathrm{tgh}\kappa\pi < -\alpha/2$ *je funkce* $f^-(\kappa)$ *ostře rostoucí a probíhá interval* $(1, +\infty)$,

– *pro* $\kappa \cdot \mathrm{tgh}\kappa\pi > -\alpha/2$ *je* $f^-(\kappa) < -\cosh\kappa\vartheta$ *pro všechna* $\kappa > 0$ *a* $\vartheta \in (0,\pi)$.

Důsledek 6.

- *Je-li* $\alpha \geq 0$, *pak* H^+ *nemá záporné vlastní hodnoty.*
- *Je-li* $\alpha < 0$, *pak* H^+ *má právě jednu vlastní hodnotu, která leží nalevo od záporného spektrálního pásu a zároveň napravo od záporně vzaté druhé mocniny řešení rovnice* $\kappa \cdot \mathrm{tgh}\kappa\pi = -\alpha/2$.

3.2 Spektrum H^-

Postup při vyšetřování spektra operátoru H^- odpovídajícího sudé složce vlnové funkce je zcela analogický postupu, jaký byl použit u operátoru H^+. Jediným rozdílem jsou okrajové podmínky na nultém kroužku, které jsou dány:

$$\psi_0\left(\frac{\pi-\vartheta}{2}\right) = 0\,, \quad \varphi_0\left(\frac{\pi+\vartheta}{2}\right) = 0\,.$$

Jak lze snadno ukázat, spektrální podmínka je vyjádřená rovností

$$-\cos\sqrt{E}\vartheta = -\cos\sqrt{E}\pi + \frac{\sin^2\sqrt{E}\pi}{\frac{\alpha}{4\sqrt{E}}\sin\sqrt{E}\pi \pm \sqrt{\left(\cos\sqrt{E}\pi + \frac{\alpha}{4\sqrt{E}}\sin\sqrt{E}\pi\right)^2 - 1}}\,,$$

tedy rozdíl oproti odpovídající podmínce u operátoru H^+ spočívá v opačném znaménku u kosinu na levé straně. Chování výrazu napravo už známe (viz Tvrzení 4), rovnou tedy můžeme popsat kladnou část bodového spektra H^-:

Tvrzení 7. *V každé spektrální mezeře operátoru* H^- *existuje buď právě jedna vlastní hodnota, anebo žádná, přičemž druhá možnost nastává v případě, kdy zároveň platí* $|\cos\sqrt{E}\pi + \frac{\alpha}{4\sqrt{E}}\sin\sqrt{E}\pi| = 1$.

Záporná část bodového spektra H^- je určena podmínkou

$$-\cosh\kappa\vartheta = -\cosh\kappa\pi - \frac{\sinh^2\kappa\pi}{\frac{\alpha}{4\kappa}\sinh\kappa\pi \pm \sqrt{\left(\cosh\kappa\pi + \frac{\alpha}{4\kappa}\sinh\kappa\pi\right)^2 - 1}}\,,$$

kde $\sqrt{E} = \mathrm{i}\kappa^2$ pro $\kappa \in \mathbb{R}^+$. Nyní stačí jen využít už vyslovené tvrzení 4. Z něj okamžitě plyne, že v posledním vztahu nikdy nenastává rovnost, tedy operátor H^- nemá žádné záporné vlastní hodnoty.

3.3 Shrnutí: spektrum $H = H^+ \oplus H^-$

V předchozích kapitolách jsme ukázali, že uvažovaná perturbace má za následek vznik vlastních hodnot uvnitř spektrálních mezer neperturbovaného systému. Uvedli jsme, že každé komponentě, tj. jak operátoru H^+, tak i H^-, může v každé mezeře patřit jedna nebo žádná vlastní hodnota. Taktéž jsme vysvětlili, že situace, kdy operátor H^+ nebo H^- nemá uvnitř spektrální mezery vlastní hodnotu, odpovídá právě tomu, že poloha

příslušného bodu padne do hranice spektrálního pásu. Na druhou stranu, vzhledem k rychlosti růstu resp. poklesu funkce $\cos(\sqrt{E}\vartheta)$, která je bez ohledu na volbu ϑ vždy nižší než rychlost růstu a poklesu funkce označené symbolem $f(\sqrt{E})$, je snadno zřejmé, že nemůže dojít k tomu, aby v jedné spektrální mezeře takto vymizely obě vlastní hodnoty.

Dále je třeba podotknout, že není vyloučen případ, kdy operátorům H^+ a H^- v dané spektrální mezeře přísluší tatáž vlastní hodnota. V takovém případě dochází ke zdvojnásobení její násobnosti. Přímým výpočtem lze mimochodem ukázat, že vlastními hodnotami tohoto typu mohou být jen řešení rovnice

$$\sqrt{E} \cdot \mathrm{tg}\sqrt{E}\pi = \frac{\alpha}{2}\,.$$

Shrňme tedy výsledek výpočtu do věty.

Věta 8.
- *Body spektra neperturbovaného systému i jejich charakter se zachovávají i v perturbovaném systému.*
- *Perturbovaný systém má navíc vlastní hodnoty ve spektrálních mezerách, přičemž:*
  - *na kladné poloose je v každé spektrální mezeře buď jedna vlastní hodnota o násobnosti 1, nebo dvě vlastní hodnoty o násobnosti 1, nebo jedna vlastní hodnota o násobnosti 2,*
  - *je-li parametr δ-vazby záporný, pak se navíc na záporné poloose vlevo od posledního spektrálního pásu nachází jedna vlastní hodnota o násobnosti 1.*

Literatura

[1] S. Albeverio, F. Gesztesy, R. Hoegh-Krohn, H. Holden *Solvable Models in Quantum Mechanics*. Springer Verlag, Berlin, (1988).

[2] J. Blank, P. Exner, M. Havlíček M. *Lineární operátory v kvantové fyzice*. Karolinum, Praha, (1993).

[3] P. Kuchment. *Quantum graphs: I. Some basic structures*. Waves Random Media **14**, (2004), S107-S128.

[4] M. Reed, B. Simon. *Methods of Modern Mathematical Physics, IV. Analysis of Operators*. Academic Press, New York, (1978).

A Quantum Dot with Impurity in the Lobachevsky Plane

Matěj Tušek*

2nd year of PGS, email: tusekm1@km1.fjfi.cvut.cz
Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU
advisor: Pavel Šťovíček, Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, CTU

Abstract. The curvature effect on a quantum dot with impurity is investigated. The model is considered on the Lobachevsky plane. The confinement and impurity potentials are chosen so that the model is explicitly solvable. The Green function as well as the Krein Q-function are computed.

Abstrakt. Článek pojednává o vlivu křivosti na kvantovou tečku s nečistotou. Konkrétně uvažujeme kvantovou tečku v Lobachevského rovině. Vazebný potenciál a potenciál pro nečistotu volíme tak, že výsledný model je řešitelný. Získáme tak explicitní vyjádření pro Greenovu a Kreinovu Q-funkci.

1 Introduction

Physically, quantum dots are nanostructures with a charge carriers confinement in all space directions. They have an atom-like energy spectrum which can be modified by adjusting geometric parameters of the dots as well as by the presence of an impurity. Thus the study of these dependencies may be of interest from the point of view of the nanoscopic physics.

A detailed analysis of three-dimensional quantum dots with a short-range impurity in the Euclidean space can be found in [4]. Therein, the harmonic oscillator potential was used to introduce the confinement, and the impurity was modeled by a point interaction (δ-potential). The starting point of the analysis was derivation of a formula for the Green function of the unperturbed Hamiltonian (i.e., in the impurity free case), and application of the Krein resolvent formula jointly with the notion of the Krein Q-function.

In the present paper, we make use of the same method to investigate quantum dots with impurity in the Lobachevsky plane. We will introduce an appropriate Hamiltonian in a manner quite analogous to that of [4] and derive an explicit formula for the corresponding Green function. In this sense, our model is solvable, and so its properties may be of interest also from the mathematical point of view.

During the computations to follow, the spheroidal functions appear naturally. Unfortunately, the notation in the literature concerned with this type of special functions is not yet uniform (see, e.g., [2] and [8]). This is why we supply, for the reader's convenience,

*in co-operation with V. Geyler from Mordovian State University, Saransk, Russia

a short appendix comprising basic definitions and results related to spheroidal functions which are necessary for our approach.

2 A quantum dot with impurity in the Lobachevsky plane

2.1 The model

Denote by (ϱ, ϕ), $0 < \varrho < \infty$, $0 \leq \phi < 2\pi$, the geodesic polar coordinates on the Lobachevsky plane. Then the metric tensor is diagonal and reads

$$(g_{ij}) = \mathrm{diag}\Big(1, a^2 \sinh^2 \frac{\varrho}{a}\Big)$$

where a, $0 < a < \infty$, denotes the so called curvature radius which is related to the scalar curvature by the formula $R = -2/a^2$. Furthermore, the volume form equals $dV = a \sinh(\rho/a) \mathrm{d}\rho \wedge \mathrm{d}\phi$. The Hamiltonian for a free particle of mass $m = 1/2$ takes the form

$$H^0 = -\left(\Delta_{LB} + \frac{1}{4a^2}\right) = -\frac{1}{\sqrt{g}} \frac{\partial}{\partial x^i} \sqrt{g} g^{ij} \frac{\partial}{\partial x^j} - \frac{1}{4a^2}$$

where Δ_{LB} is the Laplace-Beltrami operator and $g = \det g_{ij}$. We have set $\hbar = 1$.

The choice of a potential modeling the confinement is ambiguous. We naturally require that the potential takes the standard form of the quantum dot potential in the flat limit ($a \to \infty$). This is to say that, in the limiting case, it becomes the potential of the isotropic harmonic oscillator $V = \frac{1}{4}\omega^2\rho^2$. However, this condition clearly does not specify the potential uniquely. Having the freedom of choice let us discuss the following two possibilities:

$$\text{a)} \quad V_a(\rho) = \tfrac{1}{4}\, a^2\omega^2 \tanh^2 \tfrac{\varrho}{a}, \tag{1}$$

$$\text{b)} \quad U_a(\rho) = \tfrac{1}{4}\, a^2\omega^2 \sinh^2 \tfrac{\varrho}{a}. \tag{2}$$

Potential V_a is the same as that proposed in [9] for the classical harmonic oscillator on the Lobachevsky plane. With this choice, it has been demonstrated in [9] that the model is superintegrable, i.e., there exist three functionally independent constants of motion. Let us remark that this potential is bounded, and so it represents a bounded perturbation to the free Hamiltonian. On the other hand, the potential U_a is unbounded. Moreover, as shown below, the stationary Schrödinger equation for this potential leads, after the partial wave decomposition, to the differential equation of spheroidal functions. The current paper concentrates exclusively on case b).

The impurity is modeled by a δ-potential which is introduced with the aid of self-adjoint extensions and is determined by boundary conditions at the base point. We restrict ourselves to the case when the impurity is located in the centre of the dot ($\rho = 0$). Thus we start from the following symmetric operator:

$$\begin{aligned} H &= -\left(\frac{\partial^2}{\partial\varrho^2} + \frac{1}{a}\coth\Big(\frac{\varrho}{a}\Big)\frac{\partial}{\partial\varrho} + \frac{1}{a^2}\sinh^{-2}\Big(\frac{\varrho}{a}\Big)\frac{\partial^2}{\partial\phi^2} + \frac{1}{4a^2}\right) + \frac{1}{4}\, a^2\omega^2 \sinh^2\Big(\frac{\rho}{a}\Big), \\ \mathrm{Dom}(H) &= C_0^\infty((0,\infty)\times S^1) \subset L^2\left((0,\infty)\times S^1, a\,\sinh\Big(\frac{\varrho}{a}\Big)\,\mathrm{d}\varrho\,\mathrm{d}\phi\right). \end{aligned} \tag{3}$$

2.2 Partial wave decomposition

Substituting $\xi = \cosh(\varrho/a)$ we obtain

$$H = \frac{1}{a^2}\left[(1-\xi^2)\frac{\partial^2}{\partial\xi^2} - 2\xi\frac{\partial}{\partial\xi} + (1-\xi^2)^{-1}\frac{\partial^2}{\partial\phi^2} + \frac{a^4\omega^2}{4}(\xi^2-1) - \frac{1}{4}\right] =: \frac{1}{a^2}\tilde{H}, \tag{4}$$
$$\mathrm{Dom}(H) = C_0^\infty((1,\infty)\times S^1) \subset L^2\left((1,\infty)\times S^1, a^2\mathrm{d}\xi\,\mathrm{d}\phi\right).$$

Using the rotational symmetry which amounts to a Fourier transform in the variable ϕ, $\tilde{H}$ may be decomposed into a direct sum as follows

$$\tilde{H} = \bigoplus_{m=-\infty}^{\infty} \tilde{H}_m,$$
$$\tilde{H}_m = -\frac{\partial}{\partial\xi}\left((\xi^2-1)\frac{\partial}{\partial\xi}\right) + \frac{m^2}{\xi^2-1} + \frac{a^4\omega^2}{4}(\xi^2-1) - \frac{1}{4},$$
$$\mathrm{Dom}(\tilde{H}_m) = C_0^\infty(1,\infty) \subset L^2((1,\infty),\mathrm{d}\xi).$$

Note that $\tilde{H}_m$ is a Sturm-Liouville operator.

Proposition 1. *$\tilde{H}_m$ is essentially self-adjoint for $m \neq 0$, $\tilde{H}_0$ has defect indices $(1,1)$.*

Proof. The operator $\tilde{H}_m$ is symmetric and semibounded, and so the defect indices are equal. If we set

$$\mu = |m|,\ 4\theta = -\frac{a^4\omega^2}{4},\ \lambda = -z - \frac{1}{4},$$

then the eigenvalue equation

$$\tilde{H}_m\psi = z\psi \tag{5}$$

takes the standard form of the differential equation of spheroidal functions:

$$(1-\xi^2)\frac{\partial^2\psi}{\partial\xi^2} - 2\xi\frac{\partial\psi}{\partial\xi} + \left[\lambda + 4\theta(1-\xi^2) - \mu^2(1-\xi^2)^{-1}\right]\psi = 0. \tag{6}$$

According to chapter 3.12, Satz 5 in [8], for $\mu = m \in \mathbb{N}_0$ a fundamental system $\{y_{\mathrm{I}},\ y_{\mathrm{II}}\}$ of solutions to equation (5) exists such that

$$y_{\mathrm{I}}(\xi) = (1-\xi)^{m/2}\,\mathfrak{P}_1(1-\xi), \quad \mathfrak{P}_1(0) = 1,$$
$$y_{\mathrm{II}}(\xi) = (1-\xi)^{-m/2}\mathfrak{P}_2(1-\xi) + A_m\,y_{\mathrm{I}}(\xi)\log(1-\xi),$$

where, for $|\xi-1|<2$, $\mathfrak{P}_1, \mathfrak{P}_2$ are analytic functions in ξ, λ, θ; and A_m is a polynomial in λ and θ of total order m with respect to λ and $\sqrt{\theta}$; $A_0 = -1/2$.

Suppose that $z \in \mathbb{C}\setminus\mathbb{R}$. For $m=0$, every solutions to (5) is square integrable near 1; while for $m \neq 0$, y_{I} is the only one solution, up to a factor, which is square integrable in a neighbourhood of 1. On the other hand, by a classical analysis due to Weyl, there exists exactly one linearly independent solution to (5) which is square integrable in a neighbourhood of ∞, see Theorem XIII.6.14 in [7]. In the case of $m=0$ this obviously implies that the defect indices are $(1,1)$. If $m \neq 0$ then, by Theorem XIII.2.30 in [7], the operator $\tilde{H}_m$ is essentially self-adjoint. □

Define the maximal operator associated to the formal differential expression

$$L = -\frac{\partial}{\partial\xi}\left((\xi^2-1)\frac{\partial}{\partial\xi}\right) + \frac{a^4\omega^2}{4}(\xi^2-1) - \frac{1}{4}$$

as follows

$$\begin{aligned}\mathrm{Dom}(H_{max}) = \Big\{ & f \in L^2((1,\infty),\mathrm{d}\xi) :\ f, f' \in AC((1,\infty)), \\ & -\frac{\partial}{\partial\xi}\left((\xi^2-1)\frac{\partial f}{\partial\xi}\right) + \frac{a^4\omega^2}{4}(\xi^2-1)f \in L^2((1,\infty),\mathrm{d}\xi)\Big\}, \\ H_{max}f = Lf.\end{aligned}$$

According to Theorem 8.22 in [10], $H_{max} = \tilde{H}_0^\dagger$.

Proposition 2. *Let $\kappa \in (-\infty,\infty]$. The operator $\tilde{H}_0(\kappa)$ defined by the formulae*

$$\mathrm{Dom}(\tilde{H}_0(\kappa)) = \{f \in \mathrm{Dom}(H_{max}) :\ f_1 = \kappa f_0\}\,,\ \tilde{H}_0(\kappa)f = H_{max}f,$$

where

$$f_0 := -4\pi a^2 \lim_{\xi\to 1+} \frac{f(\xi)}{\log(\xi-1)}\,,\ f_1 := \lim_{\xi\to 1+} f(\xi) + \frac{1}{4\pi a^2}\, f_0 \log(\xi-1),$$

is a self-adjoint extension of $\tilde{H}_0$. There are no other self-adjoint extensions of $\tilde{H}_0$.

Proof. The methods to treat δ like potentials are now well established [1]. Here we follow an approach described in [5], and we refer to this source also for the terminology and notations. Near the point $\xi = 1$, each $f \in \mathrm{Dom}(H_{max})$ has the asymptotic behaviour

$$f(\xi) = f_0\, F(\xi,1) + f_1 + o(1) \quad \text{as } \xi \to 1+$$

where $f_0, f_1 \in \mathbb{C}$ and $F(\xi,\xi')$ is the divergent part of the Green function for the Friedrichs extension of $\tilde{H}_0$. By formula (13) which is derived below, $F(\xi,1) = -1/(4\pi a^2)\log(\xi-1)$. Proposition 1.37 in [5] states that $(\mathbb{C},\Gamma_1,\Gamma_2)$, with $\Gamma_1 f = f_0$ and $\Gamma_2 f = f_1$, is a boundary triple for H_{max}.

According to theorem 1.12 in [5], there is a one-to-one correspondence between all self-adjoint linear relations κ in $\mathbb{C}$ and all self-adjoint extensions of $\tilde{H}_0$ given by $\kappa \longleftrightarrow \tilde{H}_0(\kappa)$ where $\tilde{H}_0(\kappa)$ is the restriction of H_{max} to the domain of vectors $f \in \mathrm{Dom}(H_{max})$ satisfying

$$(\Gamma_1 f, \Gamma_2 f) \in \kappa. \tag{7}$$

Every self-adjoint relation in $\mathbb{C}$ is of the form $\kappa = \mathbb{C}v \subset \mathbb{C}^2$ for some $v \in \mathbb{R}^2$, $v \neq 0$. If (with some abuse of notation) $v = (1,\kappa)$, $\kappa \in \mathbb{R}$, then relation (7) means that $f_1 = \kappa f_0$. If $v = (0,1)$ then (7) means that $f_0 = 0$ which may be identified with the case of $\kappa = \infty$, and then the corresponding self-adjoint extension is nothing but the Friedrichs extension. □

2.3 The Green function

Let us consider the Friedrichs extension of the operator $\tilde{H}$ in $L^2\left((1,\infty)\times S^1, \mathrm{d}\xi\,\mathrm{d}\phi\right)$ which was introduced in (4). The resulting self-adjoint operator is in fact the Hamiltonian for the impurity free case. The corresponding Green function $\mathcal{G}_z$ is the generalised kernel of the Hamiltonian, and it should obey the equation

$$(\tilde{H}-z)\mathcal{G}_z(\xi,\phi;\xi',\phi') = \delta(\xi-\xi')\delta(\phi-\phi') = \frac{1}{2\pi}\sum_{m=-\infty}^{\infty}\delta(\xi-\xi')\mathrm{e}^{im(\phi-\phi')}.$$

If we suppose $\mathcal{G}_z$ to be of the form

$$\mathcal{G}_z(\xi,\phi;\xi',\phi') = \frac{1}{2\pi}\sum_{m=-\infty}^{\infty}\mathcal{G}_z^m(\xi,\xi')\mathrm{e}^{im(\phi-\phi')}, \tag{8}$$

then, for all $m\in\mathbb{Z}$,

$$(\tilde{H}_m - z)\mathcal{G}_z^m(\xi,\xi') = \delta(\xi-\xi'). \tag{9}$$

Let us consider an arbitrary fixed ξ', and set

$$\mu = m,\ 4\theta = -\frac{a^4\omega^2}{4},\ \lambda = -z-\frac{1}{4}.$$

Then for all $\xi\neq\xi'$ equation (9) takes the standard form of the differential equation of spheroidal functions (6). As one can see from the following asymptotic formulae

$$\begin{aligned}
&S_\nu^{\mu(3)}(\xi,\theta) = \frac{1}{2}\,\theta^{-1/2}\xi^{-1}\mathrm{e}^{i(2\theta^{1/2}\xi-\nu\pi/2-\pi/2)}[1+O(|\xi|^{-1})],\\
&\text{for } -\pi<\arg(\theta^{1/2}\xi)<2\pi,\\
&S_\nu^{\mu(4)}(\xi,\theta) = \frac{1}{2}\,\theta^{-1/2}\xi^{-1}\mathrm{e}^{-i(2\theta^{1/2}\xi-\nu\pi/2-\pi/2)}[1+O(|\xi|^{-1})],\\
&\text{for } -2\pi<\arg(\theta^{1/2}\xi)<\pi,
\end{aligned} \tag{10}$$

the solution which is square integrable near infinity equals $S_\nu^{|m|(3)}(\xi,-a^4\omega^2/16)$. Furthermore, the solution which is square integrable near $\xi=1$ equals $Ps_\nu^{|m|}(\xi,-a^4\omega^2/16)$ as one may verify with the aid of the asymptotic formula

$$P_\nu^m(\xi)\sim\frac{\Gamma(\nu+m+1)}{2^{m/2}\,m!\,\Gamma(\nu-m+1)}\,(\xi-1)^{m/2}\quad\text{as }\xi\to1+,\ \text{for } m\in\mathbb{N}_0.$$

We conclude that the mth partial Green function equals

$$\mathcal{G}_z^m(\xi,\xi') = -\frac{1}{(\xi^2-1)\mathscr{W}(Ps_\nu^{|m|},S_\nu^{|m|(3)})}\,Ps_\nu^{|m|}\left(\xi_<,-\frac{a^4\omega^2}{16}\right)S_\nu^{|m|(3)}\left(\xi_>,-\frac{a^4\omega^2}{16}\right), \tag{11}$$

where the symbol $\mathscr{W}(Ps_\nu^{|m|},S_\nu^{|m|(3)})$ denotes the wronskian, and $\xi_<$, $\xi_>$ are respectively the smaller and the greater of ξ and ξ'. By the general Sturm-Liouville theory, the factor $(\xi^2-1)\mathscr{W}(Ps_\nu^{|m|},S_\nu^{|m|(3)})$ is constant. Since $\mathcal{G}_z^m=\mathcal{G}_z^{-m}$ decomposition (8) may be simplified,

$$\mathcal{G}_z(\xi,\phi;\xi',\phi') = \frac{1}{2\pi}\,\mathcal{G}_z^0(\xi,\xi') + \frac{1}{\pi}\sum_{m=1}^{\infty}\mathcal{G}_z^m(\xi,\xi')\cos\left[m(\phi-\phi')\right]. \tag{12}$$

2.4 The Krein Q-function

The Krein Q-function plays a crucial role in the spectral analysis of impurities. It is defined at a point of the configuration space as the regularised Green function evaluated at this point. Here we deal with the impurity located in the centre of the dot ($\xi = 1$, ϕ arbitrary), and so, by definition,

$$Q(z) := \mathcal{G}_z^{reg}(1,0;1,0).$$

Due to the rotational symmetry,

$$\mathcal{G}_z(\xi) := \mathcal{G}_z(\xi,\phi;1,0) = \mathcal{G}_z(\xi,\phi;1,\phi) = \mathcal{G}_z(\xi,0;1,0) = \frac{1}{2\pi}\,\mathcal{G}_z^0(\xi,1),$$

and hence

$$(\tilde{H}_0 - z)\mathcal{G}_z(\xi) = 0, \quad \text{for } \xi \in (1,\infty).$$

Let us note that from the explicit formula (11), one can deduce that the coefficients $\mathcal{G}_z^m(\xi,1)$ in the series in (12) vanish for $m = 1,2,3,\ldots$. The solution to this equation is

$$\mathcal{G}_z(\xi) \propto S_\nu^{0(3)}\left(\xi, -\frac{a^4\omega^2}{16}\right).$$

The constant of proportionality can be determined with the aid the following theorem which we reproduce from [6].

Theorem 3. *Let $d(x,y)$ denote the geodesic distance between points x, y of a two-dimensional manifold X of bounded geometry. Let*

$$U \in \mathcal{P}(X) := \left\{U :\ U_+ := \max(U,0) \in L^{p_0}_{loc}(X),\ U_- := \max(-U,0) \in \sum_{i=1}^{n} L^{p_i}(X)\right\}$$

for an arbitrary $n \in \mathbb{N}$ and $2 \le p_i \le \infty$. Then the Green function $\mathcal{G}_U$ of the Schrödinger operator $H_U = -\Delta_{LB} + U$ has the same on-diagonal singularity as that for the Laplace-Beltrami operator itself, i.e.,

$$\mathcal{G}_U(\zeta;x,y) = \frac{1}{2\pi}\log\frac{1}{d(x,y)} + \mathcal{G}_U^{reg}(\zeta;x,y)$$

where $\mathcal{G}_U^{reg}$ is continuous on $X \times X$.

Let us denote by $\mathcal{G}_z^H$ and $Q^H(z)$ the Green function and the Krein Q-function for the Friedrichs extension of H, respectively. Since $\tilde{H} = a^2 H$ and $(\tilde{H} - z)\mathcal{G}_z = \delta$, we have

$$\mathcal{G}_z^H(\xi,\phi;\xi',\phi') = a^2\mathcal{G}_{a^2z}(\xi,\phi;\xi',\phi'),\ Q^H(z) = a^2 Q(a^2 z).$$

One may verify that

$$\log d(\rho,0;\vec{0}) = \log\rho = \log(a\,\mathrm{arg\,cosh}\,\xi) = \frac{1}{2}\log(\xi-1) + \log(\sqrt{2}a) + O(\xi-1)$$

as $\rho \to 0+$ or, equivalently, $\xi \to 1+$. Finally, for the divergent part $F(\xi, \xi')$ of the Green function $\mathcal{G}_z$ we obtain the expression

$$F(\xi, 1) = -\frac{1}{4\pi a^2} \log(\xi - 1). \tag{13}$$

From the above discussion, it follows that the Krein Q-function depends on the coefficients α, β in the asymptotic expansion

$$S_\nu^{0(3)}\left(\xi, -\frac{a^4\omega^2}{16}\right) = \alpha \log(\xi - 1) + \beta + o(1) \quad \text{as } \xi \to 1+, \tag{14}$$

and equals

$$Q(z) = -\frac{\beta}{4\pi a^2 \alpha}\,. \tag{15}$$

To determine α, β we need relation

$$S_\nu^{0(3)} = \frac{1}{i\cos(\nu\pi)}\left(S_{-\nu-1}^{0(1)} + i\,\mathrm{e}^{-i\pi\nu} S_\nu^{0(1)}\right).$$

for the radial spheroidal function of the third kind. Formulae

$$\begin{aligned}
&S_\nu^{\mu(1)}(\xi, \theta) = \pi^{-1}\sin[(\nu - \mu)\pi]\mathrm{e}^{-i\pi(\nu+\mu+1)} K_\nu^\mu(\theta) Qs_{-\nu-1}^\mu(\xi, \theta),\\
&S_n^{m(1)}(\xi, \theta) = K_n^m(\theta) Ps_n^m(\xi, \theta),
\end{aligned}$$

imply that

$$\begin{aligned}
&S_\nu^{0(1)}(\xi, \theta) = \frac{\sin(\nu\pi)}{\pi}\,\mathrm{e}^{-i\pi(\nu+1)} K_\nu^0(\theta) Qs_{-\nu-1}^0(\xi, \theta),\\
&S_{-\nu-1}^{0(1)}(\xi, \theta) = \frac{\sin(\nu\pi)}{\pi}\,\mathrm{e}^{i\pi\nu} K_{-\nu-1}^0(\theta) Qs_\nu^0(\xi, \theta),\\
&S_n^{0(1)}(\xi, \theta) = K_n^0(\theta) Ps_n^0(\xi, \theta),\\
&S_{-n-1}^{0(1)}(\xi, \theta) = K_{-n-1}^0(\theta) Ps_{-n-1}^0(\xi, \theta).
\end{aligned} \tag{16}$$

Here $\nu \in \mathbb{C} \setminus \mathbb{Z},\ n \in \mathbb{Z}$.

Applying the symmetry relation for the expansion coefficients of the spheroidal functions (for those expansions see [2])

$$a_{\nu,r}^\mu(\theta) = a_{-\nu-1,-r}^\mu(\theta) = \frac{(\nu - \mu + 1)_{2r}}{(\nu + \mu + 1)_{2r}}\, a_{\nu,r}^{-\mu}(\theta),$$

we derive that

$$\begin{aligned}
Qs_{-\nu-1}^0(\xi, \theta) &= \sum_{r=-\infty}^{\infty} (-)^r a_{-\nu-1,r}^0(\theta) Q_{-\nu-1+2r}^0(\xi)\\
&= \sum_{r=-\infty}^{\infty} (-)^r a_{\nu,r}^0(\theta) Q_{-\nu-1-2r}^0(\xi).
\end{aligned}$$

Using the asymptotic formulae (see [2])

$$Q_\nu^0(\xi) = -\frac{1}{2}\log\frac{\xi-1}{2} + \Psi(1) - \Psi(\nu+1) + O\left((\xi-1)\log(\xi-1)\right),$$
$$P_n^0(\xi) = 1 + O\left((\xi-1)\right), \quad \text{as } \xi \to 1+,$$

the series expansions

$$Ps_\nu^\mu(\xi,\theta) = \sum_{r=-\infty}^{\infty} (-)^r a_{\nu,r}^\mu(\theta) P_{\nu+2r}^\mu(\xi),$$
$$Qs_\nu^\mu(\xi,\theta) = \sum_{r=-\infty}^{\infty} (-)^r a_{\nu,r}^\mu(\theta) Q_{\nu+2r}^\mu(\xi)$$

and formulae (16), we deduce that, as $\xi \to 1+$,

$$S_\nu^{0(1)}(\xi,\theta) \sim -\frac{\sin(\nu\pi)}{\pi}\,\mathrm{e}^{-i\pi(\nu+1)} K_\nu^0(\theta) \times \left[s_\nu^0(\theta)^{-1}\left(\frac{1}{2}\log\frac{\xi-1}{2} - \Psi(1) + \pi\cot(\nu\pi)\right) + \Psi s_\nu(\theta)\right],$$
$$S_{-\nu-1}^{0(1)}(\xi,\theta) \sim -\frac{\sin(\nu\pi)}{\pi}\,\mathrm{e}^{i\pi\nu} K_{-\nu-1}^0(\theta) \times \left[s_\nu^0(\theta)^{-1}\left(\frac{1}{2}\log\frac{\xi-1}{2} - \Psi(1)\right) + \Psi s_\nu(\theta)\right],$$
$$S_n^{0(1)}(\xi,\theta) \sim K_n^0(\theta) s_n^0(\theta)^{-1},$$
$$S_{-n-1}^{0(1)}(\xi,\theta) \sim K_{-n-1}^0(\theta) s_{-n-1}^0(\theta)^{-1} = K_{-n-1}^0(\theta) s_n^0(\theta)^{-1},$$

where the coefficients $s_n^\mu(\theta)$ stand for $s_\nu^\mu(\theta) = \left[\sum_{r=-\infty}^{\infty}(-1)^r a_{\nu,r}^\mu(\theta)\right]^{-1}$,

$$\Psi s_\nu(\theta) := \sum_{r=-\infty}^{\infty} (-)^r a_{\nu,r}^0(\theta)\Psi(\nu+1+2r),$$

and where we have made use of the following relation for the digamma function: $\Psi(-z) = \Psi(z+1) + \pi\cot(\pi z)$.

We conclude that

$$S_\nu^{0(3)}(\xi,\theta) \sim \alpha\log(\xi-1) + \beta + O\left((\xi-1)\log(\xi-1)\right) \quad \text{as } \xi \to 1+,$$

where

$$\alpha = \frac{i\tan(\nu\pi)}{2\pi s_\nu^0(\theta)}\left(\mathrm{e}^{i\pi\nu} K_{-\nu-1}^0(\theta) - \mathrm{e}^{-i\pi(2\nu+3/2)} K_\nu^0(\theta)\right),$$
$$\beta = \alpha\left(-\log 2 - 2\Psi(1) + 2\Psi s_\nu(\theta) s_\nu^0(\theta)\right) + \mathrm{e}^{-2i\pi\nu} s_\nu^0(\theta)^{-1} K_\nu^0(\theta).$$

For the integer values $\nu = n \in \mathbb{Z}$ it holds

$$S_n^{0(3)}(\xi,\theta) \sim s_n^0(\theta)^{-1}\left(K_n^0(\theta) - i(-)^n K_{-n-1}^0(\theta)\right) \quad \text{as } \xi \to 1+.$$

The substitution for α, β into (15) yields

$$Q(z) = -\frac{1}{4\pi a^2}\left(-\log 2 - 2\Psi(1) + 2\,\Psi s_\nu\left(-\frac{a^4\omega^2}{16}\right) s_\nu^0\left(-\frac{a^4\omega^2}{16}\right)\right) + \frac{1}{2a^2\tan(\nu\pi)}\left(\mathrm{e}^{i\pi(3\nu+3/2)}\,\frac{K^0_{-\nu-1}(-\frac{a^4\omega^2}{16})}{K^0_\nu(-\frac{a^4\omega^2}{16})} - 1\right)^{-1} \tag{17}$$

where ν is chosen so that

$$\lambda_\nu^0\left(-\frac{a^4\omega^2}{16}\right) = -z - \frac{1}{4}\,. \tag{18}$$

2.5 The spectrum of a quantum dot with impurity

The Green function of the Hamiltonian describing a quantum dot with impurity is given by the Krein resolvent formula

$$\mathcal{G}_z^{H(\chi)}(\xi,\phi;\xi',\phi') = \mathcal{G}_z^H(\xi,\phi;\xi',\phi') - \frac{1}{Q^H(z)-\chi}\,\mathcal{G}_z^H(\xi,0;1,0)\mathcal{G}_z^H(1,0;\xi',0).$$

The parameter $\chi := a^2\kappa \in (-\infty,\infty\,]$ determines the corresponding self-adjoint extension $H(\chi)$ of H. In the physical interpretation, this parameter is related to the strength of the δ interaction. Recall that the value $\chi = \infty$ corresponds to the Friedrichs extension of H representing the case with no impurity. This fact is also apparent from the Krein resolvent formula.

As is well known (see, for example, [3]), for the confinement potential tends to infinity as $\rho \to \infty$, the resolvent of $H(\infty)$ is compact and the spectrum of $H(\infty)$ is discrete. The same is also true for $H(\chi)$ for any $\chi \in \mathbb{R}$ since, by the Krein resolvent formula, the resolvents for $H(\chi)$ and $H(\infty)$ differ by a rank one operator. Moreover, the multiplicities of eigenvalues of $H(\chi)$ and $H(\infty)$ may differ at most by ± 1 (see [10, Section 8.3]).

A more detailed analysis given in [4] can be carried over to our case almost literally. Denote by σ the set of poles of the function $Q^H(z)$ depending on the spectral parameter z. Note that σ is a subset of $\mathrm{spec}(H(\infty))$. Consider the equation

$$Q^H(z) = \chi. \tag{19}$$

Theorem 4. *The spectrum of $H(\chi)$ is discrete and consists of four nonintersecting parts S_1, S_2, S_3, S_4 described as follows:*

1. *S_1 is the set of all solutions to equation (19) which do not belong to the spectrum of $H(\infty)$. The multiplicity of all these eigenvalues in the spectrum of $H(\chi)$ equals 1.*

2. *S_2 is the set of all $\lambda \in \sigma$ that are multiple eigenvalues of $H(\infty)$. If the multiplicity of such an eigenvalue λ in $\mathrm{spec}(H(\infty))$ equals k then its multiplicity in the spectrum of $H(\chi)$ equals $k-1$.*

3. *S_3 consists of all $\lambda \in \mathrm{spec}(H(\infty)) \setminus \sigma$ that are not solutions to equation (19). the multiplicities of such an eigenvalue λ in $\mathrm{spec}(H(\infty))$ and $\mathrm{spec}(H(\chi))$ are equal.*

4. *S_4 consists of all $\lambda \in \mathrm{spec}(H(\infty)) \setminus \sigma$ that are solutions to equation (19). If the multiplicity of such an eigenvalue λ in $\mathrm{spec}(H(\infty))$ equals k then its multiplicity in the spectrum of $H(\chi)$ equals $k+1$.*

Hence the eigenvalues of $H(\chi)$, $\chi \in \mathbb{R}$, different from those of the unperturbed Hamiltonian $H(\infty)$ are solutions to (19). As far as we see it, this equation can be solved only numerically. We have postponed a systematic numerical analysis of equation (19) to a subsequent work. Note that the Krein Q-function (17) is in fact a function of ν, and hence dependence (18) of the spectral parameter z on ν is fundamental.

3 Conclusion

We have proposed a Hamiltonian describing a quantum dot in the Lobachevsky plane to which we added an impurity modeled by a δ potential. Formulae for the corresponding Q- and Green functions have been derived. Further analysis of the energy spectrum may be accomplished for some concrete values of the involved parameters (by which we mean the curvature a and the oscillator frequency ω) with the aid of numerical methods.

References

[1] S. Albeverio, F. Gesztesy, R. Høegh-Krohn and H. Holden. *Solvable Models in Quantum Mechanics*. Springer-Verlag, (1988).

[2] H. Bateman and A. Erdélyi. *Higher Transcendental Functions III*. McGraw-Hill Book Company, (1955).

[3] F. A. Berezin, and M. A. Shubin. *The Schrödinger Equation*. Kluwer Academic Publishers, (1991).

[4] J. Brüning, V. Geyler, and I. Lobanov. *Spectral Properties of a Short-range Impurity in a Quantum Dot*. J. Math. Phys. **46** (2004), 1267-1290.

[5] J. Brüning, V. Geyler, and K. Pankrashkin. *Spectra of Self-adjoint Extensions and Applications to Solvable Schrödinger Operators*. arXiv:math-ph/0611088 (2007).

[6] J. Brüning, V. Geyler, and K. Pankrashkin. *On-diagonal Singularities of the Green Function for Schrödinger Operators*. J. Math. Phys. **46** (2005), 113508.

[7] N. Dunford and J.T. Schwartz. *Linear Operators. Part II: Spectral theory. Self Adjoint Operators in Hilbert Space*. Wiley-Interscience Publication, (1988).

[8] J. Meixner and F.V. Schäfke. *Mathieusche Funktionen und Sphäroidfunktionen*. Springer-Verlag, (1954).

[9] M. F. Rañada and M. Santader. *On Harmonic Oscillators on the Two-dimensional Sphere S^2 and the Hyperbolic Plane H^2*. J. Math. Phys. **43** (2002), 431-451.

[10] J. Weidman. *Linear Operators in Hilbert Spaces*. Springer, (1980).