ГОДИШНИК
НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“

Факултет по математика и информатика

ANNUAL
OF SOFIA UNIVERSITY “ST. KLIMENT OHRIDSKI”

Faculty of Mathematics and Informatics

Том/Volume 111

УНИВЕРСИТЕТСКО ИЗДАТЕЛСТВО „СВ. КЛИМЕНТ ОХРИДСКИ“
ST. KLIMENT OHRIDSKI UNIVERSITY PRESS
СОФИЯ • 2024 • SOFIA

Годишник на Софийския университет „Св. Климент Охридски“
Факултет по математика и информатика

Annual of Sofia University "St. Kliment Ohridski"
Faculty of Mathematics and Informatics

Address for correspondence

Faculty of Mathematics and Informatics
Sofia University "St. Kliment Ohridski"
5 J. Bourchier Blvd., POB 48
1164 Sofia, BULGARIA
email: annuaire@fmi.uni-sofia.bg

Aims and Scope. The *Annual* is the oldest Bulgarian journal, founded in 1904, devoted to pure and applied mathematics, mechanics and computer science. It is reviewed by *Zentralblatt für Mathematik*, *Mathematical Reviews* and the Russian *Referativnii Jurnal*. The *Annual* publishes significant and original research papers of authors both from Bulgaria and abroad in some selected areas that comply with the traditional scientific interests of the Faculty of Mathematics and Informatics at the Sofia University "St. Kliment Ohridski", i.e., algebra, geometry and topology, analysis, probability and statistics, mathematical logic, theory of approximations, numerical methods, computer science, classical, fluid and solid mechanics, and their fundamental applications.

ISSN 1313–9215 (Print)
ISSN 2603–5529 (Online)

CONTENTS

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“
ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Том 111

ANNUAL OF SOFIA UNIVERSITY "ST. KLIMENT OHRIDSKI"
FACULTY OF MATHEMATICS AND INFORMATICS
Volume 111

IN MEMORIAM
IORDAN IORDANOV (1948–2021)

NIKOLAY CHERVENOV, VLADIMIR GEORGIEV AND ANDREY VASSILEV

Iordan Iordanov passed away unexpectedly on October 3, 2021. His loss is deeply felt by the academic community and friends around the world. The current work provides an overview of his academic career and achievements and surveys his research in mathematics, economics and finance.

Keywords: Iordan Iordanov, variational inequalities, partial differential equations, mathematical modelling in economics, mathematical finance and insurance, history of Bulgarian mathematics

2020 Mathematics Subject Classification: 01A70, 35J88, 35L67, 91B64, 91B72, 62H05, 91Gxx

DOI: 10.60063/GSU.FMI.111.5-24

1. Introduction: Iordan Iordanov's life and work

Iordan Iordanov suddenly passed away on October 3, 2021. Bulgarian academia lost a dedicated educator, passionate researcher, erudite scholar and dear friend. His absence will be especially felt in the community specializing in mathematical modelling applied to economics and finance, as Iordan was one of the pioneers of the field in post-communist Bulgaria.

Iordan Iordanov was born on February 27, 1948 in Sofia. He grew up in the town of Sapareva Bania and, after graduating from high school in Stanke Dimitrov (present-day Dupnitsa), went on to study mathematics at Sofia University "St. Kliment Ohridski". There he was noticed by Prof. Todor Genchev, who encouraged his interest in the field of partial differential equations. Iordan was actively involved in academic life even before graduating in mathematics in 1971, teaching seminars in his student days. Following his graduation, he assumed a teaching position in the Department of Differential Equations of the Faculty of Mathematics and Informatics at Sofia University.

Iordan defended his PhD dissertation titled "Variational inequalities for operators in partial derivatives" in 1980 and obtained habilitation as associate professor at the Faculty of Mathematics and Informatics of Sofia University "St. Kliment Ohridski" in 1983. At the time his research focused on variational inequalities and analysis of singularities of solutions to certain classes of partial differential equations. More details on his mathematical works are provided in Section 2.

As a teacher of mathematics, Iordan had a reputation of being stern, yet fully devoted to passing on the science and the craft. He taught a number of courses, including on partial differential equations, mathematical analysis for students in Physics and theory of distributions. He was noted for always striving to make accessible even the most complex concepts and seeking to present the basics of the respective subjects in line with modern practices and approaches. One of the authors of the present work, Vladimir Georgiev, recollects the following:

> Iordan had a really well-prepared and attractive course on Partial Differential Equations. Typically, this course was not considered very easy by the students. However, Iordan succeeded to attract several students. Among them was Angel Ivanov, who did a master's thesis with Iordan Iordanov and then won a PhD position at University of Pisa. This was the first Bulgarian PhD student at University of Pisa. The role of Iordan in the preparation of Angel as a PhD student was really crucial.

In the early 1990s Iordan Iordanov embarked on a new journey. He was one of the group of enthusiasts who with a lot of effort and energy re-established the Faculty of Economics and Business Administration at Sofia University in 1990. He was very active both academically and administratively, serving as vice-dean and acting dean of the faculty during various periods.

During the early years of the building of the Faculty of Economics and Business Administration Iordan made key contributions to the creation of the curricula of the two mathematics courses and the quantitative methods course taught at the faculty, in line with state-of-the-art practices in teaching these disciplines. He is the main author of the textbook that is still used in the teaching of mathematics at the Faculty of Economics and Business Administration. Iordan contributed to assembling and expanding the team of lecturers at the faculty and to introducing French-language programmes. He actively supported inter-university cooperation through student and lecturer exchanges.

In the 2000s Iordan redirected his efforts to a new project – the development of a master's programme on mathematical modelling in economics at the Faculty of Mathematics and Informatics. Together with a team of collaborators he researched best practices in the teaching of mathematical modelling applied to economics. He tirelessly worked to put together a curriculum that is appropriate for the audience at the Faculty of Mathematics and Informatics, as well as to ensure there is a team of lecturers providing the teaching. The program was launched successfully with Iordan as its first director. It continues to run to this day.

As a fellow academic, Iordan Iordanov was extremely cooperative and supportive. He was always ready to discuss, provide hints or materials, explain the finer points of an intricate issue and generally spare the time to delve into a problem and push it forward. His colleagues, students and collaborators frequently noted his generosity, meticulousness and drive to attain useful, high-quality results. These qualities carried over to organisational and administrative tasks, where he was hardworking, result-oriented and always caring and respectful to people.

This is what one of the authors, Nikolay Chervenov, shares about working with Iordanov:

> I met Iordanov for the first time in Vasil Tsanov's office, when I was a master's student in Equations of Mathematical Physics. He needed a student to assist him with the seminars for his courses. We connected right away, and with his impressive teaching flair, I soon became his assistant for Calculus, Differential Equations and Partial Differential Equations, Complex analysis. Two years later, I naturally chose him as my doctoral supervisor, as we had already established a good rapport.
>
> Copula functions, which were the core of most of our joint works, caught his interest after he saw Prof. Leda Minkova's open notes. He was drawn to the expression of the copula density as its mixed partial derivative. As a differential equations expert, he quickly wondered about the existence of this derivative and how to build a family of copulas as a solution to differential equations. A result that would be completely new in the field of copulas.

Iordan Iordanov did not confine his efforts to popularise mathematical modelling in economics to Sofia University only. He taught various courses at the Technical University of Sofia and wrote textbooks on macroeconomics and mathematical models in economics ([43, 44]). Iordan also published a number of papers on various economic and finance issues, including on the methodology of teaching economics through a quantitative lens (as detailed in Sections 3 and 4). He consulted for various institutions, including the Bulgarian National Bank, the Agency for Economic Analysis and Forecasting, as well as private entities.

After formally retiring from Sofia University, Iordan never lost momentum. He continued teaching at the Faculty of Mathematics and Informatics and at the Technical University of Sofia. He also assumed a position at the International Business School, where he contributed to the curriculum on mathematics and quantitative methods.

Iordan's interests and erudition extended far beyond mathematics and economic modelling. He spoke several foreign languages, dabbled in psychiatry and health practices, and was surprisingly skilled at gardening and different crafts. He was an avid lover of travel and nature, never missing an opportunity to experience new sights and places.

2. Mathematical contributions

Iordan's initial research interests included the regularity of solutions to some elliptic variational problems. The works [6–12] treat different aspects of this topic.

Typically, the problem is associated with a bounded domain $\Omega \subset \mathbb{R}^n$. The elliptic operator of type

$$Lv \equiv -\sum_{i,j=1}^{n} \left(a_{ij}(x)v_{x_i}\right)_{x_j} + c(x)v$$

defines the bilinear form

$$a(u,v) \equiv \int_\Omega \left(\sum_{i,j=1}^{n} a_{ij} u_{x_i} v_{x_j} \right) \mathrm{d}x + \int_\Omega cuv \,\mathrm{d}x. \tag{2.1}$$

A classical result of Lions and Stampacchia [37] treats a generic quadratic form a in a Hilbert space H. In the case of one convex set K in H the result in this article gives a unique element $u \in K$, such that

$$a(u, v-u) \geq 0 \quad \text{for any} \quad v \in K.$$

As second important point in his study Iordan chose the comparison approach proposed in the work of Lewy and Stampacchia [36], where the following comparison notion is introduced.

Let E be a closed set of $\overline{\Omega}$. An element u of $H^1(\Omega)$ will be said to be nonnegative on E in the sense of $H^1(\Omega)$ if there exists a sequence $\{u_m\}$ of functions of $C^1(\overline{\Omega})$ such that

(i) $u_m \geq 0$ on E,

(ii) $u_m \to u$ in $H^1(\Omega)$ as $m \to \infty$.

More generally, we say that, for $u \in H^1(\Omega)$, $v \in H^1(\Omega)$, $u \geq v$ on E in the sense of $H^1(\Omega)$ if $u - v \geq 0$ on E in the sense of $H^1(\Omega)$.

In the previous definition condition (ii) may be replaced by (ii)′ $u_m \to u$ weakly in $H^1(\Omega)$. In fact, if there exists a sequence $\{u_m\}$ of $C^1(\overline{\Omega})$ functions satisfying (i) and (ii)′, then, by a well-known theorem of Banach-Saks, there exists a sequence $\{u'_m\}$ of positively weighted means of $\{u_m\}$ which satisfies conditions (i) and (ii).

We turn to the quadratic form $a(u, v)$ defined in (2.1) for $u, v \in H^1(\Omega)$. In a series of works [6–12] Iordan studied appropriate generalization of existence and uniqueness of solutions of variational inequalities for the case of two closed subsets E and F in Ω with nonzero measures.

More precisely, given any $\psi \in H^1_0(\Omega)$ one can define

$$N = \sup_E \psi, \quad M = \inf_F \psi. \tag{2.2}$$

Iordan used the following variant of the Stampacchia comparison definition. One can say that $v, u \in H^1_0(\Omega)$ satisfy the relation $v \geq u$ in sense of $H^1_0(\Omega)$ in some subsets in Ω, if there exist sequences $\{v_n\}$ and $\{u_n\}$ approximating v and u in $H^1(\Omega)$ and for all n satisfying $v_n \geq u_n$ in the chosen subset. Then one defines the following convex subsets in $H^1_0(\Omega)$:

$$K_1 = \left\{v \mid v \in H^1_0(\Omega),\ v \geq \psi \text{ in } E \text{ in the sense of } H^1_0(\Omega)\right\}, \tag{2.3}$$

$$K_2 = \left\{v \mid v \in H^1_0(\Omega),\ v \leq \psi \text{ in } F \text{ in the sense of } H^1_0(\Omega)\right\}. \tag{2.4}$$

In this way Iordan faced a situation different from the classical case of one convex set K in a Hilbert space H analysed in [37]. In fact, in his works Iordan generalized this result for the case of two convex sets K_1, K_2.

The results obtained in the works of Iordan give existence and uniqueness of solution $u \in K_2$ to the following variational inequality

$$a(u, v - u) \geq 0 \quad \text{for any} \quad v \in K_2. \tag{2.5}$$

The main difficulty is the control of the approximation sequence for the variational inequality (2.5) near $E \cap F$ and the weak regularity of the boundary of this set. In a series of lemmas Iordan succeeded to define the approximating sequence and with appropriate extension outside the boundary of $E \cap F$.

The case when the Sobolev space is $H^{1,p}$ type is considered in [12–14, 38]. Concerning the regularity of the solution Iordan studied and applied the method developed by G. Stampacchia in [41].

Iordan obtained the following important result.

Theorem 2.1. *Assume the boundary $\partial\Omega$ of the domain Ω is in the class $C^{1,1}$ and that $\varphi, \psi_1 \in H^{1,p}(\Omega)$ with $p > n$. We assume further that*

$$\varphi < \psi \text{ on } \Gamma = \overline{\Omega \cap \partial\left(E \cap F\right)}$$

and $\Gamma \neq \varnothing$. *We assume further that the boundaries* $\partial\left(\overline{\Omega} \setminus E\right)$ *and* $\partial\left(\overline{\Omega} \setminus F\right)$ *are Lipschitz continuous and that the coefficients* $a_{ij} = a_{ji}$, $i, j = 1, \ldots, n$, *are Lipschitz continuous in* $\overline{\Omega}$.

Then the solution u *to the problem* (2.5) *is in the space*

$$C^{0,\alpha}(\overline{\Omega}), \quad \alpha \in \left(0, 1 - \frac{n}{p}\right].$$

This result and its generalizations can be found in [14].

Furthermore, Iordan Iordanov was interested in the problem of creation of singularities for some non-strictly hyperbolic systems. At this time he was discussing often the propagation of singularities of solutions of PDE's with Vesselin Petkov.

The work [16] deals with the Cauchy problem for the semilinear non-strictly hyperbolic system of type

$$\left(\partial_t + \lambda_j(x,t)\partial_x\right) u_j = g_j\left(x, t, u_1, \ldots, u_n\right), \quad j = 1, \cdots, n, \tag{2.6}$$

where

$$\lambda_j \in C^\infty\left(\mathbb{R}^2\right), \quad g_j \in C^\infty\left(\mathbb{R}^{n+2}\right).$$

At $t = 0$ the first speed velocity coefficient satisfies

$$\lambda_1 = t^{2q-1}\alpha(x,t), \quad \alpha \in C^\infty\left(\mathbb{R}^2\right), \quad \alpha(x,t) \geq \alpha_1 = \text{const} > 0, \quad q \in \mathbb{N}.$$

The second term λ_2 is identically 0. For other speed velocities the assumption of non-degeneracy is the following one

$$\inf_{\substack{i,j=1,\ldots,n \\ i<j,\,(i,j)\neq(1,2)}} \left|\lambda_i(x,t) - \lambda_j(x,t)\right| \geq \delta = \text{const} > 0.$$

Iordan studied the propagation of anomalous singularities; these can exist along the forward characteristics from the points of singular support of the initial data.

The interaction of the singularities propagating along two characteristics C_1, C_2 crossing at some point could give rise of new singularities (the so-called anomalous singularities) propagating along the outgoing transversal characteristics starting from that point. The solution is supposed smooth along the corresponding backward transversal characteristics. The anomalous singularities are weaker than the initial ones and in the case of weakly hyperbolic semilinear systems the larger the order of contacts between the characteristics C_1, C_2 carrying out the initial singularities, the more regular the solutions.

In the works [16, 39] an example of a weakly hyperbolic system is studied. In a sufficiently small neighbourhood of the origin in $\mathbb{R}^2_{(x,t)}$ the following weakly hyperbolic system is considered:

$$\begin{aligned}
&Xu \equiv \left(\partial_t + pt^{p-1}\partial_x\right) u = 0, \\
&\partial_t v = u, \\
&Dz \equiv \left(\partial_t + \partial_x\right) z = f(u,v),
\end{aligned}$$

where p is a positive integer and the initial data are given at $t=-T$, $T>0$,

$$u(x,-T)=u_0(x), \quad v(x,-T)\equiv z(x,-T)\equiv 0.$$

The characteristics $x=t^p+\text{const}$ of X and $x=\text{const}$ of ∂_t have order of contact $p-1$ at points on the line $t=0$ and the characteristics of $D(x=t+\text{const})$ are transversal to them. Denoting by C_i, $i=1,2,3$, the characteristics $x=t^p$, $x=0$ and $x=t$, passing through the origin $(0,0)$ one can define C_i^+ as the part of C_i for $t>0$. It is assumed that the function $u_0(x)$ has a jump discontinuity of order k ($k\geq 0$ integer) at $x_0=T^p$.

Iordan valued highly working with young researchers. The work of Angel Ivanov at University of Pisa referenced in Section 1 was done practically in collaboration with Iordan. As a result Angel had two important publications [4,5].

Consider the following Schrödinger equation with potential perturbation:

$$i\partial_t u-\Delta u+Vu=0, \quad \Delta=\partial_{x_1}^2+\partial_{x_2}^2+\partial_{x_3}^2, \tag{2.7}$$

$$u(0,x)=u_0(x), \quad x\in\mathbb{R}^3. \tag{2.8}$$

Here $V=V(x)$ is a real-valued potential that satisfies the assumption

$$\|V\|_{L^{(3/2,\infty)}}\leq \delta_0, \tag{2.9}$$

where $L^{(p,q)}$ are standard Lorentz spaces, $L^{(p,\infty)}$ is the weak L^p space. For $\delta_0>0$ sufficiently small one can define the bilinear form

$$Q(u,v)=(\nabla u,\nabla v)_{L^2(\mathbb{R}^3)}+\int_{\mathbb{R}^3}V(x)u(x)\overline{v(x)}\,\mathrm{d}x \tag{2.10}$$

on $\dot{H}^1(\mathbb{R}^3)$. The Friedrichs extension of the quadratic form, implies that $-\Delta+V$ has dense domain $D=H^2(\mathbb{R}^3)$ such that, if $f\in D$, then $(-\Delta+V)f\in L^2$ and $Vf\in L^2$. Moreover, $-\Delta+V$ with dense domain D is a self-adjoint operator. The existence and mapping properties of the wave operators

$$\Omega_\pm=s-\lim_{t\to\pm\infty}e^{itH}e^{-itH_0}, \tag{2.11}$$

where $H_0=-\Delta$, $H=-\Delta+V$, are presented in [33]. The crucial assumption is that V belongs to the Kato class (see, for example, [33]). It is easy to see that potential of type

$$V(x)=\frac{W\left(\frac{x}{|x|}\right)}{|x|^2}, \tag{2.12}$$

where $W\in L^\infty(S^2)$, does not belong to the Kato's class (for example we can take $W(x/|x|)=1$).

The main goal of this line of work was to prove the existence and mapping properties of the wave operator $\Omega_\pm$ for the larger class of potentials V, satisfying only (2.9). This problem is closely connected with another classical problem, namely the Strichartz-type estimates for the corresponding inhomogeneous Cauchy problem

$$i\partial_t u - \Delta u = F, \quad u(0) = f. \tag{2.13}$$

We shall call the pair $\left(\frac{1}{p}, \frac{1}{q}\right)$ sharp admissible (see [35] for this notion and the properties of sharp admissible pairs), if it satisfies the condition

$$\frac{3}{4} = \frac{1}{p} + \frac{3}{2q}, \quad 2 \le p \le \infty. \tag{2.14}$$

The main result is given in the next theorem.

Theorem 2.2. *If (p,q) and $(\tilde{p},\tilde{q})$ satisfy (2.14), then the solution to the Cauchy problem*

$$\begin{gathered} i\partial_t u - \Delta u + Vu = F, \quad (t,x) \in \mathbb{R}_t \times \mathbb{R}^3_x, \\ u(0,x) = f(x), \end{gathered} \tag{2.15}$$

satisfies the estimate

$$\|u\|_{L^p(\mathbb{R}_t;L^{(q,2)}_x)} + \|u\|_{C(\mathbb{R}_t;L^2)} \le C(\|F\|_{L^{\tilde{p}'}(\mathbb{R}_t;L^{(\tilde{q}',2)}_x)} + \|f\|_{L^2}), \tag{2.16}$$

where $\tilde{q}'$ is such that $1/\tilde{q} + 1/\tilde{q}' = 1$.

Another direction of research, where Iordan also participated together with Angel Ivanov and Vladimir Georgiev, concerned the two-dimensional equivariant wave map problem with inhomogeneous source term in special domains. The initial results are presented in [4]. The result shows that if the initial data are in a better space than the natural energy space, then the corresponding norm of the solution blows up as t goes to 0.

After the PhD thesis defence of Angel Ivanov in Pisa, he returned to Sofia, Bulgaria, and took a position at Technical University of Sofia. The work on Strichartz estimates for potential perturbations was continued.

The new project started with Iordan and Angel was really challenging. More precisely, Iordan, Angel and Vladimir Georgiev started to study the Laplace operator perturbed by a potential $V = V(x)$, $x \in \mathbb{R}^n$

$$\Delta_V = \sum_{j=1}^n \partial_j \partial_j - V = \sum_{j=1}^n \partial_{x_j} \partial_{x_j} - V \tag{2.17}$$

and their goal was to study the dispersive properties of the corresponding wave equation

$$\begin{cases} \partial_{tt} u - \Delta_V u = F, \quad t \in \mathbb{R}, \quad x \in \mathbb{R}^n \\ u(0,x) = f_0(x), \ \partial_t u(0,x) = f_1(x). \end{cases} \tag{2.18}$$

The main assumption concerning the potential V is the following one.

Assumption 2.1. *There exists $C > 0$ and p, $2n/3 \le p \le n$, such that we have*

$$\sum_{k\in\mathbb{Z}} 2^{2k-nk/p} \|V(x)\|_{L^p_{\{|x|\sim 2^k\}}} \le C. \tag{2.19}$$

The following estimate was obtained.

Theorem 2.3. *If $n \ge 4$, then for any potential V satisfying* Assumption 2.1 *there exists $C > 0$, such that for any $f_0, f_1 \in S(\mathbb{R}^n)$ and any $F(t,x) \in S(\mathbb{R}\times\mathbb{R}^n)$ the solution $u(t,x)$ to (2.18) satisfies the endpoint estimate*

$$\|u\|_{L^2_t L^{p_0}_x} \le C\|f_0\|_{\dot H^\gamma_x} + C\|f_1\|_{\dot H^{\gamma-1}_x} + C\|F\|_{L^2_t L^{p_0'}_x}, \tag{2.20}$$

where $p_0 = 2(n-1)/(n-3)$ and

$$\gamma = \frac{n-1}{2} - \frac{n}{p_0} = \frac{n+1}{2(n-1)}.$$

The work was not completed, since after an incident in Sofia Angel passed away in 2008. Angel's demise took away unexpectedly a young and very promising researcher, who was one of the key members of the group working on dispersive PDE at Technical University. His colleagues and friends, among which Iordan Iordanov, George Venkov, Mirko Tarulli and Vladimir Georgiev, were deeply saddened and shocked by the death of Angel.

Over the next years (between 2008 and 2019) Iordan and Vladimir Georgiev had various contacts on research topics. The focus of discussion was mathematical models in finance and especially the use of copulas in these models.

The above shows the extremely creative approach of Iordan, his deep mathematical preparation and ability to work with Master and PhD students, and stimulate in this way the progress and development of mathematical research in Bulgaria.

3. Economic modelling

Iordan's work on re-establishing the Faculty of Economics and Business Administration at Sofia University naturally piqued his curiosity about the mathematical toolkit used in economic modelling. This was initially motivated by attempts to build cutting-edge curricula in mathematics and quantitative methods for economists but soon spilled over to research questions. As an expert in partial differential equations, Iordan was underwhelmed by their use in the economic sciences. He was well aware of developments around the Black-Scholes option pricing approach but found the use of PDE tools there almost accidental, a by-product of a different construct. His starting point was that the natural use of PDEs would be to model phenomena evolving through both space and time. This stimulated his interest in spatial economics.

It quickly became apparent that spatial economics, while conceptually closer to Iordan's views, does not abound with examples of applications of PDEs. Instead, a central theme of the spatial economics literature in the late 1990s and early 2000s was the analysis of models featuring transportation costs and potentially other costs arising from geographical separation. This type of model inspired Iordan's early work in spatial economics in collaboration with Stoyan Stoyanov and Andrey Vassilev.

3.1. Strategic trade models

The first contributions to the analysis of spatial models came in the form of two linked papers [20,42]. They study the adjustment of prices towards an equilibrium in a model of trade between two regions, respectively in discrete-time and in continuous-time settings. Trade takes place under conditions of strategic interactions, taking as given output, financial resources and transportation costs.

More precisely, the consumer in region I (consumer I) exogenously receives money income $Y_1 > 0$ in each period and so does the consumer in region II (consumer II), who receives money income $Y_2 > 0$. For each period t, in region i a fixed quantity $q_i > 0$ of a certain good is supplied at a price $p_{i,t}$. The consumers place orders for the desired quantities in each region, observing their budget constraints and incurring symmetric transportation costs $\rho > 0$ per unit of shipment from the "foreign" region.

For each period t the above situation is modelled as a static noncooperative game of complete information. The orders placed by consumer I in regions I and II, respectively, are denoted by α and β. Similarly, γ and δ stand for the orders of consumer II in regions I and II. In period t consumer I's strategy space S_1 is determined by the budget constraint and the nonnegativity restrictions on the orders:

$$S_1 = \{\alpha p_{1,t} + \beta(p_{2,t} + \rho) \leq Y_1,\ \alpha, \beta \geq 0\}. \tag{3.1}$$

Consumer II's strategy space in period t is

$$S_2 = \{\gamma(p_{1,t} + \rho) + \delta p_{2,t} \leq Y_2,\ \gamma, \delta \geq 0\}. \tag{3.2}$$

The payoff function for consumer I is given by

$$P_1(\alpha, \beta, \gamma, \delta) = \min(\alpha, q_1) + \min(\beta, q_2 - \min(\delta, q_2)) \tag{3.3}$$

and that for consumer II by

$$P_2(\alpha, \beta, \gamma, \delta) = \min(\gamma, q_1 - \min(\alpha, q_1)) + \min(\delta, q_2). \tag{3.4}$$

The motivation behind the above formulation of the payoff functions comes from a restriction in market access taking the form of full protection for the local consumer. More specifically, when total orders for the respective region exceed the quantity available, the order of the local consumer is executed first to the extent possible and then the remaining quantity, if any, is allocated to the consumer from the other region.

In discrete time, a price is adjusted downwards if the quantity available in the respective region has not been entirely consumed. A price is adjusted upwards if a part of the orders placed in the respective region has not been satisfied. In particular, prices evolve according to the equation

$$\frac{q_i - q_i^{\text{cons}}}{q_i} = -\frac{p_{i,t+1} - p_{i,t}}{p_{i,t}} \quad \text{or} \quad p_{i,t+1} q_i = p_{i,t} q_i^{\text{cons}} \tag{3.5}$$

when $q_i > q_i^{\text{cons}}$ and according to

$$\frac{Y_i^{\text{res}} - p_{i,t} q_i}{p_{i,t} q_i} = \frac{p_{i,t+1} - p_{i,t}}{p_{i,t}} \tag{3.6}$$

when $Y_i^{\text{res}} > p_{i,t} q_i$ with q_i^{cons} denoting the total amount consumed in region i and Y_i^{res} standing for the part of the region i's income left unspent.

In continuous time the counterpart of equations (3.5) and (3.6) is given by the differential equation system

$$\frac{\dot{p}_i}{p_i} = -\frac{q_i - q_i^{\text{cons}}}{q_i} + \frac{Y_i - Y_i^{\text{cons}}}{p_i q_i}, \quad i = 1, 2. \tag{3.7}$$

For the above setup, the existence of Nash equilibria is proved and selection rules that pinpoint a unique equilibrium are formulated for edge cases. The Nash equilibria in the model are classified based on the relations among quantities produced and financial resources. The different price paths that can be obtained are also analysed and classified to establish the various possibilities of obtaining a price equilibrium. It turns out that in certain cases the laws governing price dynamics in discrete time lead to a zero price in one of the regions, which can be interpreted as a breakdown of economic activity in the region. This pathology is ruled out in the case of continuous-time price dynamics.

The model from [20] and [42] is generalised in [28, 30, 31]. The main novelty is that local consumers enjoy only partial protection in terms of preferential access to the good on their home market. Additionally, the formulation is general and dispenses with certain symmetry assumptions, allowing for differences between regions in terms of parameters such as income, size of the local market supply, degree of protection and transportation costs.

The partial protection framework is formalised through the following payoff function for consumer 1:

$$\begin{aligned} P_1(\alpha, \beta, \gamma, \delta) = \min(\beta, (1 - \epsilon_2) q_2 + \max(0, \epsilon_2 q_2 - \delta)) \\ + \min(\alpha, \max(q_1 - \gamma, \min(\alpha, \epsilon_1 q_1))). \end{aligned} \tag{3.8}$$

Similarly, the payoff function for consumer 2 is

$$\begin{aligned} P_2(\alpha, \beta, \gamma, \delta) = \min(\gamma, (1 - \epsilon_1) q_1 + \max(0, \epsilon_1 q_1 - \alpha)) \\ + \min(\delta, \max(q_2 - \beta, \min(\delta, \epsilon_2 q_2))). \end{aligned} \tag{3.9}$$

The interpretation of (3.8) and (3.9) is that a fixed share $\epsilon_1 \in (0, 1]$ of the quantity q_1 is reserved for consumer 1 and, similarly, a share $\epsilon_2 \in (0, 1]$ of the quantity q_2 is preferentially available to consumer 2. Consumers have the right to buy the respective quantities $\epsilon_1 q_1$ and $\epsilon_2 q_2$ but are not obliged to do so. After the local consumer buys a part or all of the preferentially available quantity, the remaining quantity of the good is offered to the foreign consumer. In turn, the foreign consumer can purchase part or all of this remainder and, if there is anything left, it is again offered to the local consumer.

In the extended model price dynamics are studied in discrete time and consequently several versions of (3.5) and (3.6) are used to describe the price changes over time. For this model, the existence of Nash equilibria is proved and a classification of the players' best-reply functions and the obtained Nash equilibria is presented. Price dynamics are studied by means of numerical simulations. The results show that different price adjustment rules used in conjunction with different parameter sets can produce very diverse outcomes. Examples range from trivial one-period "jumps" of prices to a steady state value to much less regular behaviour such as transitional dynamics settling down on a steady state value or cyclical behaviour. Overall, the specification of the price adjustment rules is crucial for the type of dynamics obtained and the interpretation of the results.

3.2. Spatial relocation models

Iordan's quest for an economic model with "genuine space" continued and an initial formulation appeared around 2006. The model considered is in essence microeconomic and studies the dynamic consumption-saving behaviour of an individual who is allowed to move in a suitable economic space in search of better job opportunities. The introduction of an explicit economic space that provides an abstraction for the physical, geographic space where agents operate was a first step in an agenda aimed at constructing an aggregate description of an economy with both spatial and temporal dimensions.

One version of the formulation is the following [22,23]. There is a consumer who, given an initial location in space x_0 and asset level a_0, supplies inelastically a unit of labour in exchange for a location-dependent wage $w(x(t))$, and chooses consumption $c(t)$ and spatial location $x(t)$ over time. The consumer has a finite lifetime T at the end of which a bequest in the form of assets is left. This bequest provides utility to the consumer. More precisely, for ρ, r, η, ξ and p – positive constants, and $\theta \in (0,1)$, the consumer's choices are described by the optimal control problem of maximizing with respect to $c(t), z(t) \in \Delta$ the objective functional

$$J(c(t), z(t)) := \int_0^T e^{-\rho t} \left(\frac{c(t)^{1-\theta}}{1-\theta} - \eta z^2(t) \right) \mathrm{d}t + e^{-\rho T} \frac{a(T)^{1-\theta}}{1-\theta} \tag{3.10}$$

subject to

$$\dot{a}(t) = ra(t) + w(x(t)) - pc(t) - \xi z^2(t), \tag{3.11}$$

$$\dot{x}(t) = z(t), \tag{3.12}$$

$$a(0) = a_0 \geq 0, \quad x(0) = x_0 \in [0,1],$$

where $a(t)$, $x(t)$ are the state variables, assumed to be absolutely continuous, and $c(t)$, $z(t)$ are the control variables. The set of admissible controls Δ consists of all pairs of functions $(c(t), z(t))$ which are measurable in $[0, T]$ and satisfy the conditions

$$0 \leq c(t) \leq C, \tag{3.13}$$

$$|z(t)| \leq Z, \tag{3.14}$$

$$a(T) \geq 0. \tag{3.15}$$

The constants C and Z satisfy appropriately defined conditions that are technical in nature but do not affect the economic analysis of the problem. The constant $\rho > 0$ is a time discount parameter and $\theta \in (0, 1)$ is the utility function parameter. The control $c(t)$ represents physical units of consumption and the control $z(t)$ governs the speed of relocation in space. Relocation in space brings about two type of consequences: subjective disutility due to habit formation with respect to the current location and monetary relocation costs. These are captured through the speed of movement in space $\dot{x}(t)$ or, equivalently, $z(t)$, transformed through a quadratic function. The parameters $\eta, \xi \geq 0$ multiplying this function measure respectively the subjective disutility from changing one's location in space and the relocation costs in monetary terms. The parameters $p > 0$ and $r > 0$ stand for the price of a unit of consumption and the interest rate, respectively.

The above optimal control problem presents a mathematical challenge, since its set of generalised speeds is not convex, which is a key element of standard proofs of existence of solutions. This situation requires a different approach to show existence, which is done in [21] for a generalised version of the problem (3.10)–(3.15) that dispenses with explicit functional forms and works with general convexity and concavity assumptions instead.

Further analysis looks at the basic case where economic space is represented by the real line or, more specifically, a subset of it, the interval $[0, 1]$. This is modelled by taking the initial location $x_0 \in [0, 1]$ and requiring the location-dependent wage, which is positive in $(0, 1)$, to be negative outside $[0, 1]$ (as a way of circumventing the need to introduce state constraints in the model) and to satisfy additional assumptions. Necessary conditions for optimality based on Pontryagin's maximum principle are derived and, based on them, various model properties (e.g., bounds on terminal assets $a(T)$) are established. The implications of particular functional forms for the wage distribution $w(x)$ are investigated. A first set of results confirms that for a single wage maximum economically intuitive behaviour can be obtained, namely, moving towards regions where higher wages can be earned. Further results for the case of a wage distribution with two local maxima, included in Andrey Vassilev's PhD dissertation, reveal more complex behaviour, including the possibility of moving towards the smaller local maximum depending on the length of the planning horizon, as well as the possibility of having multiple solutions.

The paper [1] further extends this line of research by looking at the case of two-dimensional economic space. This work provides results on existence and necessary conditions that are the counterpart of the ones obtained in [23]. The analysis also reveals the limitations of using a negative wage as a substitute for state constraints for two-dimensional domains of more complex (e.g., non-convex) shape. In particular, it turns out that in two-dimensional economic space it is possible to have an optimal trajectory in the spatial variables that passes through a region with negative wage values in order to reach faster a region with high positive wage values. Numerical investigation of the model shows that even in the simplest case of a two-dimensional wage distribution with a single maximum intuitive but nontrivial results can be obtained. More specifically, depending on the exact form of the wage distribution, the agent may have incentives to relocate in economic space along various nonlinear trajectories. This finding suggests that for more complicated two-dimensional wage distributions significantly more complex types of behaviour can be expected.

3.3. Models of small open economies

Iordan Iordanov also worked on models that help understand macroeconomic developments in small open economies. These models can be classified as belonging to the dynamic stochastic general equilibrium strand of the economic literature. A first attempt to develop such a model is documented in [24], which modifies a standard model of a small open economy to fit the case of an economy operating under a currency board arrangement. The model is tailored to the Bulgarian economy both in terms of how the currency board is introduced and also in terms of its calibration to data and empirical facts. It also specifies the government behaviour in a way that is consistent with its prominent role in situations where monetary policy is passive. Model properties are illustrated by studying dynamics under three scenarios: a shock in the prices of energy resources, a shock in the productivity of the non-tradable sector and a shock in tradable goods prices. The authors conclude that the responses of the model variables with respect to various shocks that affect the economy are plausible and correspond to usual consequences entailed by the main economic mechanisms built into the model but further work is needed to improve the specification of the government sector and introduce nominal rigidities and adjustment costs in the analytical framework.

The directions for model improvement identified in [24] are tackled in [26]. In this work the government is modelled as following a set of fiscal rules that aim to stabilise output fluctuations while maintaining fiscal stability. This approach more closely corresponds to guiding principles and best practices for conducting fiscal policy. Additionally, it reflects public discussions on fiscal policy principles that were ongoing in Bulgaria at the time of writing. The paper also improves on the framework adopted in [26] by introducing monopolistic competition and Calvo pricing to account for nominal rigidities and adjustment costs. The simulation results demonstrate the potential of using fiscal rules to stabilise the economy when monetary policy is constrained because of the currency board arrangement.

A final contribution in the modelling of small open economies appeared in [27]. This work suggests a way of extending the popular Galí-Monacelli (G-M) model to the case of a world economy consisting of several regions, interpreted as monetary unions. The paper provides the counterparts of selected derivations from the G-M model which lead to a representation of domestic output as a function of domestic and foreign prices, and exchange rates. It also develops an alternative index form for aggregators that yields qualitatively similar results to the usual one with the added benefit of having explicit dimensionality consistency.

3.4. Educational contributions

Besides his purely research-oriented contributions to economic modelling, Iordan was always interested in pushing the boundary where the tools for conducting research and presenting economic models are concerned. One of his pet peeves was with the way many economics papers present results by providing a model formulation and then stating the final form of the optimality conditions describing the purported solution without any reference to the method for deriving these conditions. Frequently in such cases reference is made to a Lagrangian formalism but, if the derivation is shown at all, it is little more than the mechanical application of a recipe. More precise formulations and justifications exist but are scattered throughout the literature. To partially remedy this, the paper [25] collects and presents some results on necessary conditions and sufficient conditions for optimality in dynamic optimization problems with explicit controls when one approaches the solution through a Lagrangian formalism.

Other contributions to the teaching of mathematical methods in economics appear in [29] and [32]. Both follow an agenda outlined in the preface to [44], where the role of using numerical methods and computer simulations to study economic models is emphasised. The paper [29] offers an approach to deepening the students' understanding of the Solow growth model by presenting an interactive software application that helps students get acquainted with the main properties of the model. The application is implemented in Python on the basis of the IPython kernel for the Jupyter Notebook. An advantage of this approach is that it hides the implementation details from the typical user while still allowing advanced students to access the application internals and learn from the underlying program code.

The paper [32] is similar in spirit to [29] in that it employs the same interactive approach and computational framework in order to study the popular IS-LM model.

4. Modelling in finance and insurance

Iordan Iordanov also researched problems related to modelling applications in finance and insurance. His papers in this field are focused on copula functions and are co-authored with his PhD students Nikolay Chervenov and Boyan Kostadinov.

Copulas appear as a modern tool in creating flexible probability models with more than one random variable. Any multivariate model could be constructed by means of copulas. They are widely applied in fields such as insurance, finance, or any field where there are interacting processes (risks) and there is a need to measure their dependence. Iordanov's works on copulas were innovative and unlike the most papers related to copulas, where considerations are restricted to functions which are well-defined in each point and benefit from copulas being Lipschitz functions on their domain, he and his co-authors considered copulas under more general assumptions.

The paper [17] deals with 2-dimensional copula functions and proposes a method to simplify the otherwise complicated task of verifying the 2-increasing of C-volume of the copula by using simple differentiation, followed by a method to obtain a class of copulas as a solution of a boundary value problem in a suitable Sobolev space.

Precisely, a weakly 2-increasing function is defined by using weak derivatives (derivatives in the sense of distribution theory), with $(H_{xy}, \varphi) \geq 0$, for any test function $\varphi \geq 0$ in $\mathscr{D}(\mathbb{R}^2)$. It is proved that, for $H \in \mathscr{D}'(\mathbb{R}^2) \cap C^0(\mathbb{R}^2)$, the standard 2-increasing and weakly 2-increasing properties are equivalent.

Furthermore, despite some restrictions imposed by Sobolev spaces, in this work a family of copulas C is constructed based on a given probability density $\partial_{uv} C$, using the concept of weak derivative. The main result is solving the boundary value problem

$$\begin{aligned} &\partial_{uv} C(u,v) = f(u,v) \text{ in } I^2 = [0,1] \times [0,1] \text{ (in weak sense);} \\ &C(u,0) = 0 = C(0,v); \\ &C(u,1) = u,\ C(1,v) = v \text{ for all } u,v \in I = [0,1], \end{aligned}$$

under certain conditions on f. First the case when f is smooth function is considered and later it is generalized to $f \in W^{-1,p}(I^2)$, $p > 2$. One might initially think of this problem as a Dirichlet problem for the wave equation, but it turns out to be a Goursat problem for hyperbolic equations. However, it is important to note that this is an ill-posed boundary value problem.

While in [17] the results are given rather schematically, in [18] detailed proofs and numerous examples are given for the same problems. The paper pays special attention to the *a priori* estimate, which ensures the uniqueness of the solution C. Namely, there exists a constant M, which does not depend on f, such that $\|C\|_{W^{1,p}(I^2)} \leq M\|f\|_{L_p(I^2)}$. Next, the work provides detailed proofs of theorems on the existence of the solution in both the smooth $f \in L_p(I^2)$ and generalized $f \in W^{-1,p}(I^2)$, $p > 2$, cases.

Further works extend the 2-dimensional case to the n-dimensional, and study the n-increasing property, existence and uniqueness of the solution of the boundary value problem when $n > 2$.

The paper [2] defines the Goursat problem over the unit cube I^n. It is proved that there exists a unique solution of the problem

$$(-1)^n (C, \varphi_{x_1,\ldots,x_n}) = (f, \varphi)$$

for all $\varphi \in C_0^\infty(\mathbb{R}^n \setminus \bigcup S_j)$ restricted over I^n, where $S_j = I^n \bigcap \{(u_1,\ldots,u_n) \in \mathbb{R}^n \mid u_j = 1\}$ are the sides of I^n passing through the vertex $(1,\ldots,1)$ and the function C satisfies the usual boundary conditions for a copula function $C(u_1,\ldots,u_n) = 0$, if at least one of the coordinates $u_1,\ldots,u_n$ equals zero.

The article [3] was Iordanov's final work on the topic. The paper presents two generalizations of the concept of an n-increasing function. The main result on the subject states that a function $H \in W^{1,p}(I^n)$ is weakly n-increasing in I^n if the inequality

$$(-1)^n (H, f_{x_1,\dots,x_n}) \geq 0$$

is fulfilled for all $f \geq 0$ in $W^{n-1,q}(I^n)$ such that f and its derivatives vanish on the sides of I^n passing through the vertex $(1, \dots, 1)$.

This makes it easy to obtain several important results for copulas. In particular, the authors apply their approach to the bivariate Archimedean copulas, which are a special class of copulas. The proof of the characteristic property for Archimedean copula is completely new in the literature.

Another problem studied in [3] relates to the construction of n-dimensional copulas by using values of their derivatives $\frac{\partial^n}{\partial x_1 \dots \partial x_n}$. To this end, the paper uses the generalized concept of an n-increasing function and expands the results of a version of the Goursat problem in n-dimensional cube in $\mathbb{R}^n$ proved in [2]. It is worth mentioning that the local conditions required to solve the boundary value problem in the 2-dimensional case are equivalent to those imposed in the n-dimensional case.

A notable application of the results presented in this section can be found in Nikolay Chervenov's PhD dissertation. The work considers a real-life problem for insurance risk assessment and solves it with numerical methods. The desired copula was constructed using the approach of Iordanov and Chervenov, and was obtained as a solution of a boundary value problem using real data from a Bulgarian insurance company.

Acknowledgements

This work has benefited from conversations and materials provided by Tatiana Chernogorova, Vaia Rakidzi and Tzenka Videnova.

References

[1] I. Brayanov, I. Iordanov and A. Vassilev, Spatial relocation decisions of dynamically optimizing consumers in two-dimensional economic space, International Journal of Technology, Modeling and Management 1(1) (2010) 75–84.

[2] N. Chervenov, I. Iordanov and B. Kostadinov, Goursat problem over unit cube in first quadrant of $\mathbb{R}^n$ (with applications to existence of copulas), AIP Conference Proceedings 2048, 040022, 2018, https://doi.org/10.1063/1.5082094.

[3] N. Chervenov, I. Iordanov and B. Kostadinov, n-dimensional copulas and weak derivatives, Serdica Math. J. 44 (2018) 413–438.

[4] V. Georgiev and A. Ivanov, Concentration of local energy for two-dimensional wave maps, Rend. Istit. Mat. Univ. Trieste 35(1–2) (2003), 195–235.

[5] V. Georgiev and A. Ivanov, Existence and mapping properties of the wave operator for the Schrödinger equation with singular potential, Proc. Amer. Math. Soc. 133(7) (2005) 1993–2003.

[6] I. Iordanov, Regularity of the solution of a certain elliptic variational inequality with two constraints, Annuaire Univ. Sofia Fac. Math. Méc. 67 (1972/73) 169–190 (in Bulgarian).

[7] I. V. Iordanov, Régularité de la solution d'une inéquation variationnelle elliptique à deux contraintes, C. R. Acad. Bulg. Sci. 26(12) (1973) 1587–1589.
[8] I. V. Iordanov, Inéquation variationnelle associée à un opérateur elliptique dégénéré, C. R. Acad. Bulg. Sci. 27(11) (1974) 1473–1476.
[9] I. V. Iordanov, Problème non-coercif pour une inéquation variationnelle elliptique à deux contraintes, Serdica 1(3) (1975) 261–268.
[10] I. V. Iordanov, Deux notes sur quelques inéquations variationnelles elliptiques, Serdica 4(2–3) (1978) 232–240.
[11] I. V. Iordanov, Conditions de régularité de la solution d'un problème non-coercif [Conditions for regularity of the solution of a noncoercive problem], Pliska Stud. Math. Bulgar. 3 (1981) 61–66.
[12] I. V. Iordanov, Solutions höldériennes d'inéquations variationnelles à contraintes discontinues. I. [Hölder solutions of variational inequalities with discontinuous constraints. I], Serdica 8(3) (1982) 296–306.
[13] I. V. Iordanov, Hypoellipticity of quasilinear elliptic operators, C. R. Acad. Bulg. Sci. 35(12) (1982) 1627–1630.
[14] I. V. Iordanov, Solutions höldériennes d'inéquations variationnelles à contraintes discontinues. II. [Hölder solutions of variational inequalities with discontinuous constraints. II], Serdica 9(3) (1983), 289–300.
[15] I. V. Iordanov, Anomalous singularities of semilinear hyperbolic systems, C. R. Acad. Bulg. Sci. 41(10) (1988) 13–15.
[16] I. Iordanov, Anomalous singularities of semilinear nonstrictly hyperbolic system, J. Math. Pures Appl. 70(4) (1991) 393–426.
[17] I. Iordanov and N. Chervenov, Copulas on Sobolev spaces, C. R. Acad. Bulg. Sci. 68(1) (2015) 11–18.
[18] I. Iordanov and N. Chervenov, Copulas on Sobolev spaces, Serdica Math. J. 42 (2016) 335–360.
[19] I. Iordanov and P. Popivanov, Weakly hyperbolic systems having infinite order of tangency between two characteristics – regularizing effect on the anomalous singularities, C. R. Acad. Bulg. Sci. 56(12) (2003) 11–16.
[20] I. V. Iordanov, S. V. Stoyanov and A. A. Vassilev, Price dynamics in a two region model with strategic interaction, in: Proceedings of the Spring Conference of the Union of Bulgarian Mathematicians (2004) 144–149.
[21] I. V. Iordanov and A. A. Vassilev, Existence of solutions in a class of dynamic spatial consumption models, in: Proceedings of the Spring Conference of the Union of Bulgarian Mathematicians (2007) 252–256.
[22] I. V. Iordanov and A. A. Vassilev, Optimal relocation strategies for spatially mobile consumers, C. R. Acad. Bulg. Sci. 60(11) (2007) 1163–1166.
[23] I. V. Iordanov and A. A. Vassilev, Optimal relocation strategies for spatially mobile consumers, Journal of Concrete and Applicable Mathematics 6(1) (2008) 101–126.
[24] I. V. Iordanov and A. A. Vassilev, A small open economy model with a currency board feature — the case of Bulgaria, Bulgarian National Bank Discussion Paper DP/63/2008, 2008.
[25] I. Iordanov and A. Vassilev, Notes on the solution of dynamic optimization problems with explicit controls in discrete-time economic models, in: Proceedings of the Spring Conference of the Union of Bulgarian Mathematicians (2011) 187–192.
[26] I. V. Iordanov and A. A. Vassilev, The stabilizing role of fiscal policy rules under the Bulgarian currency board arrangement, C. R. Acad. Bulg. Sci. 65(11) (2012) 1499–1506.

[27] I. V. Iordanov and A. A. Vassilev, Extending a small open economy model to a multi-regional setting: domestic output determination, C. R. Acad. Bulg. Sci. 67(12) (2014) 1621–1628.

[28] I. V. Iordanov and A. A. Vassilev, Strategic trade with partial local consumer protection, in: Proceedings of the 46th Spring Conference of the Union of Bulgarian Mathematicians (2017) 181–186.

[29] I. V. Iordanov and A. A. Vassilev, Teaching the Solow growth model — an interactive approach via IPython in the Jupyter notebook, Fourteenth International Scientific Conference "Development of Higher Schools in the Context of European Requirements for Quality of Education Services", International Business School (2017) 553–560.

[30] I. Iordanov and A. Vassilev, Strategic trade between two countries — exploring the case of partial local consumer protection, Serdica Journal of Computing 11(1) (2017) 31–43.

[31] I. V. Iordanov and A. A. Vassilev, Strategic trade between two regions with partial local consumer protection — General setup and Nash equilibria, AIP Conference Proceedings 1910, 020012, 2017, https://doi.org/10.1063/1.5013949.

[32] I. Iordanov and A. Vassilev, A computational approach to teaching the IS-LM model, International Conference on High Technology for Sustainable Development (HiTech) 2018, https://doi.org/10.1109/HiTech.2018.8566441.

[33] A. Jensen and S. Nakamura, Mapping properties of functions of Schrödinger operators between L^p-spaces and Besov spaces. Spectral and scattering theory and applications, Adv. Stud. Pure Math. 23 (1994) 187–209.

[34] A. Jensen and S. Nakamura, L^p-mapping properties of functions of Schrödinger operators and their applications to scattering theory, J. Math. Soc. Japan 47(2) (1995) 252–273.

[35] M. Keel and T. Tao, Endpoint Strichartz estimates, Amer. J. Math. 120(5) (1998) 955–980.

[36] H. Lewy and G. Stampacchia, On the regularity of the solution of a variational inequality, Comm. Pure Appl. Math. 22 (1969) 153–188.

[37] J.-L. Lions and G. Stampacchia, Variational inequalities, Comm. Pure Appl. Math. 20 (1967) 493-–519.

[38] P. R. Popivanov and I. V. Iordanov, Paradifferential operators and propagation of singularities for nonlinear P.D.E., Report MATH 83-6, Akademie der Wissenschaften der DDR, Institut für Mathematik, Berlin (1983).

[39] P. Popivanov and I. Iordanov, Anomalous singularities of the solutions to several classes of weakly hyperbolic semilinear systems. Examples, Progress in Analysis, Vol. I, II (Berlin, 2001), World Sci. Publ., River Edge, NJ (2003) 697–703.

[40] P. Popivanov and I. Iordanov, On the anomalous singularities of the solutions to some classes of weakly hyperbolic semilinear systems. Examples, Evolution Equations (Warsaw, 2001), Banach Center Publ. 60, Polish Acad. Sci. Inst. Math., Warsaw (2003) 209–218.

[41] G. Stampacchia, Contributi alla regolarizzazione delle soluzioni dei problemi al contorno per equazioni del secondo ordine ellitiche, Ann. Scuola Norm. Sup. Pisa Cl. Sci. 12(3) (1958) 223–245.

[42] A. A. Vassilev, I. V. Iordanov and S. V. Stoyanov, A strategic model of trade between two regions with continuous-time price dynamics, C. R. Acad. Bulg. Sci. 58(4) (2005) 361–366.

[43] A. Vassilev and I. Iordanov, Introduction to macroeconomics, Sofia University "St. Kliment Ohridski" Publishing House, 2016 (in Bulgarian).

[44] G. Venkov and I. Iordanov, Mathematical models in economics, Technical University-Sofia Publishing House, 2016 (in Bulgarian).

Received on June 3, 2024
Accepted on September 17, 2024

Nikolay Chervenov

BNP Paribas Cardif Bulgaria and Romania
29, Hristo Belchev Str.
1000 Sofia
BULGARIA

E-mail: `nikolay.chervenov@bnpparibas.com`

Vladimir Georgiev

Department of Mathematics
University of Pisa
Largo Bruno Pontecorvo 5
56127 Pisa
ITALY

E-mail: `georgiev@dm.unipi.it`

Andrey Vassilev

Faculty of Economics and Business Administration
Sofia University "St. Kliment Ohridski"
125 Tsarigradsko Shosse Blvd., Bl. 3
1113 Sofia
BULGARIA

E-mail: `a.vassilev@feb.uni-sofia.bg`

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“
ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Том 111

ANNUAL OF SOFIA UNIVERSITY “ST. KLIMENT OHRIDSKI”
FACULTY OF MATHEMATICS AND INFORMATICS
Volume 111

CAPEC ONTOLOGY GENERATOR

Vladimir Dimitrov

CAPEC is an effort coordinated by MITRE Corporation. Its aim is attack pattern database structured in taxonomies. CAPEC is available as XML document from its project site. CAPEC structure and content are under permanent change and development. It is still not mature database but may be never will.
CAPEC, CWE, and CVE are databases devoted to attacks, weaknesses, and vulnerabilities. They refer each other forming a knowledge ecosystem in cybersecurity area.
Traditional approach for knowledge presentation as information does not work well with conceptualizations under dynamics of this ecosystem and particularly of CAPEC. In this paper, an alternative approach to CAPEC knowledge presentation is proposed, as ontology. First, CAPEC structure and content are discussed and then ontology structure is introduced. CAPEC as ontology opens doors to “open world” concept that is more adequate to ecosystem dynamics.
CAPEC ontology is programmatically generated from CAPEC database.
CAPEC ontology generator in implemented in Python.

Keywords: cybersecurity, attack patterns, ontology, CAPEC, OWL, Python

2020 Mathematics Subject Classification: 68M25, 68T30, 68U35

CCS Concepts:
- Security and privacy~Formal methods and theory of security~Formal security models;
- Security and privacy~Formal methods and theory of security~Logic and verification

1. Attack patterns

CAPEC (Common Pattern Enumeration and Classification) [5] is an effort coordinated by MITRE Corporation. Its aim is attack pattern database structured in taxonomies.

CAPEC is freely available in XML format from the site.

MITRE Corporation maintains two more initiatives CVE [4] and CWE [11]. These are vulnerability and weakness databases.

DOI: 10.60063/GSU.FMI.111.25-33

CVE is mature enough to be maintained officially by NIST as NVD [7].

Original CVE is still maintained by MITRE Corporation for new vulnerability registration.

NVD contains only analyzed vulnerabilities augmented with additional analytic information supplied by NIST.

CWE and CAPEC are still not mature. They are under active development.

CVE, CWE, and CAPEC refer to each other forming knowledge ecosystem in cybersecurity area.

Vulnerabilities are revealed errors, faults, or gaps in specific products fixed by its version and execution environment. CVE entries are numerous – currently 197676 in March 2023.

Weaknesses (CWE) are vulnerability types. Vulnerability processing includes type assignment – one or more CWEs are associated with vulnerability under consideration.

Attacks (CAPEC) exploit vulnerabilities. More precisely, attack patterns exploit one or more weaknesses.

CWE and CAPEC have structures organized by several taxonomies, while CVE is simply a list.

CWE and CAPEC are still not mature and they contain some discrepancies in their notation, structure and content.

CWE and CAPEC are distributed as XML documents, while CVE – as JSON documents.

Our approach is to present CVE, CWE, and CAPEC in OWL [12]. Semantic web is more suitable for formal knowledge presentation than XML. This is especially valid when the knowledge base is under development – not mature.

The term "formal knowledge presentation" refers to conceptualized knowledge presentation, for example via OWL.

CVE, CWE, and CAPEC ontologies are presented in several other publications.

Subject of this paper is CAPEC ontology generation from its XML presentation.

2. CAPEC ontology generator

CAPEC ontology generator does not require parallelism. On 02.01.2023 attack patterns are 555.

CAPEC ontology generator is published at the following address `https://github.com/VladimirDimitrov1957/CAPEC-ontology-generator`. It is implemented in Python.

CAPEC generator presentation below follows its control flow.

The generator by default works with input data downloaded in "`data`" subdirectory. With parameter "`-d`" or "`-download`" fresh copy of CAPEC database can be downloaded from CAPEC site.

Procedure `main(download)` manages the whole generation process.

Procedure `downloadCAPEC()` downloads database copy from CAPEC site. This copy is compressed in zip format, so the procedure decompresses downloaded file. CAPEC database is XML document.

Every time validation of input XML document is done using CAPEC XSD schema. For this purpose, `lxml` package is used. It is important to match used packages because some of them simply does not work or are not usable.

In the next step, XML document presenting the database is parsed to internal format. This parsed CAPEC database is used in the following operations.

Procedure `generateIndividuals(root)` performs CAPEC ontology generation. Parameter "`root`" presents CAPEC database in parsed internal format.

This procedure copies and modifies ontology shell from file `shell.owl`. The result is written in file `capec.owl`. External references in the last file are presented as annotations.

Then procedure `generateIndividuals(root)` generates attack patterns, categories, and views with:

`generateAttackPatternIndividual(item, out_file)`,
`generateCategoryIndividual(item, out_file)`, and
`generateViewIndividual(item, root, out_file)`
presented below.

All catalog elements are scanned and corresponding generation procedure is called.

Finally, procedure ends with generation of object property objects. It is described in more details in subsections about `AttackPattern`, and `Individual` class. Figure 1 shows class structures.

Procedure `generateAttackPatternIndividual(item, out_file)` has parameter "`item`" that contains an internal XML presentation of an attack pattern. For this attack pattern, the procedure generates an individual description.

Parameter "`out_file`" is the file in which the ontology is written. Here, more specifically, the attack pattern individual description is written.

`AttackPattern` class plays central role for attack pattern individual generation. An object of this class collects all individual characteristics. Then these characteristics are serialized as a string and written in the ontology file. The class is described in separate section below.

Procedure `generateCategoryIndividual(item, out_file)` is simple. It follows category subelements structure and generates a category individual description.

Procedure `generateViewIndividual(item, root, out_file)` is hardcoded by catalog views. The last are very specific to be implemented in some common generation algorithm.

`Individual` is the other class in CAPEC generator. It collects information about the target objects for object properties. This class has extent, i.e. supports set of all `Individual` objects.

At the end of procedure `generateIndividuals(root)`, objects from `Individual` extent are serialized and appended to the result file.

Class `AttackPattern` and `Individual` are presented in Figure 1 as UML class diagram.

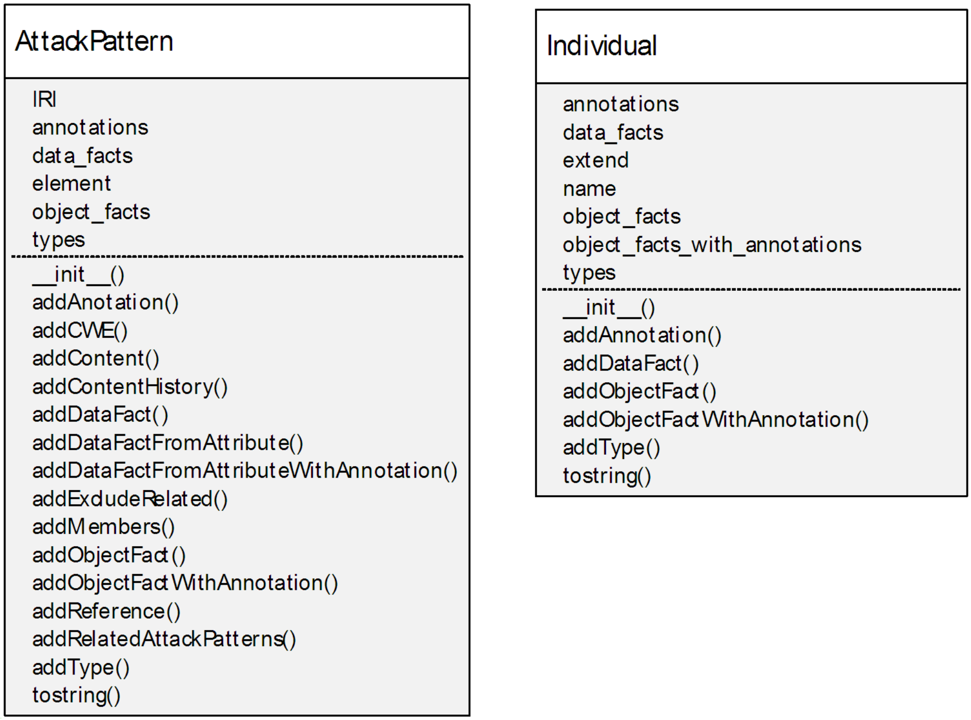


Figure 1. CAPEC generator classes

2.1. AttackPattern class

`AttackPattern` instance variables are “`element`” – internal XML presentation of attack pattern, individual IRI, dictionary of annotations, dictionary of data facts, dictionary of object facts, and set of individual types. Data facts present individual data property instances. Object facts present individual object property instances.

When an object of the class is instantiated, its instance variables are accordingly initiated.

Method `addType(self, aName)` adds a type to the individual type set. This operation is hardcoded for categories because their type is more specifically extracted.

Method `addDataFact(self, tag, path = "", structured = False)` adds a data fact.

The data fact is extracted from a subelement of “`element`”. The relative path to this subelement is given by parameter “`path`”.

Content of a data fact element can be structured text. It is marked by the parameter “`structured`”.

If data fact content is structured text, function `stext(s, tag)` is used to remove line formatting and codes the text as OWL literal. For the last operation, function `flat(s)` is used, but that function is used regularly for non-structured text.

Method `addDataFactFromAttribute(self, att)` extracts data fact from “`element`” attribute. Parameter “`att`” contains attribute name and it is a data property name too.

Method `addDataFactFromAttributeWithAnnotation(self, el, att, path, aName)` extracts data fact from attributes of “`element`” subelements. Parameter “`el`” contains a subelement tag. If the subelement is not a child of “`element`”, then the relative path to it is given by parameter “`path`”.

It is possible the subelement to have several instances. In that case, the data fact should be with several values.

Parameter “`aName`” is a dictionary with keys that are data fact values. Dictionary value is a list of annotations applicable for this data fact. In this method, only list with one annotation name is used, but presentation is unified with all other methods.

Method `addDataFactWithAnnotation(self, tag, aTag, path = "", name = None, aName = None, structured = False)` is as previous one, but it extracts data facts from “element” subelements.

Attributes are always extracted from subelement content – not from attribute values.

Parameters “`tag`” and “`aTag`” contents are data fact and annotation element tags correspondingly. In most of the cases, tag and data property share the same name. The same is valid for annotation element tag and annotation name. If it is not true, then data fact and annotation name are set by parameters “`name`” and “`aName`”.

Parameter “`path`” is used when subelements are not children of “`element`”. This is as in the previous method.

Parameter “`structured`” has the same purpose as in method `addDataFact (self, tag, path = "", structured = False)`.

Method `addAnnotation(self, tag, name = None, path = "", structured = False)` extracts annotations form “`element`” subelements. If the annotation name differs from the subelement tag, then parameter “`name`” is used. The other parameters have the same meaning and usage as in above methods.

Method `addReferences(self)` is more specific. It extracts references from “`element`” – from subelements `References` and `Reference`. Annotations are formatted as literals. All references are annotations.

Method `addContentHystory(self)` is specific. It extracts and adds as annotations the content history.

Method `addObjectFact(self, path, oName, cName, cADict)` extracts object facts form “`element`” subelements.

If subelements are not “`element`” children, then parameter “`path`” sets the relative path to them.

Parameter “`oName`” contains the object property name.

The target individual type is given by the parameter “`cName`”.

Parameter "`cADict`" is a dictionary. It is used for target individual generation. Dictionary key is the subelement tag and that attribute is set by "`oName`" parameter.

Method `addObjectFactWithAnnotation(self, path, oName, cName, cADict = {}, cSDict = {}, cANDict = {}, references = False, note = False)` extracts object facts with annotations. This is the most complex method. All parameters including "`cADict`" have the same purpose and meaning as in the above method.

Parameter "`cSDict`" is a dictionary used for extraction of data facts from subelements of some "`element`" subelements. The dictionary key is the subelement tag, and the dictionary value is the data fact name. Observed examples are more specific case because they may reference CVE individuals or simply to be `Reference` annotations.

Parameter "`cANDict`" is a dictionary. It used for annotation extraction for the target individual from attributes. Target individual is an individual that is generated. The object fact (object property) points to this target individual. The dictionary key is the subelement tag and the dictionary value is the annotation name. Annotations usually are added to the target individual, but in case of techniques and attack identifiers, annotations are added to the object facts of the target individual.

Method `addCWE(self)` adds referenced CWE individuals from attack patterns elements as object facts.

Method `addExcludeRelated(self, category)` adds excluded category ancestors of the attack. The ancestor category is given with parameter "`category`".

Method `tostring(self)` serializes its object into a string containing individual description.

Method `addMembers(self, relationships = False)` is used to add views and categories. Link type is determined via "`relationships`" parameter – by default, category members are processed.

Method `addRelatedAttackPatterns(self)` adds the object facts for related attack patterns. Here, excluded categories are taken in account.

Method `addContent(self, capecID)` is specialized for `Has_Member` object fact with attack identifier for attack pattern. Parameter "`capecID`" contains this attack identifier.

Methods in `AttackPattern` class are combinations of with / without annotations, from attributes / subelements, from children / other descendants, for data / object properties. Here, only combinations that exist in CAPEC ontology and CAPEC XSD structure are implemented. This note is applicable for `Individual` class too.

2.2. Individual class

`Individual` class is a simplified implementation of `AttackPattern` class. It collects the characteristics of the target individuals of the object facts.

Here, class extend is supported (class variable "`extend`"), from which the description of all target individuals are serialized at the end of procedure `generateIndividuals (root)`.

Methods `addType(self, t)`, `addDataFact(self, d, v)`, `addAnnotation (self, a, v)`, `addObjectFact(self, d, v)`, `addObjectFactWithAnnotations (self, d, v, an, av)`, and `tostring(self)` are simplified implementations of the corresponding `AttackPattern` methods.

3. Conclusion

CAPEC generator implementation is similar to that of CWE generator. In both implementations, the control flow follows corresponding XSD schemas. Both implementations have two classes `AttachPattern`/`Weakness` and `Individual`. CWE generator classes are presented as UML class diagram in Figure 2.

From complier/translator point of view, ecosystem generators for CPE, CVE/NVD, CWE, and CAPEC ontologies are compilers from XML/JSON to OWL.

The components of every compiler are lexical, syntactical, semantical analyzers, and code generator. In this case, the analyzer functionality is derived from Python libraries. Only code generators are coded.

For CPE and CVE/NVD generators, concurrent implementations have been adapted to achieve acceptable production timings. In these implementations, all available cores and memory are used for processing.

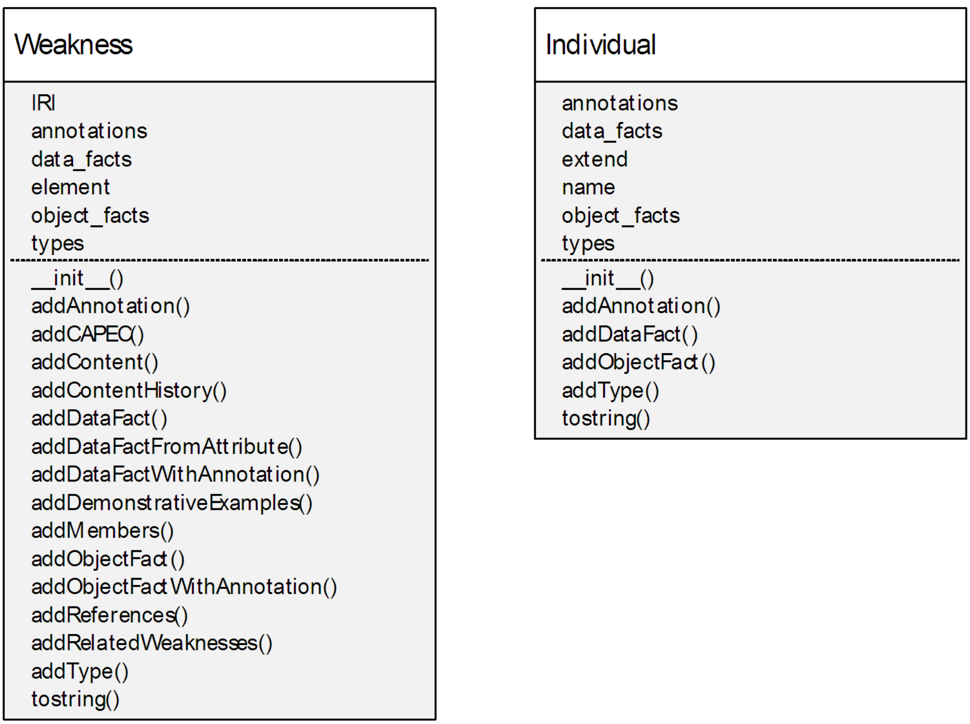


Figure 2. CWE generator classes

Many applications have no option to use available computer hardware resources. Usually, they work with one core and limited memory, independently of the job. This problem, here, in the above mentioned generators, is roundabout by manual programming. Today, we are still far away from full automatization of concurrent programming. Therefore, every change, especially in concurrent programs, is a difficult task.

Every change in XSD schemas reflect in generator implementations. This happened at least one time in the year.

Another problem is the change of generator data sources. For example, NIST plan in September 2023 to retire NVD and CPE data feed supplies and to start only JSON RESTfull services for that purpose. Therefore, the corresponding generators have to be rewritten to meet this change.

All ecosystem ontologies have been presented in OWL Manchester syntax [9]. This syntax is more human readable, but it is not used as input source for currently available triple stores. The last ones support Turtle syntax [10]. Therefore, the generators have to be rewritten to generate in Turtle syntax. This process is going on and some generators available on published yet addresses generate their ontologies in Turtle syntax. From programming point of view, it is not dramatic change. Something more, generator codes get simplified.

Ontologies in Turtle syntax for presentation purpose can be easily converted into Manchester syntax using some tools like Protégé [6] or Robot [8].

Ecosystem ontologies must be available in a triple store if we want the ecosystem to be used. The most important task is to load and update ecosystem ontologies into some triple store. However, CVE/NVD ontology is really huge. Currently, ecosystem ontologies contain approximately eight billion triples. Experiment shows that these ontologies can be loaded in a triple store on modern desktop computer for two months.

Experiments have been done with BrightstarDB [3] and Apache Jena [1]. These triple stores have their advantages and disadvantages, but their common problem is the slow batch load. Both triple stores do not use all available hardware resources. Apache Jena is somehow better than BrightstarDB, but still is not a solution.

BraghtstarDB is well integrated in .NET environment. Automatically, OWL ontologies can be presented as C# classes and used for object-oriented development. At the same time, there is access with SPARQL to its ontologies, but that is all.

Apache Jena is a complex of many Semantic web tools, but their integration is achieved via Java programming. Therefore programming in Apache Jena is not very automated.

The problem with CVE/NVD ontology actualization can be solved following NIST proposition. Vulnerabilities are organized in chunks by years. Therefore, they can be loaded once. Then, NIST offers data feed for modified vulnerabilities. This patch can be applied to update quickly CVE/NVD content.

Acknowledgements

This paper is prepared with the support of MIRACle: Mechatronics, Innovation, Robotics, Automation, Clean Technologies – Establishment and Development of a Center for Competence in Mechatronics and Clean Technologies – Laboratory Intelligent Urban Environment, funded by the Operational Program Science and Education for Smart Growth 2014–2020, Project BG 05M2OP001-1.002-0011.

References

[1] Apache Jena, A free and open source Java framework for building Semantic Web and Linked Data applications, https://jena.apache.org.

[2] S. Barnum and A. Sethi, Attack patterns as a knowledge resource for building secure software, Cigital, Inc., 2007, https://capec.mitre.org/documents/Attack_Patterns-Knowing_Your_Enemies_in_Order_to_Defeat_Them-Paper.pdf.

[3] BrightstarDB, A native RDF database for the .NET platform, https://brightstardb.com.

[4] E. Gamma, R. Helm, R. Johnson and J. Vlissides, Design patterns: Elements of reusable object-oriented software, Boston, MA, Addison-Wesley, 1995.

[5] MITRE Corporation, CAPEC (Common Pattern Enumeration and Classification), 2023, http://capec.mitre.org.

[6] M. A. Musen, The Protégé project: A look back and a look forward, AI Matters 1(4) (2015) 4–12, https://doi.org/10.1145/2757001.2757003.

[7] NIST, NVD (National Vulnerability Database), 2023, http://nvd.nist.gov.

[8] Open Biological and Biomedical Ontology Foundry (OBO), ROBOT, ROBOT is an OBO Tool, http://robot.obolibrary.org.

[9] W3C, OWL 2 Web Ontology Language, Manchester Syntax (Second Edition), W3C Working Group Note, 11 December 2012, https://www.w3.org/TR/owl2-manchester-syntax.

[10] W3C, RDF 1.1 Turtle, Terse RDF Triple Language, W3C Recommendation, 25 February 2014, https://www.w3.org/TR/turtle.

[11] W3C, Shapes Constraint Language (SHACL), W3C Recommendation, 20 July 2017, http://www.w3.org/TR/shacl.

[12] W3C, Semantic Web, Web Ontology Language (OWL), 2023, https://www.w3.org/OWL.

Received on January 3, 2024
Accepted on February 27, 2024

Vladimir Dimitrov

Faculty of Mathematics and Informatics
Sofia University "St. Kliment Ohridski"
5, James Bourchier Blvd.
1164 Sofia
BULGARIA

E-mail: cht@fmi.uni-sofia.bg

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“
ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Том 111

ANNUAL OF SOFIA UNIVERSITY “ST. KLIMENT OHRIDSKI”
FACULTY OF MATHEMATICS AND INFORMATICS
Volume 111

EXPLORING SOFTWARE ENGINEERING KNOWLEDGE DOMAINS

DILYAN GEORGIEV

Software engineering, as the primary value-based process, is influenced by the culture a given company establishes, formal and informal procedures, and technological improvements introduced during the implementation. Having this in mind it could be articulated further how these innovative methods could result in best practices, communication flows, strong-bonded teams, and successful projects. By adding an organisational domain to the model, the relationships between management's organizational aspects and technological development approaches are discussed.
This paper aims to explore and classify the software engineering domains in order to facilitate the process of managing knowledge within ICT companies. Based on a literature overview, a model is proposed which combines the multiple perspectives for adopting knowledge management practices more efficiently.

Keywords: software engineering, knowledge management, software process improvement, knowledge domains, overview

2020 Mathematics Subject Classification: 68N01, 68N30

CCS Concepts:
• Software and its engineering~Software organization and properties~Contextual software domains

1. Introduction

Software engineering is a relatively new knowledge domain that has been through many changes during the last few decades. Considering the multiple perspectives having its influence, it emerges out of the ICT, also recognized as an industry about other industries because of the role it has in their growth and evolution. Furthermore, software engineering as a “knowledge-intensive activity” [6] results in solutions, expected to play an even more critical role in society and the economy, for example,

DOI: 10.60063/GSU.FMI.111.35-53

by implementing artificial intelligence technologies in new cyber-physical solutions such as robotised and autonomous systems – self-driving cars, unmanned aerial vehicle (UAV), intelligent systems, and many more. Enabling software companies to improve the way they manage their knowledge processes is increasingly important in this specific moment when hardware-based infrastructures and systems are being upgraded with digital functionalities. Suppose this is put through a more global perspective. In that case, digital solutions aim to tackle complex socio-economic and ecological problems that, when solved, will cause a radical improvement in the way our environment is being constructed. Concepts like smart cities, green economies, and meta worlds could provide a new way of handling those problems and set up a new course of the way software engineering is progressing.

Knowledge management, on second thought, introduces many instruments, methods, practices, and approaches for value creation, especially in knowledge-intensive industries which "has spurred an exponential increase in publications covering a broad spectrum of diverse and overlapping research areas" [43]. The processes of codification and personalisation lay in the basis of those instruments since the company should support knowledge generation, sharing, and transferring among the employees. This is crucial, especially in software processes that require high expertise since much of the knowledge used for a given solution is wrapped as intangible assets. However, stimulating these processes could provide a valuable setup for the project's growth.

Furthermore, new software engineering methods and concepts are introduced to validate requirements beforehand and to apply methodologies at the team level. Hence, a value co-creation system could be established, involving different stakeholders, and coordinating between complex business models. The user participates more actively and provides feedback that helps developers work on well-defined functionalities. On the other hand, new solutions require better cohesion between the architecture of a given product, communication channels, and handling changes through time. Considering how dynamic and sometimes unpredictable the software industry is about technologies and paradigms, additional expertise is needed regarding specific solutions and management on different levels for processes, experiences, and integrations with other systems.

All this demonstrates the need for improved KM processes in software engineering. Thus, a new paradigm is proposed that could provide bigger clarity on how critical it is to create relations between the different management aspects, on the one hand, and the software development, on the other hand. Through that perspective, knowledge seems even more valuable highlighting how important it is to introduce a new understanding of how it can be efficiently managed within the software engineering aspect and the organizational and socio-economical contexts.

The main goal of this research is to overview the latest methods, approaches, and trends in software process improvement, focusing on the domain and organisational concepts of management. To do that, two critical questions are answered:

- How the organisation handles its intellectual capital to achieve its purposes and goals?

- How is different knowledge being handled in different domains?

After analysing how knowledge evolves within these two aspects, the idea is developed further by inspecting previous systematic literature reviews. The concepts that lay the foundations of current ICT evolution are highlighted and a simple model is introduced, which represents how knowledge can be categorised through the organisational and domain perspectives (Section 2). This will introduce an advanced view of areas in which studies have been focused. Then, after overviewing papers retrieved from different resources, many methods, architectures, and methodologies are identified that are used to handle the issues related to knowledge management (Section 3). A short discussion is presented in Section 4, and conclusions and predictions are made in Section 5.

2. Theoretical background foundations

There have been many changes in how knowledge is perceived in the software industry. At first, knowledge was considered a possession, "something that could be captured" [6]. However, currently, it is much more important to have the knowledge applied than to possess it – it is called the knowledge-in-action concept, and it is even more critical if not involved in the software process. Being context-dependent, applying the knowledge is an activity that needs additional preparation, as internal stakeholders must be aware of the available explicit and tacit knowledge [39]. "Tailoring software development" by applying agile concepts draws the development team's attention, strengthens the decision-making process, increases inter-team coordination, and builds collective-code ownership [31]. The creation, storage, transfer, and retrieval activities must be introduced to generate new ideas that could support the product, expand its scope, and implement its functionalities according to the stakeholders' expectations. Regarding product delivery, there are frameworks like DevOps that could guarantee product quality and simplify and fasten integration processes with the production environment [10]. This corresponds with the goals regarding customer expectations – building trust, bringing satisfaction, and establishing the belief that "working software is the primary measure of progress" [34].

To define the SE Knowledge domain, a simple concept representation is made, which includes 16 systematic reviews and their main topics (Figure 1). By comparing different aspects, an analysis of the connections between domains is made, and subdomains are identified and highlighted.

First, two significant distinctions must be made between KM and SE concepts. When discussing organisational topics, the organisational culture and structure should be considered, as well as the mission, vision, learning, and growth. The organisational structure supports transferring tacit knowledge through its levels of management [3]. The culture company supports, on the other hand, is critical for sharing between teams and departments. The mission and vision dictate the overall motivation among the employees, and learning and growth could help to introduce new opportunities and challenges.

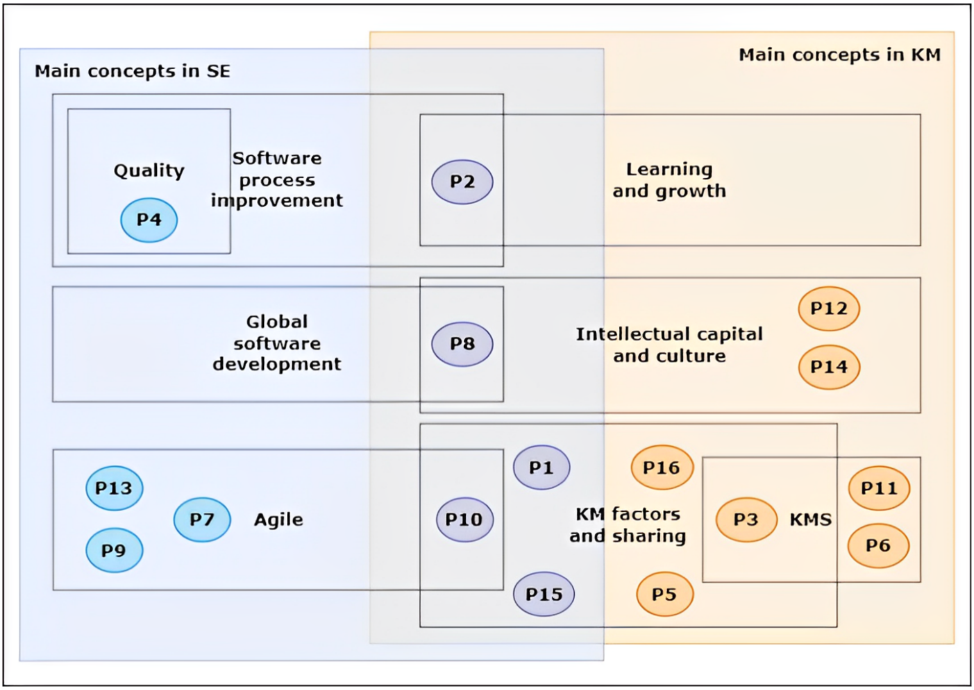


Figure 1. Main SE and KM concepts found in research data

In terms of domains, the focus is on the knowledge that comes from the specific environment and requires expertise that could provide solutions to a problem or set of problems. In that manner, Knowledge management is used to drive at least one of the five directions of development [43]:

- Ontology of Knowledge;
- Knowledge Management Systems (KMS);
- Role of Information Technologies;
- Managerial and Social issues;
- Knowledge Measurement.

All these directions aim to support the codification and personalisation processes – previously defined in the SECI model, which describes the different types of transformations between tacit and explicit knowledge – “organisational knowledge creation involves continual interaction” between those two types of knowledge [28]. Those models can be added to the models defined by Nonaka and Takeuchi, Davenport and Prusak, and Stewart, who define it as the “foundation of Knowledge management and Intellectual Capital field” [50] and later try to set up a starting point for the current overview.

2.1. Domain Knowledge in SE

The term domain is considered abstract since it expands the development field by adding processes supporting a given product in its early-stage development. These processes are critical in software engineering since they contribute to the scope definition, architecture styles, product development, and delivery.

Approaches that establish concepts with higher agility coefficients affirm the influence of the domain significantly when the scope is constantly changing, or the software is planned to be supported in long-term exploitation. In both cases, all decisions must be consistent within the context, technology trends, and other factors.

Most of the papers focus on how different requirements – functional and non-functional affect the decisions regarding the software structure and the technology stack used for its implementation. It is also necessary to distinguish between methodologies focused on the development process and design patterns related to the product's overall performance. Having this in mind, Requirements engineering is defined as the main driver in SE, on one hand, and Software architecture on the other hand.

To clarify the approaches applied when generating knowledge, the domain knowledge is split into Product and Process knowledge fields of research. Thus, it will be acknowledged how a product could be described and normed according to the client's expectations and how the process defined during the project's implementation could contribute to the team's efficiency by establishing a stable and easy-to-follow workflow.

2.1.1. Product knowledge

When a product is described, it is essential to define it as a documented set of components with their functionalities and relations. Descriptive languages often provide a machine-readable domain specification that could be enriched by introducing meta-data and using conflict-detecting mechanisms based on a three-step analytic process for validation, verification, and performance metrics.

In terms of the state KM is being introduced, this approach is used to transform, for example, legacy systems "into easily accessible, well described, and interoperable, modular services" [27]. Later, these descriptions could allow the team to apply changes more quickly, especially within the agile software delivery process. Thus, descriptive languages like UML, VEL [5], and WADL have a supportive role in different development designs such as Quality-driven Architecture Design (QDAD), Quality-Aware Rapid Software Development Design (Q-Rapids) [16], Transformation models in dynamic environments [56] and older pipeline dev models like PLUSS (Product Line Use case modelling for Systems and Software engineering) [1].

Meta-based models represent another perspective – focusing on the software design, the business process, and its persistence, different data objects could be categorised into relationships and correspondence models depending on the type, usability, and consistency [11], prioritization [36], effort estimation, predictions, and size measurement [42]. Compliance modelling, on the other hand, is focused on the abstract level of requirements engineering and identifies challenges in system modelling, such as "missing linkage with the business use cases" [57] or difficulties in understanding the connection between compliance requirements and design rules. This adds new ways of understanding and could be developed further into theories that could provide formalization and focus on stakeholders, opportunities, human resources, working practices, and the software system itself [14]. All these models and methods could be monitored via classification methods such as Gamification, which, when applied during the development process, improves the team's overall performance [18], Prototyping, which focuses on data science projects and its three main elements: Experimentation, Development, and self-discipline [2], or CASE which considers Technical Debt when trying to reconstruct already existing solutions [24].

In product-centric companies, knowledge networks could provide critical information via internal and external communication. For example, PISA models are based on feedback and use different rating systems so the company can "identify improvement indications for next releases" [12]. Including the user in the development process is also applied in "User-Centred Designs", used for software products whose end users have different expertise. Using the software is defined as part of their working routine [60]. Another way to establish effective product management is the product road mapping approach [35] or the Visual Milestone Planning (VMP) [15] approach, which supports the product discovery process and the agile transformation, including the stakeholders in features prioritization. This applies in industries such as Video Gaming, where virtual environment simulation is a crucial activity when validating new ideas.

2.1.2. Process knowledge

Capturing the knowledge that a given software process generates is a challenge for every team leader. Different methods that support this knowledge-gathering activity are identified by separating the Software Development Lifecycle (SDLC). There are various solutions, some focusing on how data is being handled, others – on monitoring changes and product engineering.

In the early stages of the product, the main goal is to clarify requirements and come up with a solution based on previous experiences. The portfolio-driven development model is an advanced model of APLE (Agile Product Line Engineering) and demands a more agile and rapid approach using "a collection of projects, programs, and the other operations for achieving the business goals" [25]. Based on that, collection managers can provide more accurate estimations, especially when working with fixed resources.

During the implementation phase, the focus is on the value provided and how the development process is executed – Waterfall, Iterative, V-Shaped, or Spiral. There are many models with different priorities. However, when it comes to continuous improvement, there is a need for a transition model like LACE (Lean-Agile Centre of Excellence) [8], which could remove impediments in large-scale projects.

Later during the quality assessment phase, non-functional requirements like security, maintainability, transparency, and endurance are considered. The result is a set of maturity models and concepts that target the SDLC, like the Capability Maturity Model, which aims to "evaluate and assess security engineering practices", Cybersecurity Capability Maturity Model, which is designed to help organisations "to improve their cybersecurity programs", and Software Assurance Maturity Model, which is an open framework for practice evaluation [37] which targets self-adaptive systems and their business goals. Other concepts focus on user feedback and how it is introduced to the requirements evolution [29], how the agile process's efficiency is measured [44], what metrication should be applied in order to increase the transparency in customer-relationships-based platforms [40], microservice applications [7], business intelligence systems [13] and industrial software products [45], how cybersecurity vulnerabilities could be identified via different communication models [26], and even how QA could be evaluated statistically via algorithms like Majority voting, ZenCrowd and Naïve Bayes [17].

2.2. Organisational knowledge in SE

One of the fundamental research projects is a state-of-the-art report which identifies the need of investigating the connection between how knowledge is managed and how software processes are coordinated [46]:

Needs regarding behavioural KM:

- The need for domain knowledge;
- The need for knowledge capture and share;
- The need for knowledge about local policies;
- The need for knowledge about who knows what;

Needs regarding technocratic KM:

- The need for knowledge about new technologies;
- The need for knowledge about distance collaboration;
- The need for knowledge about new challenges and opportunities;

Having that in mind, different kinds of research could be further categorised into two significant types of knowledge management. Similar classifications are discussed through the years by adding more and more details regarding relationships and possible outcomes.

2.2.1. Technocratic knowledge

Major studies point out the need for a more technocratic approach toward KM. This includes cognitive analysis, a closer look at communication within a virtual reality, event modelling, and data clustering based on common features [38]. This is essential in project management since codification and documentation are vital activities when monitoring [4]. Knowledge modelling focused on applying the conceptual and computational models when extracting functions or processing clustered data – perceived complexity and managed risks contribute to using knowledge artifacts. So, several aspects of knowledge management can be classified within this technocratic approach.

First, starting on a more global level, knowledge management systems should be focused on software products like audit tools, HR management tools, and collaboration software. For example, audit tools help the organisation with crucial activities via assets mapping, landscape mapping, flowcharts, competitive analysis, diagnostics, critical function analysis, and benefit assessment [20]. The whole audit modelling is fundamental if the company wants to improve its culture, establish an effective working environment, and use its knowledge beneficially. From a technological perspective, the KMS could be centralized and decentralized and apply techniques like knowledge discovery using surveys and audits, introducing knowledge inventorying and mapping, broadcasting knowledge, competitive analysis, and diagnostics [21].

Second, going further into info structures and data analysis, the main drivers are the knowledge itself, characterised by a specific domain, a business process with a given goal, and a context contributing to the process to be executed [22]. The system's design should be considered according to the scope, developing methodology, human resources, additional assets, KM practices, and infrastructure elements [23]. Another critical issue that should be discussed is the monitoring process which should provide abstract, meta-oriented views reflecting the organisational structure [55].

2.2.2. Behavioural knowledge

Regarding the behavioural approach, two main directions should be analysed, one wrapped around team modelling and finding patterns within larger-scale projects. The other is based on the Agile framework and how it could transform more prominent companies into strongly connected knowledge communities.

Team patterns. Team patterns usually try to identify specific behaviour types within a team that could be mapped with improved activities during work. Some of the papers observed are focused on teamwork in educational institutions where students are grouped and work in similar environments. Thus, the researchers could identify different approaches in work by prioritizing the main drivers in software development. For example, some of the process patterns include engineering-driven development, where the main drivers are planning and design, code-driven development, where coming up with prototypes is more efficient, and ad-hoc development, where the priority is dynamic and could go towards engineering, coding or verification and validation [19].

Research shows that human factors are under deep analysis, and according to literature overviews, the main clusters describe how they are introduced effectively in the SE process and how they improve it via education and motivation [32]. A similar analysis applied on a large scale identifies the need to audit existing methods and practices so teams can be more prepared and adaptable to any context. For example, Global Software Development focuses on more abstract principles defining software development as a "human-centric and socio-technical activity" and taking cultural context into account when managing larger projects [33].

Agile on a larger scale. Since agile was successfully integrated into smaller companies, managing to build compact and effective teams that could quickly deliver value according to clients' expectations, bigger enterprises try to adapt this simple framework on a larger scale. Methods like Large Scale Scrum (LeSS), which could be applied to up to "10 Scrum teams (of seven people)", and LeSS Huge, which could be applied to a "few thousand people working on one product" [48, 52] are influential, especially when companies do not pay attention at early stages and grow up rapidly. Scaled Agile Framework (SAFe) and Disciplined Agile Delivery (DAD), on the other hand, define four levels: Team level, Program level, Portfolio level, and Value stream level [41]. Some measurements identified via surveys are lead cycle and release time, the value provided, the number of defects found, velocity, automation, and predictability [30]. Others focus on maturity modelling, where levelling criteria are strongly related to the agile process's performance [49].

However, the main driver of the agile transformation is the agile coach himself – his responsibilities include "building teams by providing realistic support during implementation of agile processes, leading the team towards self-organisation" and "creating guidelines, setting goals and roadmaps" [53]. Lastly, agile transformation must be evaluated via success factors like shared product vision, shared responsibilities, shared knowledge, feedback, and ownership, and failure factors like lack of middle management support, barriers to the production environment, excessive control by the higher management, and lack of understanding from the stakeholders [51].

3. Model representation

Considering the different paradigms and aspects described in the previous section, four groups can be formed depending on the context (Table 1), and key concepts that include these knowledge domains can be identified.

The first group combines everything related to the product – studies are focused on different approaches in the development setup that correspond to the context,

Table 1. Knowledge domain groups

Domain		**Organisation**	
Product	**Process**	**Behavioural**	**Technocratic**
Design	Change	Team	Knowledge
Architecture	Improvement	Agile	Systems
Engineering	Model	Approach	Projects
Systems	Metrics	Development	Model
Software	Quality attributes	Methodologies	Practices
Data	Analysis	Scrum	Management

from architecture-leading technologies like the Internet of things and Artificial Intelligence, through domain-based concepts like Automobile industries and eLearning, to fundamental processes of software engineering like Requirements elicitation and Design patterns. This forms a wide field of discoveries about the domain that could be patterned.

The second group is formed from the idea of Software Process Improvement. These studies centre the process as a systematically arranged set of activities, each contributing to the product's development. This sets up requirements regarding quality and introduces metrics that help validate and verify the software. Deployment and Maintenance are also very critical phases of every project – approaches like DevOps and Meta-Modelling seriously impact how people can manage the product after its release.

The third group is concentrated on human resources. The need to manage intellectual capital triggers another narrative in scientific discoveries, which includes experiments on all types of projects and their teams. These experiments investigated how teams are formed in agile development, how technologies can be introduced and later improved to reflect a team's productivity, and how methodologies like Lean and Scrum tackle significant problems like low motivation, knowledge loss, and impediment documentation.

The last group is more abstract than the others – its goal is to categorize and systematize any knowledge related to the other three domains and use it strategically toward a successful closure. This includes Business patterns, Knowledge audits, and Information systems that support the organisation.

The model is represented in Figure 2, reviewing how different types of knowledge correspond to each other and formalise communications on higher and lower scales. It identifies knowledge domains and how they could be introduced according to the

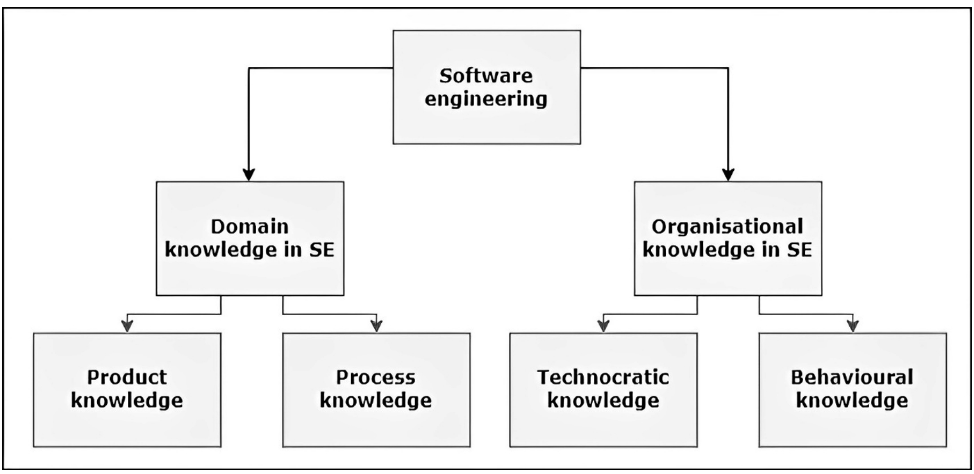


Figure 2. Knowledge types

perspectives described above, but also identifies the main aspects that drive the communication flow both horizontally and vertically through the organisation:

- **Technologies** represent all technological knowledge used for the solutions company provides to the market. Technologies depend on the set of problems developers want to tackle, the difficulty that should be reached to deliver a Minimum viable product (MVP), and the trends that influence the environment. The model compares the tech perspective to the organisation's mission and goals, as well as previous solutions developed;

- **Methodologies** represent all procedures the company applies regarding how business processes should be executed, how knowledge is transferred from one process to another, and how problems are handled to mitigate casualties. Methodologies depend on the overall dynamics, the organisational structure, what efficiency metrics have been established and how the product is being represented as components that need to be delivered. If the deployment plan requires bigger accuracy and quality is a primary driver of development, the sprints are more extended with fewer deficiencies found afterward. However, if the goal is value-oriented, shorter sprints with more frequent feedback are expected;

- **Culture** represents how the organisation introduces the working environment, and what values support its progress. This depends on the company's mission and vision and the socio-economic factors that influence the market. If it is an international company, the culture is more globally considered since it should apply to as many people as possible. But if the company is a local enterprise that works with people within a given country, the culture should be oriented according to the local mentality;

- **Human resources** represent all employees' experience and abilities as assets that should be monitored and used to fulfil company's goals. They depend on the technology stack defined, the culture the organisation establishes, and the environment considered internally and externally.

4. Discussion

Several trends can be analysed, further developed, and enriched. First, KM has a variety of applications on different levels, both from managerial and functional perspectives. These aspects are cross-examined by describing the four different types of knowledge, which shows the approaches and trends discussed in the previous sections and how they support the KM process (Figure 3).

Technocratic knowledge a given company relies on supports the process knowledge by setting up procedures everybody should know, and quality standards teams should strictly comply with. By introducing a knowledge modelling structure that correlates to the organisational one, different methodologies applied on a larger scale, like Agile, could confirm a well-described, balanced, and customised set of workflows,

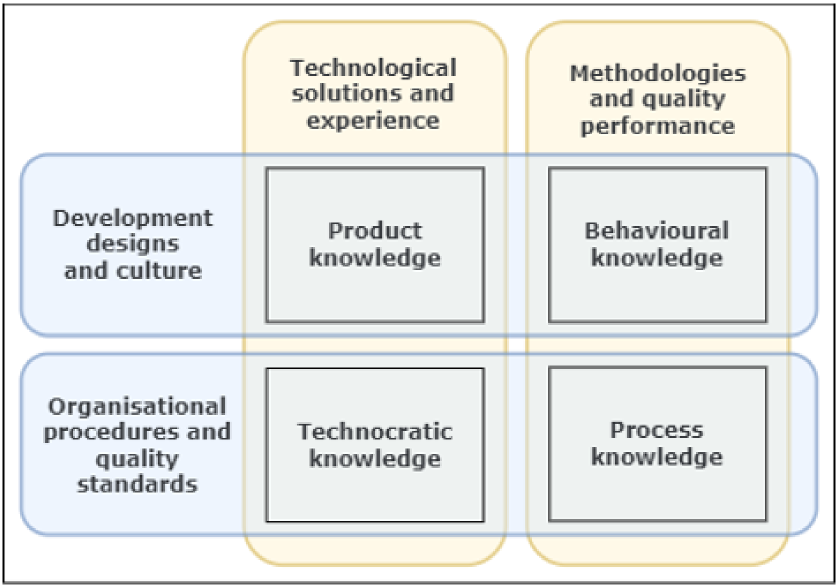


Figure 3. Knowledge types relations

metrics, and instructions managers can use when starting a new project. This sets the foundation that could guarantee a good start for every development process.

Technocratic knowledge could also support product knowledge by suggesting technological solutions based on previous experiences and saving new experiences, using them to improve the company's reputation according to the dynamic ICT environment. This is critical for industries whose progress counts on R&D and concurrency between different enterprises.

On the other hand, behavioural knowledge focuses on how these solutions mentioned above correspond to the organisational culture and the company's vision. In terms of product knowledge, this results in development designs and teams that split functionalities according to a given semantics and tackle technological problems by setting up a CI process that supports the client's expectations. In process knowledge, however, the essential part stands for quality assessment and how methodologies guarantee the proper execution of the development itself.

By further discussing how this categorisation could be expanded via local and global perspectives, problems are meant to be handled from top-down and bottom-up approaches. As Table 2 shows, knowledge is used locally for specific projects and on a global scale by adding new experiences or improving overall performance. This KM concept is relatively simple yet effective since there is a clear division between the management and the development that could be handled by a customised communication strategy, for which bigger experience is needed.

Table 2. Knowledge representation

		Technocratic knowledge		Behavioural knowledge	
		Global scale	Local scale	Global scale	Local scale
Product knowledge	Global scale	Experience and portfolio knowledge		Organisational culture	
	Local scale		Technological solutions		Development designs
Process knowledge	Global scale	Quality standards		Quality performance	
	Local scale		Organisational procedures		Methodologies

5. Conclusion and future work

In conclusion of this overview, Knowledge management has a meaningful role in data flows not only on an organisational level but also on a project level. It can be argued, especially in small and medium-sized companies, that the organisational structure aims to evolve towards the communication channels – a process described by Conway's law. This corresponds to the vital process that started in the 90s with the first methodologies, developed further in the 2000s, and formalized fully nowadays.

Requirements should be more descriptive and considered when making software structure decisions. When starting a project, most of the problems are related to how teams describe the domain, how the context is perceived, and how communications are established. If the scope is clear and well-defined, future problems could be predicted not only about the development process but also about human resources. This is a problem KM supports by introducing different types of solutions related to the technological perspective on the one hand – KMS, automation, and knowledge migration, and to the cultural perspective on the other – patterns, procedures, and learning by sharing methods.

It is critical to highlight the role of the business itself – more and more companies delegate resources to investigate their data flows and how knowledge has been spread among team members. This could be used for updating the organisation's culture, mission, and goals. In addition, big companies generate statistical data that could be used for academic research, resulting in theories and taxonomies described in the public literature.

In the future, these relations will be observed not only on the theoretical level but also in practice. Knowledge contributes critical value to a company's success and will be further observed as enabling factor of shared learning and growth.

Appendix

Id	Citation	Keywords
P1	Ragab and Arisha [43]	knowledge measurement, knowledge management
P2	Bjørnson and Dingsøyr [6]	software engineering, knowledge management, learning software organization, software process improvement, systematic review
P3	Centobelli et al. [9]	entrepreneurship, factors affecting KM, KMSs, knowledge management, performance, start-up firms, scalability
P4	Céspedes et al. [10]	systematic literature review, DevOps, product quality, ISO/IEC 25000
P5	M. Asrar-ul-Haq et al. [3]	knowledge management, knowledge sharing, antecedents, trends
P6	Iskandar et al. [27]	knowledge management system, KMS, current issues, systematic literature review, big data issue in KMS
P7	Kiv et al. [30]	agile manifesto, agile methods, agile methods adoption, partial agile adoption, systematic literature review
P8	Marinho et al. [32]	global software development, global teams, culture, systematic literature review
P9	Mora et al. [33]	agile paradigm tenets, agile ITSM methods, agile software engineering methods, FitSM, IT4IT, representative literature analysis
P10	Ouriques et al. [38]	knowledge management, agile software development, knowledge processes
P11	Saad and Zainudin [47]	computational thinking, project-based learning, PBL-CT, teaching and learning strategies
P12	Serenko and Bontis [50]	knowledge management, process management, intellectual capital
P13	Stray et al. [53]	agile coaching, skills, tasks, systematic literature review, agile transformation, software development practices
P14	Theobald et al. [54]	agile leadership, agile management, agile organization, motivation, systematic literature review
P15	Venkitachalam and Busch [58]	tacit knowledge, implicit knowledge, knowledge management, research
P16	Wang and Noe [59]	knowledge sharing, knowledge exchange, knowledge management

References

[1] F. Ahmed and L. F. Capretz, The software product line architecture: An empirical investigation of key process activities, Inf. Softw. Technol. 50(11) (2008) 1098–1113, https://doi.org/10.1016/j.infsof.2007.10.013.

[2] T. Aho, O. Sievi-Korte, T. Kilamo, S. Yaman and T. Mikkonen, Demystifying data science projects: A look on the people and process of data science today, in: Product-Focused Software Process Improvement (PROFES 2020), ed. by M. Morisio et al., LNCS 12562, Springer, Cham, 2020, 153–167, https://doi.org/10.1007/978-3-030-64148-1_10.

[3] M. Asrar-ul-Haq, S. Anwar and T. Nisar, A systematic review of knowledge management and knowledge sharing: Trends, issues, and challenges, Cogent Business & Management 3(1) https://doi.org/10.1080/23311975.2015.1127744.

[4] V. C. Ayarza and S. Bayona-Oré, Cluster monitoring and integration in technology company, in: Trends and Applications in Software Engineering (CIMPS 2019), ed. by J. Mejia et al., Adv. Intell. Syst. Comput. 1071, Springer, Cham, 2020, 253–265, https://doi.org/10.1007/978-3-030-33547-2_19.

[5] A. Bagnato et al., Showcasing Modelio and pure:variants Integration in REVaMP^2 Project, in: Product-Focused Software Process Improvement (PROFES 2019), ed. by X. Franch et al., LNCS 11915, Springer, Cham, 2019, 590–595, https://doi.org/10.1007/978-3-030-35333-9_43,

[6] F. O. Bjørnson and T. Dingsøyr, Knowledge management in software engineering: A systematic review of studied concepts, findings and research methods used, Inf. Softw. Technol. 50(11) (2008) 1055–1068.

[7] J. Bogner, S. Schlinger, S. Wagner, and A. Zimmermann, A Modular Approach to Calculate Service-Based Maintainability Metrics from Runtime Data of Microservices, in: Product-Focused Software Process Improvement (PROFES 2019), ed. by X. Franch et al., LNCS 11915, Springer, Cham, 2019, 489–496, https://doi.org/10.1007/978-3-030-35333-9_34.

[8] J. Bowring and M. Paasivaara, Keeping the Momentum: Driving Continuous Improvement after the Large-Scale Agile Transformation, in: Product-Focused Software Process Improvement (PROFES 2021), ed. by L. Ardito et al., LNCS 13126, Springer, Cham, 2021, 66–82, https://doi.org/10.1007/978-3-030-91452-3_5.

[9] P. Centobelli, R. Cerchione and E. Esposito, Knowledge management in startups: Systematic literature review and future research agenda, Sustainability 9(3) (2017) 361, https://doi.org/10.3390/su9030361.

[10] D. Céspedes, P. Angeleri, K. Melendez and A. Dávila, Software product quality in DevOps contexts: A systematic literature review, in: Trends and Applications in Software Engineering (CIMPS 2019), ed. by J. Mejia et al., Advances in Intelligent Systems and Computing 1071, Springer, Cham, 2019, 51–64.

[11] M. El Hamlaoui, S. Bennani, S. Ebersold, M. Nassar, and B. Coulette, AHM: Handling heterogeneous models matching and consistency via MDE, in: Evaluation of Novel Approaches to Software Engineering (ENASE 2018), ed. by E. Damiani et al., Communications in Computer and Information Science 1023, Springer, Cham, 2018, 288–313, https://doi.org/10.1007/978-3-030-22559-9_13.

[12] F. Falcini and G. Lami, Embracing software process improvement in automotive through PISA model, in: Product-Focused Software Process Improvement (PROFES 2019), ed. by X. Franch et al., LNCS 11915 Springer, Cham, 2019, 73–88, https://doi.org/10.1007/978-3-030-35333-9_5.

[13] I. Figalist, C. Elsner, J. Bosch and H. Olsson, An end-to-end framework for productive use of machine learning in software analytics and business intelligence solutions, in: Product-Focused Software Process Improvement (PROFES 2020), ed. by M. Morisio et al., LNCS 12562, Springer, Cham, 2020, 217–233, https://doi.org/10.1007/978-3-030-64148-1_14.

[14] J. Fischbach, J., Frattini, D. Mendez, M. Unterkalmsteiner, H. Femmer and A. Vogelsang, How do practitioners interpret conditionals in requirements?, in: Product-Focused Software Process Improvement (PROFES 2021), ed. by L. Ardito et al., LNCS 13126, Springer, Cham, 2021, 85–102, https://doi.org/10.1007/978-3-030-91452-3_6.

[15] D. Fontdevila, M. Genero, A. Oliveros and N. Paez, Evaluating the utility of the usability model for software development process and practice, in: Product-Focused Software Process Improvement (PROFES 2019), ed. by X. Franch et al., LNCS 11915, Springer, Cham, 2019, 741–757, https://doi.org/10.1007/978-3-030-35333-9_57.

[16] X. Franch, L. Lopez, S. Martínez-Fernández, M. Oriol, P. Rodríguez and A. Trendowicz, Quality-aware rapid software development project: The Q-rapids project,

in: Software Technology: Methods and Tools (TOOLS 2019), ed. by M. Mazzara et al., LNCS 11771, Springer, Cham, 2019, 378–392, https://doi.org/10.1007/978-3-030-29852-4_32.

[17] T. Fredriksson, D. I. Mattos, J. Bosch and H. H. Olsson, Data labeling: An empirical investigation into industrial challenges and mitigation strategies, in: Product-Focused Software Process Improvement (PROFES 2020), ed. by M. Morisio et al., LNCS 12562, Springer, Cham, 2020, 202–216, https://doi.org/10.1007/978-3-030-64148-1_13.

[18] G. A. García-Mireles and M. E. Morales-Trujillo, Gamification in software engineering: A tertiary study, in: Trends and Applications in Software Engineering (CIMPS 2019), ed. by J. Mejia et al., Adv. Intell. Syst. Comput. 1071, Springer, Cham, 2020, 116–128, https://doi.org/10.1007/978-3-030-33547-2_10.

[19] E. Germain and P. N. Robillard, Towards software process patterns: An empirical analysis of the behaviour of student teams, Inf. Softw. Technol. 50(11) (2008) 1088–1097, https://doi.org/10.1016/j.infsof.2007.10.018.

[20] E. Gourova, A. Antonova and Y. Goleminova, Knowledge audit concepts, processes and practice, WSEAS Trans. Bus. Econ. 6 (2009) 605–619.

[21] E. Gourova and M. Dragomirova, Design of knowledge management info-structures, in: EuroPLoP'15: Proc. 20th Eur. Conf. on Pattern Languages of Programs, Art. No 15 (2015) 9 pp.

[22] E. Gourova and Y. Todorova, Knowledge audit data gathering and analysis, in: EuroPLoP'10, Art. No 14 (2010) 7 pp.

[23] E. Gourova and K. Toteva, Design of knowledge management systems. VikingPLoP 2014, Art. No 3 (2014) 15 pp.

[24] D. Guamán, J. Pérez, J. Garbajosa and G. Rodríguez, A systematic-oriented process for tool selection: The case of green and technical debt tools in architecture reconstruction, in: Product-Focused Software Process Improvement (PROFES 2020), ed. by M. Morisio et al., LNCS 12562, Springer, Cham, 2020, 237–253, https://doi.org/10.1007/978-3-030-64148-1_15.

[25] K. Hayashi and M. Aoyama, A portfolio-driven development model and its management method of agile product line engineering applied to automotive software development, in: Product-Focused Software Process Improvement (PROFES 2020), ed. by M. Morisio et al., LNCS 12562, Springer, Cham, 2020, 88–105, https://doi.org/10.1007/978-3-030-64148-1_6.

[26] M. Hell and M. Höst, Communicating Cybersecurity Vulnerability Information: A Producer-Acquirer Case Study, in: Product-Focused Software Process Improvement (PROFES 2021), ed. by L. Ardito et al., LNCS 13126, Springer, Cham, 2021, 215–230, https://doi.org/10.1007/978-3-030-91452-3_15.

[27] S. M. Huang, Y. T. Chu, S. H. Li and D. C. Yen, Enhancing conflict detecting mechanism for Web Services composition: A business process flow model transformation approach, Inf. Softw. Technol. 50(11) (2008) 1069–1087, https://doi.org/10.1016/j.infsof.2007.10.014.

[28] K. Iskandar, M. Jambak, R. Kosala and H. Prabowo, Current Issue on Knowledge Management System for future research: A systematic literature review, Procedia Comput. Sci. 116(C) (2017) 68–80, https://doi.org/10.1016/j.procs.2017.10.011.

[29] J. O. Johanssen, A. Kleebaum, B. Bruegge and B. Paech, Feature Crumbs: Adapting usage monitoring to continuous software engineering, in: Product-Focused Software Process Improvement (PROFES 2018), ed. by M. Kuhrmann et al., LNCS

11271, Springer, Cham, 2018, 263–271, https://doi.org/10.1007/978-3-030-03673-7_19.

[30] P. Kettunen, M. Laanti, F. Fagerholm and T. Mikkonen, Agile in the era of digitalization: A finnish survey study, in: Product-Focused Software Process Improvement (PROFES 2019), ed. by X. Franch et al. LNCS 11915, Springer, Cham, 2019, 383–398, https://doi.org/10.1007/978-3-030-35333-9_28.

[31] S. Kiv, S. Heng, M. Kolp and Y. Wautelet, Agile manifesto and practices selection for tailoring software development: A systematic literature review, in: Product-Focused Software Process Improvement (PROFES 2018), ed. by M. Kuhrmann et al., LNCS 11271, Springer, Cham, 2018, 12–30, https://doi.org/10.1007/978-3-030-03673-7_2.

[32] L. Machuca-Villegas and G. P. Gasca-Hurtado, Towards a social and human factor classification related to productivity in software development teams, in: Trends and Applications in Software Engineering (CIMPS 2019), ed. by J. Mejia et al., Adv. Intell. Syst. Comput. 1071, Springer, Cham, 2020, 36–50, https://doi.org/10.1007/978-3-030-33547-2_4.

[33] M. Marinho, A. Luna and S. Beecham, Global software development: Practices for cultural differences, in: Product-Focused Software Process Improvement (PROFES 2018), ed. by M. Kuhrmann et al., LNCS 11271, Springer, Cham, 2018, 299–317, https://doi.org/10.1007/978-3-030-03673-7_22.

[34] M. Mora, F. Wang, J. M. Gómez and O. Díaz, A comparative review on the agile tenets in the IT service management and the software engineering domains, in: Trends and Applications in Software Engineering (CIMPS 2019), ed. by J. Mejia et al., Adv. Intell. Syst. Comput. 1071, Springer, Cham, 2019, 102–115, https://doi.org/10.1007/978-3-030-33547-2_9.

[35] J. Münch, S. Trieflinger and D. Lang, What's hot in product roadmapping? Key practices and success factors, in: Product-Focused Software Process Improvement (PROFES 2019), ed. by X. Franch et al. LNCS 11915, Springer, Cham, 2019, 401–416, https://doi.org/10.1007/978-3-030-35333-9_29.

[36] E. Nazaruka and J. Osis, The formal reference model for software requirements, in: Evaluation of Novel Approaches to Software Engineering (ENASE 2018), ed. by E. Damiani et al., Communications in Computer and Information Science 1023, Springer, Cham, 2019, 352–372, https://doi.org/10.1007/978-3-030-22559-9_16.

[37] P. Nikbakht, M. Höst, and M. Hell, HAVOSS: A maturity model for handling vulnerabilities in third party OSS components, in: Product-Focused Software Process Improvement (PROFES 2018), ed. by M. Kuhrmann et al., LNCS 11271, Springer, Cham, 2018, 81–97, https://doi.org/10.1007/978-3-030-03673-7_6.

[38] S. Osorio, A. P. P. Negrón, and A. E. Valdez, From a conceptual to a computational model of cognitive emotional process for engineering students, in: Trends and Applications in Software Engineering (CIMPS 2019), ed. by J. Mejia et al., Adv. Intell. Syst. Comput. 1071, Springer, Cham, 2020, 173–186, https://doi.org/10.1007/978-3-030-33547-2_14

[39] R. Ouriques, K. Wnuk, T. Gorschek and R. Berntsson, Knowledge management strategies and processes in agile software development: A systematic literature review, Int. J. Softw. Eng. Knowl. Eng. 29(03) (2019) 345–380, https://doi.org/10.1142/S0218194019500153.

[40] C. R. Prause and A. Hönle, Emperor's new clothes: Transparency through metrication in customer-supplier relationships, in: Product-Focused Software Process

Improvement (PROFES 2018), ed. by M. Kuhrmann et al., LNCS 11271, Springer, Cham, 2018, 288–296, https://doi.org/10.1007/978-3-030-03673-7_21.

[41] A. Putta, M. Paasivaara and C. Lassenius, Benefits and challenges of adopting the Scaled Agile Framework (SAFe): Preliminary results from a multivocal literature review, in: Product-Focused Software Process Improvement (PROFES 2018), ed. by M. Kuhrmann et al., LNCS 11271, Springer, Cham, 2018, 334–351, https://doi.org/10.1007/978-3-030-03673-7_24.

[42] C. Quesada-López, A. Martínez, M. Jenkins, L. C. Salas and J. C. Gómez, Automated functional size measurement: A multiple case study in the industry, in: Product-Focused Software Process Improvement (PROFES 2019), ed. by X. Franch et al., LNCS 11915, Springer, Cham, 2019, 263–279, https://doi.org/10.1007/978-3-030-35333-9_19.

[43] M. A. F. Ragab and A. Arisha, Knowledge management and measurement: a critical review, J. Knowl. Manag. 17(6) (2013) 873–901, https://doi.org/10.1108/JKM-12-2012-0381.

[44] P. Ram et al., An empirical investigation into industrial use of software metrics programs, in: Product-Focused Software Process Improvement (PROFES 2020), ed. by M. Morisio et al., LNCS 12562, Springer, Cham, 2020, 419–433, https://doi.org/10.1007/978-3-030-64148-1_26.

[45] P. Ram, P. Rodriguez and M. Oivo, Software process measurement and related challenges in agile software development: A multiple case study, in: Product-Focused Software Process Improvement (PROFES 2018), ed. by M. Kuhrmann et al., LNCS 11271, Springer, Cham, 2018, 272–287, https://doi.org/10.1007/978-3-030-03673-7_20.

[46] I. Rus, M. Lindvall and S. S. Sinha, Knowledge management in software engineering. A DACS state-of-the-art report, IEEE Software 19(3) (2002) 26–38, https://doi.org/10.1109/MS.2002.1003450.

[47] A. Saad and S. Zainudin, A review of Project-Based Learning (PBL) and Computational Thinking (CT) in teaching and learning, Learning and Motivation 78 (2022) 101802.

[48] A. Salameh and J. Bass, Influential factors of aligning Spotify squads in mission-critical and offshore projects – A longitudinal embedded case study, in: Product-Focused Software Process Improvement (PROFES 2018), ed. by M. Kuhrmann et al., LNCS 11271, Springer, Cham, 2018, 199–215, https://doi.org/10.1007/978-3-030-03673-7_15.

[49] A. Schmitt, S. Theobald and P. Diebold, Comparison of agile maturity models, in: Product-Focused Software Process Improvement (PROFES 2019), ed. by X. Franch et al., LNCS 11915, Springer, Cham, 2019, 661–671, https://doi.org/10.1007/978-3-030-35333-9_52.

[50] A. Serenko and N. Bontis, Meta-review of knowledge management and intellectual capital literature: Citation impact and research productivity rankings, Knowl. Process Manag. 11(3) (2004) 185–198.

[51] I. Signoretti et al., Success and failure factors for adopting a combined approach: A case study of two software development teams, in: Product-Focused Software Process Improvement (PROFES 2020), ed. by M. Morisio et al., LNCS 12562, Springer, Cham, 2020, 125–141, https://doi.org/10.1007/978-3-030-64148-1_8.

[52] J. P. Steghöfer, E. Knauss, J. Horkoff and R. Wohlrab, Challenges of scaled agile for safety-critical systems, in: Product-Focused Software Process Improvement

(PROFES 2019), ed. by X. Franch et al., LNCS 11915, Springer, Cham, 2019, 350–366, https://doi.org/10.1007/978-3-030-35333-9_26.

[53] V. Stray, B. Memon and L. Paruch, A systematic literature review on agile coaching and the role of the agile coach, in: Product-Focused Software Process Improvement (PROFES 2020), ed. by M. Morisio et al., LNCS 12562, Springer, Cham, 2020, 3–19, https://doi.org/10.1007/978-3-030-64148-1_1.

[54] S. Theobald, N. Prenner, A. Krieg and K. Schneider, Agile leadership and agile management on organisational level – A systematic literature review, in: Product-Focused Software Process Improvement (PROFES 2020), ed. by M. Morisio et al., LNCS 12562, Springer, Cham, 2020, 20–36, https://doi.org/10.1007/978-3-030-64148-1_2.

[55] V. Torres, M. Gil and V. Pelechano, Software knowledge representation to understand software systems, in: Product-Focused Software Process Improvement (PROFES 2019), ed. by X. Franch et al., LNCS, vol, 11915, Springer, Cham, 2019, 137–144 https://doi.org/10.1007/978-3-030-35333-9_10.

[56] S. Trieflinger, J. Münch, S. Wagner, D. Lang and B. Roling, A transformation model for excelling in product roadmapping in dynamic and uncertain market environments, in: Product-Focused Software Process Improvement (PROFES 2021), ed. by L. Ardito et al., LNCS 13126, Springer, Cham, 2021, 136–151, https://doi.org/10.1007/978-3-030-91452-3_9.

[57] M. Usman, M. Felderer, M. Unterkalmsteiner, E. Klotins, D. Mendez and E. Alégroth, Compliance requirements in large-scale software development: An industrial case study, in: Product-Focused Software Process Improvement (PROFES 2020), ed. by M. Morisio et al., LNCS 12562, Springer, Cham, 2020, 385–401, https://doi.org/10.1007/978-3-030-64148-1_24.

[58] K. Venkitachalam and P. Busch, Tacit knowledge: review and possible research directions, J. Knowl. Manag. 16(2) (2012) 356–371.

[59] Sh. Wang and R. A. Noe, Knowledge sharing: A review and directions for future research, Hum. Resour. Manag. Rev. 20 (2010) 115–131.

[60] M. Winterer, C. Salomon, G. Buchgeher, M. Zehethofer and A. Derntl, Establishing a user-centered design process for human-machine interfaces: Threats to success, in: Product-Focused Software Process Improvement (PROFES 2019), ed. by X. Franch et al., LNCS 11915, Springer, Cham, 2019, 89–102, https://doi.org/10.1007/978-3-030-35333-9_6.

Received on December 20, 2023
Accepted on March 8, 2024

Dilyan Georgiev

Faculty of Mathematics and Informatics
Sofia University "St. Kliment Ohridski"
5, James Bourchier Blvd.
1164 Sofia
BULGARIA

E-mail: diljang@fmi.uni-sofia.bg

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“
ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Том 111

ANNUAL OF SOFIA UNIVERSITY “ST. KLIMENT OHRIDSKI”
FACULTY OF MATHEMATICS AND INFORMATICS
Volume 111

THE INTEGRABILITY OF THE GENERALIZED HAMILTONIAN SYSTEM

BOGDANA A. GEORGIEVA

This paper reviews the generalized Hamiltonian system and its connection to contact transformations. The generalized Hamiltonian system is related to Herglotz variational principle in the same way in which the Hamiltonian system is related to the classical variational principle. We prove a criterion for the integrability of the generalized Hamiltonian system in terms of a complete set of first integrals, and a method of generating such first integrals. These results are due to Gustav Herglotz.

Keywords: Herglotz variational principle, integrable systems, contact transformations, integrability, complete integrability, generalized Hamiltonian system

2020 Mathematics Subject Classification: 37J06, 37J35, 37J55

1. Introduction

In 1932 Gustav Herglotz gave a series of lectures on contact transformations, the generalized Hamiltonian system

$$\begin{aligned}
\frac{d}{dt}x_j &= \frac{\partial \mathcal{H}}{\partial p_j},\\
\frac{d}{dt}z &= p_j \frac{\partial \mathcal{H}}{\partial p_j} - \mathcal{H},\\
\frac{d}{dt}p_j &= -\frac{\partial \mathcal{H}}{\partial x_j} - p_j \frac{\partial \mathcal{H}}{\partial z}, \quad j = 1, \dots, n,
\end{aligned}$$

where $\mathcal{H}$ is a function of $x_1, \dots, x_n, z, p_1, \dots, p_n$, and the relationship between them. The generalized Hamiltonian system is closely related to the variational principle, proposed by Herglotz [15, 16]. It is very powerful for giving a variational description

DOI: 10.60063/GSU.FMI.111.55-76

of nonconservative processes involving one independent variable. It is more general than the classical variational principle with one independent variable and contains it as a special case.

In the variational principle of Herglotz the functional z, whose extrema are sought, is defined by an ordinary differential equation rather than by an integral:

$$\frac{dz}{dt} = L(t, x, \dot{x}, z), \quad 0 \leq t \leq s,$$

where t is the only independent variable, $x \equiv (x^1, \dots, x^n)$ are the argument functions of t, $\dot{x} = dx/dt$. We denote $z = z[x; s]$. Herglotz showed that the value of this functional is an extremum when its argument-functions $x^k(t)$ are solutions of the generalized Euler-Lagrange equations

$$\frac{\partial L}{\partial x^k} - \frac{d}{dt}\frac{\partial L}{\partial \dot{x}^k} + \frac{\partial L}{\partial z}\frac{\partial L}{\partial \dot{x}^k} = 0, \quad k = 1, \dots, n.$$

His lectures revealed the remarkable geometry which underlines the generalized Hamiltonian system and its integrability in terms of a complete set of first integrals. They provide a method for generating first integrals for such systems. In the present paper we review these results. The summation convention on repeated indices is used throughout the paper.

Furta et al. show in [4] a close link between the Herglotz variational principle and control and optimal control theories. It is also related to contact transformations, see Guenther et al. [14]. Herglotz's work was motivated by ideas from S. Lie [17,18] and others. For historical remarks through 1935 see Caratheodory [2]. The contact transformations, which can be derived from the generalized variational principle, have found applications in thermodynamics. Mrugala shows in [20] that the processes in equilibrium thermodynamics can be described by successions of contact transformations acting in a suitably defined thermodynamic phase space. The latter is endowed with a contact structure, closely related to the symplectic structure. In [5] and [7] Georgieva et al. formulated and proved first and second Noether-type theorems which yields a first integral corresponding to a known symmetry of the functional defined by the Herglotz variational principle; and an identity corresponding to an infinite-dimensional symmetry of the Herglotz functional. For a summary of the resent results related to the variational principle of Herglotz see [9].

In [6] Georgieva, Guenther and Bodurov introduce a new variational principle, which extends the Herglotz principle to one with several independent variables. In honor of Gustav Herglotz they named it in his name. This new varational principle contains as special cases both the classical variational principle with several independent variables and the Herglotz variational principle. It can describe not only all physical processes which the classical variational principle can, but also many others for which the classical variational principle is not applicable. It can give a variational description of nonconservative processes involving physical fields.

The *generalized variational principle with several independent variables* is as follows:

Let the functional $z = z[u; s]$ of $u = u(t, x)$ be defined by an integro-differential equation of the form

$$\frac{dz}{dt} = \int_{\Omega} \mathcal{L}(t, x, u, u_t, u_x, z)\, d^n x, \quad 0 \le t \le s, \tag{1.1}$$

where t and $x \equiv (x^1, \ldots, x^n)$ are the independent variables, $u \equiv (u^1, \ldots, u^m)$ are the argument functions, $u_x \equiv (u_x^1, \ldots, u_x^m)$, $u_t \equiv (u_t^1, \ldots, u_t^m)$ and $u_x^i \equiv (u_{x^1}^i, \ldots, u_{x^n}^i)$, $i = 1, \ldots, m$, $d^n x \equiv dx^1 \ldots dx^n$, and where the function $\mathcal{L}$ is at least twice differentiable with respect to u_x, u_t and once differentiable with respect to t, x, z. Let $\eta \equiv (\eta^1(t, x), \ldots, \eta^m(t, x))$ have continuous first derivatives and otherwise be arbitrary except for the boundary conditions:

$$\begin{aligned} &\eta(0, x) = \eta(s, x) = 0, \\ &\eta(t, x) = 0 \quad \textit{for} \quad x \in \partial\Omega,\ 0 \le t \le s, \end{aligned}$$

where $\partial\Omega$ is the boundary of Ω. Then, the value of the functional $z[u; s]$ is an extremum for functions u which satisfy the condition

$$\left.\frac{d}{d\varepsilon} z[u + \varepsilon\eta; s]\right|_{\varepsilon=0} = 0.$$

The function $\mathcal{L}$, just as in the classical case, is called the *Lagrangian density*. It should be observed that when a variation $\varepsilon\eta$ is applied to u, the integro-differential equation defining the functional z must be solved with the same fixed initial condition $z(0)$ at $t = 0$ and the solution evaluated at the same fixed final time $t = s$ for all varied argument functions $u + \varepsilon\eta$.

Every function $u \equiv (u^1, \ldots, u^m)$, for which the functional z defined by the integro-differential equation (1.1) has an extremum, is a solution of

$$\frac{\partial \mathcal{L}}{\partial u^i} - \frac{d}{dt}\frac{\partial \mathcal{L}}{\partial u_t^i} - \frac{d}{dx^k}\frac{\partial \mathcal{L}}{\partial u_{x^k}^i} + \frac{\partial \mathcal{L}}{\partial u_t^i}\int_{\Omega} \frac{\partial \mathcal{L}}{\partial z}\, dx = 0, \quad i = 1, \ldots, m. \tag{1.2}$$

These equations are called (in correspondence with the classical case) the generalized Euler-Lagrange equations.

It is important to observe that the definition of the functional z by the integro-differential equation reduces to the classical definition of a functional by an integral when $\mathcal{L}$ does not depend on z. Similarly, the generalized Euler-Lagrange equations reduce to the classical Euler-Lagrange equations when $\mathcal{L}$ does not depend on z.

Many examples of physical processes described with the generalized variational principle of Herglotz are available in the papers [3, 5–10, 12]. Here we give two applications for the convenience of the reader:

The first is the set of equations which describe the propagation of electromagnetic waves in a conductive medium

$$c^2\nabla^2\mathbf{E} - \frac{\partial^2\mathbf{E}}{\partial t}^2 - \frac{\sigma}{\varepsilon}\frac{\partial \mathbf{E}}{\partial t} = 0, \tag{1.3}$$

where $\mathbf{E} = (E^1, E^2, E^3)$ is the electric field vector, c is the velocity of the electromagnetic waves, σ is the electrical conductivity and ε is the dielectric constant of the medium. Exactly the same equation holds for the magnetic field vector $\mathbf{B} = (B^1, B^2, B^3)$. These equations are direct consequence of the Maxwell's equations in conjunction with the medium's property equations $\mathbf{J} = \sigma\mathbf{E}$ and $\rho = 0$, where $\mathbf{J} = (J^1, J^2, J^3)$ is the current density and ρ is the charge density. Equation (1.3) and the equation for the magnetic vector field $\mathbf{B} = (B^1, B^2, B^3)$ *can not* be described variationally via the classical variational principle, because their Frechet derivative is not selfadjoint. Remarkably, they can be described variationally via the variational principle proposed in [6]. In more detail:

One can easily verify that this system is the system of generalized Euler-Lagrange equations for the functional z defined by the integro-differential equation with

$$\mathcal{L} = c^2 \frac{\partial E^i}{\partial x^j}\frac{\partial E^i}{\partial x^j} - \frac{\partial E^i}{\partial t}\frac{\partial E^i}{\partial t} + \alpha(x) z, \quad i, j = 1, 2, 3$$

and

$$\frac{\sigma}{\varepsilon} = \int_\Omega \alpha(x)\, d^3x = \text{const.}$$

As a second example of a physical process which *can not* be described variationally via the classical variational principle, but can be given a variational description with the variational principle of Bodurov [6] consider the nonlinear Schrödinger equation with electromagnetic interaction and losses or gains

$$i\frac{\partial \Psi}{\partial t} - \Phi\Psi + \mu\left(\frac{\partial}{\partial x^k} - iA_k\right)^2 \Psi - G(\Psi\Psi^*, x)\Psi - \beta\frac{i}{2}\Psi = 0, \quad \beta = \text{const}$$

for the wave function $\Psi(t, x^1, x^2, x^3)$ with electromagnetic interaction and losses or gains, where the summation index $k = 1, 2, 3$. Here $(\Phi(t, x^1, x^2, x^3), \mathbf{A}(t, x^1, x^2, x^3))$ is the electromagnetic potential, G is a real-valued function, and $\mathbf{A}$ is the vector potential $\mathbf{A} = (A_1, A_2, A_3)$. The losses ($\beta > 0$) or gains ($\beta < 0$) are represented with the term $-\beta\frac{i}{2}\Psi$. This equation does not have a variational description with the classical variational principle, because its Frechet derivative operator is not self-adjoint. In [3] such a description is presented for this process via the generalized variational principle of Herglotz with several independent variables, due to Bodurov. In [3] it is shown that the functional z is invariant under the gauge transformation

$$\Phi' = \Phi - \frac{\partial g}{\partial t}, \quad \mathbf{A}' = \mathbf{A} + \nabla_x g, \quad \Psi' = e^{ig}\Psi,$$

where $g = g(t, x^1, x^2, x^3)$ is an arbitrary function, and an identity is found using the main theorem in [12], due to Georgieva and Bodurov, which is satisfied by the four-potential

$$(\Phi(t, x^1, x^2, x^3), \mathbf{A}(t, x^1, x^2, x^3))$$

of the electromagnetic field. When the wave function Ψ is a solution to the nonlinear Schrödinger equation with electromagnetic interaction and losses or gains, this identity becomes

$$\frac{\partial Q_0}{\partial t} - \nabla_x \cdot \mathbf{Q} - \beta Q_0 = 0,$$

where $\mathbf{Q} = (Q_1, Q_2, Q_3)$. In the classical case when $\beta = 0$ this is the common conservation law with Q_0-conserved density and $\mathbf{Q}$ conserved current. When $\beta \neq 0$ this identity becomes a continuity law – Q_0 is not conserved but generated or dissipated (depending on the sign of β) at a rate proportional to Q_0 itself.

In [10] Georgieva introduces a method for finding the variational symmetries of the functional in the generalized variational principle with several independent variables proposed by Bodurov et al. in [6]. In [12] Georgieva and Bodurov formulate and prove a theorem which gives an identity corresponding to an infinite-dimensional symmetry of that functional.

After this brief overview of the variational principle of Herglotz and its generalization to one with several independent variables, let us return to contact transformations, the generalized Hamiltonian system and their connection to the variational principle of Herglotz.

Let S be a continuously differentiable one-to-one transformation defined on a domain of $\mathbf{R}^n \times \mathbf{R}^1 \times \mathbf{R}^n$ with range in $\mathbf{R}^n \times \mathbf{R}^1 \times \mathbf{R}^n$ which we write in the form

$$S(x, z, p) = (X(x, z, p), Z(x, z, p), P(x, z, p)).$$

We assume that both it and its inverse are sufficiently differentiable so that the computations below make sense, and that the Jacobian is distinct than zero. Such a transformation is called an *element transformation*.

Definition 1.1. A *contact transformation* is an element transformation which is one-to-one, on to, and for which $p \cdot dx - dz = 0$ implies $P \cdot dX - dZ = 0$.

Theorem 1.1. *Equation* (1.1) *represents a contact transformation if and only if there is a function* $\rho = \rho(x, z, p) \neq 0$ *such that* $P \cdot dX - dZ = \rho(p \cdot dx - dz)$.

The proof can be found in [14].

Example 1.1. The Legendre transformation in 3-dimensional space

$$X = p, \quad Y = q, \quad Z = px + qy - z, \quad P = x, \quad Q = y$$

is a contact transformation, with $\rho = -1$.

2. Special contact transformations

Definition 2.1. A contact transformation of the form

$$X = \tilde{X}(x, p), \quad Z = \tilde{Z}(x, p) + z, \quad P = \tilde{P}(x, p) \tag{2.1}$$

is called a *special contact transformation*.

Some of the most important applications of special contact transformations are to Hamiltonian systems.

Theorem 2.1. *A (general) contact transformation U in the $(n+1)$-dimensional xz-space, $\mathbf{R}^{n+1}$, can be extended to a special contact transformation $\overline{U}$ in the $(n+2)$-dimensional $\bar{x}\bar{z}$-space, $\mathbf{R}^{n+2}$, which when restricted to the subspace $\mathbf{R}^{n+1}$ of $\mathbf{R}^{n+2}$ has the same effect as U.*

Proof. Let

$$U\colon X = X(x,z,p),\ Z = Z(x,z,p),\ P = P(x,z,p) \tag{2.2}$$

be a general contact transformation in $\mathbf{R}^{n+1}$. By Theorem 1.1, there is a function $\rho = \rho(x,z,p) \neq 0$ such that

$$P \cdot dX - dZ = \rho(p \cdot dx - dz). \tag{2.3}$$

Let $\bar{x} = (\bar{x}_1, \dots, \bar{x}_n, \bar{x}_{n+1})$ be a point in $\mathbf{R}^{n+1}$, where $\bar{x}_i = x_i$ for $i = 1, \dots, n$ and $\bar{x}_{n+1} = -z$. Then $(\bar{x}, \bar{z}) \in \mathbf{R}^{n+2}$. In the image space, we adjoin an additional coordinate, $\overline{Z}$, so that $(\overline{X}, \overline{Z})$ is in a domain in $\mathbf{R}^{n+2}$, where $\overline{X} = (\overline{X}_1, \dots, \overline{X}_n, \overline{X}_{n+1})$, $\overline{X}_i = X_i$ for $i = 1, \dots, n$ and $\overline{X}_{n+1} = -Z$. Next, let $\bar{p}_{n+1}$, $\overline{P}_{n+1}$ be direction coefficients. We shall choose $\overline{P}_{n+1}$ appropriately below. Equation (2.3) becomes

$$P_i d\overline{X}_i + d\overline{X}_{n+1} = \rho\left(p_i d\bar{x}_i + d\bar{x}_{n+1}\right). \tag{2.4}$$

Let $\bar{p}_1, \dots, \bar{p}_{n+1}$ be direction coefficients, where $\bar{p}_1, \dots, \bar{p}_n$ are related to $p_1, \dots, p_n$ by $\bar{p}_i = p_i \bar{p}_{n+1}$, $i = 1, \dots, n$. Also, define $\overline{P}_1, \dots, \overline{P}_n$ by $\overline{P}_i = P_i \overline{P}_{n+1}$, $i = 1, \dots, n$, and $\overline{P}_{n+1}$ is chosen as follows. From (2.4) we have

$$\frac{\overline{P}_i}{\overline{P}_{n+1}}\, d\overline{X}_i + d\overline{X}_{n+1} = \rho\left(\frac{\bar{p}_i}{\bar{p}_{n+1}}\, d\bar{x}_i + d\bar{x}_{n+1}\right)$$

or

$$\overline{P}_i\, d\overline{X}_i = \frac{\rho \overline{P}_{n+1}}{\bar{p}_{n+1}} \bar{p}_i\, d\bar{x}_i, \quad i = 1, \dots, n+1.$$

The transformation $U\colon (x,z,p) \to (X,Z,P)$ is extended to the transformation $\overline{U}\colon (\bar{x}, \bar{z}, \bar{p}) \to (\overline{X}, \overline{Z}, \overline{P})$ by adjoining to the $2n+1$ equations defining U, the two additional equations $\overline{Z} = \bar{z}$, $\overline{P}_{n+1} = (1/\rho)\bar{p}_{n+1}$. The system of equations

$$\begin{aligned}
\overline{X}_j &= X_j(x, -x_{n+1}, p) \equiv \overline{X}_j(\bar{x}, \bar{p}), \quad j = 1, \dots, n,\\
\overline{X}_{n+1} &= -Z(x, -x_{n+1}, p) \equiv \overline{X}_{n+1}(\bar{x}, \bar{p}),\\
\overline{Z} &= 0 + \bar{z} \quad (\text{i.e. } \tilde{Z}(\bar{x}, \bar{p}) = 0),\\
\overline{P}_j &= (\bar{p}_{n+1}/\rho) P_j(x, -x_{n+1}, p) \equiv \overline{P}_j(\bar{x}, \bar{p}), \qquad j = 1, \dots, n,\\
\overline{P}_{n+1} &= (1/\rho)\bar{p}_{n+1} \equiv \overline{P}_{n+1}(\bar{x}, \bar{p})
\end{aligned} \tag{2.5}$$

is a special contact transformation in $\mathbf{R}^{n+2}$ which satisfies $\overline{P} \cdot d\overline{X} = \bar{p} \cdot d\bar{x}$. Conversely, when restricted to $\mathbf{R}^{n+1}$, (2.5) defines a contact transformation which coincides with (2.2). □

Example 2.1. The extension $\overline{U}$ of the Legendre transformation in Example 1.1 is

$$\begin{aligned}\overline{X}_i &= X_i = p_i = \frac{\bar{p}_i}{\bar{p}_{n+1}}, \quad i = 1, \dots, n,\\ \overline{X}_{n+1} &= -Z = -\left(\frac{\bar{p}_i}{\bar{p}_{n+1}}\bar{x}_i + \bar{x}_{n+1}\right),\\ \overline{Z} &= \bar{z},\\ \overline{P}_i &= P_i\overline{P}_{n+1} = x_i\overline{P}_{n+1} = \bar{x}_i\overline{P}_{n+1} = -\bar{x}_i\bar{p}_{n+1}, \quad i = 1, \dots, n,\\ \overline{P}_{n+1} &= -\bar{p}_{n+1}.\end{aligned}$$

Returning now to transformations in $\mathbf{R}^{n+1}$, we drop the bar notation.

Theorem 2.2. *An element transformation of the form* (2.1)

$$X = X(x,p), \quad Z = Z(x,p) + z, \quad P = P(x,p)$$

is a special contact transformation if and only if the equation

$$P \cdot dX - p \cdot dx = d(Z - z) = dZ \tag{2.6}$$

holds, where dZ is the total differential of a function Z of (x,p).

Condition (2.6) yields

$$\left(P_i\frac{\partial X_i}{\partial x_j} - p_j\right) dx_j + P_i\frac{\partial X_i}{\partial p_j}\, dp_j = \frac{\partial Z}{\partial x_j}\, dx_j + \frac{\partial Z}{\partial p_j}\, dp_j$$

or, comparing coefficients,

$$\begin{aligned}\frac{\partial Z}{\partial x_j} &= P_i\frac{\partial X_i}{\partial x_j} - p_j, \quad j = 1, \dots, n,\\ \frac{\partial Z}{\partial p_j} &= P_i\frac{\partial X_i}{\partial p_j}, \qquad\quad j = 1, \dots, n.\end{aligned} \tag{2.7}$$

These conditions characterize contact transformations of the form $X = X(x,p)$, $P = P(x,p)$ in the $2n$-dimensional xp-space. Such transformations are also referred to as *canonical transformations*.

Using the equivalence of the mixed second partial derivatives for Z and (2.7) one obtains conditions on $(X(x,p), P(x,p))$ that are independent of Z:

$$\begin{aligned}\frac{\partial P_i}{\partial x_k}\frac{\partial X_i}{\partial x_j} - \frac{\partial P_i}{\partial x_j}\frac{\partial X_i}{\partial x_k} &= 0, \quad j,k = 1, \dots, n,\\ \frac{\partial P_i}{\partial p_k}\frac{\partial X_i}{\partial x_j} - \frac{\partial P_i}{\partial x_j}\frac{\partial X_i}{\partial p_k} &= \delta_{jk}, \quad j,k = 1, \dots, n,\\ \frac{\partial P_i}{\partial p_k}\frac{\partial X_i}{\partial p_j} - \frac{\partial P_i}{\partial p_j}\frac{\partial X_i}{\partial p_k} &= 0, \quad j,k = 1, \dots, n,\end{aligned}$$

where δ_{jk} is the Kronecker delta.

3. Characterization of the general contact transformation

Consider the transformation $(x, z, p) \to (\bar{x}, \bar{z}, \bar{p})$, where we use the bar notation:

$$\bar{x}_i = x_i, \quad \bar{x}_{n+1} = -z, \quad \bar{p}_i = \bar{p}_{n+1} p_i, \quad i = 1, \dots, n. \tag{3.1}$$

Let $f = f(x, z, p)$, $g = g(x, z, p)$ be two differentiable functions, and

$$\begin{aligned} f(x, z, p) &= f(x_1, \dots, x_n, z, p_1, \dots, p_n) \\ &= f(\bar{x}_1, \dots, \bar{x}_n, -\bar{x}_{n+1}, \bar{p}_1/\bar{p}_{n+1}, \dots, \bar{p}_n/\bar{p}_{n+1}) \\ &\equiv \bar{f}(\bar{x}, \bar{p}) \end{aligned}$$

and similarly $g(x, z, p) \equiv \bar{g}(\bar{x}, \bar{p})$. The Poisson bracket for the pair of functions $\bar{f}$ and $\bar{g}$ is given by

$$[\bar{f}, \bar{g}]_{\bar{x}\bar{p}} = \frac{\partial \bar{f}}{\partial \bar{x}_j} \frac{\partial \bar{g}}{\partial \bar{p}_j} - \frac{\partial \bar{f}}{\partial \bar{p}_j} \frac{\partial \bar{g}}{\partial \bar{x}_j}. \tag{3.2}$$

We may now rewrite this expression in terms of the original variables

$$\frac{\partial \bar{f}}{\partial \bar{x}_i} = \frac{\partial f}{\partial x_i}, \quad \frac{\partial \bar{f}}{\partial \bar{p}_i} = \frac{1}{\bar{p}_{n+1}} \frac{\partial f}{\partial p_i}, \quad i = 1, \dots, n,$$
$$\frac{\partial \bar{f}}{\partial \bar{x}_{n+1}} = -\frac{\partial f}{\partial z}, \quad \frac{\partial \bar{f}}{\partial \bar{p}_{n+1}} = -\frac{\partial f}{\partial p_i} \frac{\bar{p}_i}{\bar{p}_{n+1}^2} = -\frac{1}{\bar{p}_{n+1}} p_i \frac{\partial f}{\partial p_i}, \quad p_i = \frac{\bar{p}_i}{\bar{p}_{n+1}},$$

and similar formulas for $\bar{g}$ hold. Then formula (3.2) takes the form

$$\begin{aligned} [\bar{f}, \bar{g}]_{\bar{x}\bar{p}} &= \frac{1}{\bar{p}_{n+1}} \left(\frac{\partial f}{\partial x_j} \frac{\partial g}{\partial p_j} - \frac{\partial f}{\partial p_j} \frac{\partial g}{\partial x_j} \right) + \frac{1}{\bar{p}_{n+1}} \left(\frac{\partial f}{\partial z} p_j \frac{\partial g}{\partial p_j} - \frac{\partial g}{\partial z} p_j \frac{\partial f}{\partial p_j} \right) \\ &= \frac{1}{\bar{p}_{n+1}} \left(\left(\frac{\partial f}{\partial x_j} + p_j \frac{\partial f}{\partial z} \right) \frac{\partial g}{\partial p_j} - \left(\frac{\partial g}{\partial x_j} + p_j \frac{\partial g}{\partial z} \right) \frac{\partial f}{\partial p_j} \right). \end{aligned} \tag{3.3}$$

The symbol

$$\{f, g\}_{xzp} = \left(\frac{\partial f}{\partial x_j} + p_j \frac{\partial f}{\partial z} \right) \frac{\partial g}{\partial p_j} - \left(\frac{\partial g}{\partial x_j} + p_j \frac{\partial g}{\partial z} \right) \frac{\partial f}{\partial p_j} \tag{3.4}$$

is called the *Mayer bracket* of f and g. Equation (3.3) in terms of the Mayer bracket takes the form

$$[\bar{f}, \bar{g}]_{\bar{x}\bar{p}} = \frac{1}{\bar{p}_{n+1}} \{f, g\}_{xzp}. \tag{3.5}$$

The Mayer bracket satisfies properties similar to those of the Poisson bracket.

Theorem 3.1. *Let f, g, h be differentiable functions of the variables (x, y, z) and let α be a constant. Then:*

i) $\{f, g\} = -\{g, f\}$, $\{f, f\} = 0$;

ii) $\{\alpha, f\} = 0$, $\{\alpha f, g\} = \alpha \{f, g\}$;

iii) $\{f + g, h\} = \{f, h\} + \{g, h\}$;

iv) $\{fg, h\} = g\{f, h\} + f\{g, h\}$;

v) *The Jacobi identity holds in the form*

$$\{f\{g, h\}\} + \{g\{h, f\}\} + \{h\{f, g\}\} + f_z\{g, h\} + g_z\{h, f\} + h_z\{f, g\} = 0.$$

Note that the subscripts xzp have been dropped.

Equation (3.5) leads to a formula describing how the Mayer bracket changes under a contact transformation.

Theorem 3.2. *The element transformation*

$$X = X(x, z, p), \quad Z = Z(x, z, p), \quad Y = Y(x, z, p)$$

is a contact transformation with multiplier ρ *if and only if up to a factor* $1/\rho$ *it leaves the Mayer bracket of two arbitrary differentiable functions invariant*

$$\{F, G\}_{XZP} = \frac{1}{\rho}\{f, g\}_{xzp}.$$

Proof. Let

$$X = X(x, z, p), \quad Z = Z(x, z, p), \quad P = P(x, z, p)$$

be contact transformation, and let

$$x = x(X, Z, P), \quad z = z(X, Z, P), \quad p = p(X, Z, P)$$

be its inverse. Set

$$F(X, Z, P) = F(X(x, z, p), Z(x, z, p), P(x, z, p)) \equiv f(x, z, p),$$
$$G(X, Z, P) = G(X(x, z, p), Z(x, z, p), P(x, z, p)) \equiv g(x, z, p).$$

Now lift the variables one dimension and set

$$\overline{X}_i = X_i, \quad \overline{X}_{n+1} = -Z, \quad \overline{P}_i = \overline{P}_{n+1}P_i, \quad \overline{P}_{n+1} = \frac{1}{\rho}\bar{p}_{n+1}.$$

We use the fact that canonical transformations preserve the form of the Poisson bracket, and formula (3.5) to get

$$\frac{1}{\bar{p}_{n+1}}\{f, g\}_{xzp} = [\bar{f}, \bar{g}]_{\bar{x}\bar{p}} = [\bar{F}, \bar{G}]_{\overline{XP}} = \frac{1}{\overline{P}_{n+1}}\{F, G\}_{XZP},$$

or since $\overline{P}_{n+1}/\bar{p}_{n+1} = 1/\rho$, $\{F, G\}_{XZP} = (1/\rho)\{f, g\}_{xzp}$. □

Theorem 3.2 suggests that the Mayer bracket plays the same role for general contact transformations as the Poisson bracket plays for the special (canonical) transformations.

Theorem 3.3. *In order for the one-to-one element transformation* $X = X(x, z, p)$, $Z = Z(x, z, p)$, $Y = Y(x, z, p)$, *which satisfies the relationship*

$$P_j\, dX_j - dZ = \rho(p_k\, dx_k - dz) \tag{3.6}$$

with $\rho(x, z, p) \neq 0$, *to be a contact transformation, it is necessary and sufficient that the following relations are satisfied*:

$$\begin{aligned} \{X_i, X_j\}_{xzp} &= 0, & i, j &= 1, \dots, n, \\ \{X_i, P_j\}_{xzp} &= \rho\delta_{ij}, & i, j &= 1, \dots, n, \\ \{X_i, Z\}_{xzp} &= 0, & i &= 1, \dots, n, \\ \{P_i, P_j\}_{xzp} &= 0, & i, j &= 1, \dots, n, \\ \{P_i, Z\}_{xzp} &= -\rho P_i, & i &= 1, \dots, n. \end{aligned} \tag{3.7}$$

Moreover, the following conditions hold:

$$\begin{aligned} \{\rho, X_j\}_{xzp} &= \rho\frac{\partial X_j}{\partial z}, \\ \{\rho, Z\}_{xzp} &= \rho\frac{\partial Z}{\partial z} - \rho^2, \\ \{\rho, P_j\}_{xzp} &= \rho\frac{\partial P_j}{\partial z}. \end{aligned} \tag{3.8}$$

Proof. Notice that $\{X_i, X_j\}_{xzp} = \rho\{X_i, X_j\}_{XZP} = 0$, $i, j = 1, \dots, n$. The rest of the equations (3.7) are obtained similarly. The derivations of equations (3.8) are lengthy and can be found in [14]. □

Corollary 3.1. *The functions* (X, P) *of a contact transformation are independent of* z *if and only if* ρ *is a constant.*

Proof. We observe that

$$\begin{aligned} \frac{\partial\rho}{\partial X_j} &= \frac{\partial P_j}{\partial z} + \frac{1}{\rho}\frac{\partial\rho}{\partial z}P_j, & j &= 1, \dots, n, \\ \frac{\partial\rho}{\partial P_j} &= -\frac{\partial X_j}{\partial z}, & j &= 1, \dots, n, \\ \frac{\partial\rho}{\partial Z} &= -\frac{1}{\rho}\frac{\partial\rho}{\partial z}. \end{aligned}$$

We will show how to obtain the second of these equations

$$\{\rho, X_j\}_{XZP} = \left(\frac{\partial\rho}{\partial X_i} + P_i\frac{\partial\rho}{\partial Z}\right)\frac{\partial X_j}{\partial P_i} - \left(\frac{\partial X_j}{\partial X_i} + P_i\frac{\partial X_j}{\partial Z}\right)\frac{\partial\rho}{\partial P_i} = -\frac{\partial\rho}{\partial P_j},$$

since $\partial X_j/\partial P_i = \partial X_j/\partial Z = 0$ and $\partial X_i/\partial X_j = \delta_{ij}$. Also, by one of the identities (3.8), $\{\rho, X_j\}_{XZP} = (1/\rho)\{\rho, X_j\}_{xzp} = \partial X_j/\partial z$. Similar calculations produce the other two equations. □

Theorem 3.4. *Let $X_1, \dots, X_n, Z$ be $n+1$ independent functions which are pairwise in involution with respect to the Mayer bracket. Then there is precisely one contact transformation for which these are the first $n+1$ functions and the remaining $n+1$ functions $P_1, \dots, P_n, \rho$ may be obtained by solving a linear system of equations.*

Proof. If the $n+1$ independent functions $Z, X_1, \dots, X_n$ of (x, z, p) are pairwise in involution, that is if they satisfy $\{Z, X_i\}_{xzp} = 0$, $\{X_i, X_j\}_{xzp} = 0$, then the functions $P_1, \dots, P_n, \rho$ can be calculated as follows. By equating coefficients in the defining identity (3.6) for a contact transformation, we obtain the system

$$\begin{aligned} \frac{P_i}{\rho}\frac{\partial X_i}{\partial x_j} - \frac{1}{\rho}\frac{\partial Z}{\partial x_j} &= p_j, \qquad j = 1, \dots, n, \\ \frac{P_i}{\rho}\frac{\partial X_i}{\partial z} - \frac{1}{\rho}\frac{\partial Z}{\partial z} &= -1, \\ \frac{P_i}{\rho}\frac{\partial X_i}{\partial p_j} - \frac{1}{\rho}\frac{\partial Z}{\partial p_j} &= 0, \qquad j = 1, \dots, n. \end{aligned} \tag{3.9}$$

Since $X_1, \dots, X_n, Z$ are functionally independent, the rank of the matrix

$$\begin{pmatrix} X_x & Z_x \\ X_z & Z_z \end{pmatrix} = \begin{pmatrix} \frac{\partial X_i}{\partial x_j} & \frac{\partial Z}{\partial x_j} \\ \frac{\partial X_i}{\partial z} & \frac{\partial Z}{\partial z} \end{pmatrix} \tag{3.10}$$

is $n+1$, so the first $n+1$ equations in the above system can be solved for P_i/ρ, $i = 1, \dots, n$, and $1/\rho$. We now must show that the last n equations in system (3.9) are satisfied identically. For that consider the expression

$$\begin{aligned} \frac{P_j}{\rho}\{X_i, X_j\} - \frac{1}{\rho}\{X_i, Z\} = \left(\frac{\partial X_i}{\partial x_k} + p_k\frac{\partial X_i}{\partial z}\right)\left(\frac{P_j}{\rho}\frac{\partial X_j}{\partial p_k} - \frac{1}{\rho}\frac{\partial Z}{\partial p_k}\right) \\ - \left(\frac{\partial X_j}{\partial x_k}\frac{P_j}{\rho} - \frac{1}{\rho}\frac{\partial Z}{\partial x_k}\right)\frac{\partial X_i}{\partial p_k} - p_k\left(\frac{\partial X_j}{\partial z}\frac{P_j}{\rho} - \frac{\partial Z}{\partial z}\frac{1}{\rho}\right)\frac{\partial X_i}{\partial p_k}. \end{aligned}$$

Taking in consideration the validity of the first two equations in system (3.9) and that $\{X_i, X_j\} = 0$ and $\{X_i, Z\} = 0$, we obtain

$$\left(\frac{\partial X_i}{\partial x_k} + p_k\frac{\partial X_i}{\partial z}\right)\left(\frac{P_j}{\rho}\frac{\partial X_j}{\partial p_k} - \frac{1}{\rho}\frac{\partial Z}{\partial p_k}\right) = 0. \tag{3.11}$$

Since the columns of the matrix (3.10) are linearly independent, identity (3.11) implies that the last n equations in system (3.9) are identically satisfied. □

4. One-parameter families of contact transformations

We now consider the special system of $2n+1$ differential equations for $2n+1$ unknowns $X = (X_1, \dots, X_n)$, Z, $X = (P_1, \dots, P_n)$

$$\dot{X} = \xi(X, Z, P, t), \quad \dot{Z} = \zeta(X, Z, P, t), \quad \dot{P} = \pi(X, Z, P, t), \tag{4.1}$$

which satisfy the initial conditions

$$X = x, \quad Z = z, \quad P = p, \quad \text{when} \quad t = 0. \tag{4.2}$$

The functions $\xi = (\xi_1, \dots, \xi_n)$, ζ, $\pi = (\pi_1, \dots, \pi_n)$ are all assumed to be continuously differentiable. The solutions to (4.1), (4.2)

$$X = X(x, z, p, t), \quad Z = Z(x, z, p, t), \quad P = P(x, z, p, t) \tag{4.3}$$

determine a family of transformations

$$S_t \colon (x, z, p) \to (X, Z, P). \tag{4.4}$$

In this section we give the necessary and sufficient conditions for the transformations (4.4) to be contact transformations uniformly in t.

Theorem 4.1. *In order for solution (4.3) of system (4.1) to represent a one-parameter family of contact transformations containing the identity, it is necessary that (4.1) be a canonical system, that is, that there exists a function,* $\mathcal{H} = \mathcal{H}(X, Z, P, t)$ *called the characteristic function, such that the system (4.1) has the form*

$$\begin{aligned} \frac{d}{dt} X_j &= \frac{\partial \mathcal{H}}{\partial P_j}, \\ \frac{d}{dt} Z &= P_j \frac{\partial \mathcal{H}}{\partial P_j} - \mathcal{H}, \\ \frac{d}{dt} P_j &= -\frac{\partial \mathcal{H}}{\partial X_j} - P_j \frac{\partial \mathcal{H}}{\partial Z}, \quad j = 1, \dots, n. \end{aligned} \tag{4.5}$$

Proof. In Section 3 we had found that the transformations must satisfy

$$P \cdot dX - dZ = \rho(p \cdot dx - dz), \quad \rho \neq 0. \tag{4.6}$$

(4.6) is supposed to hold when the differentials are calculated only with respect to the spatial variables. When X, Z, P also depend on t, then dZ is given by

$$dZ = \frac{\partial Z}{\partial x_j} dx_j + \frac{\partial Z}{\partial z} dz + \frac{\partial Z}{\partial p_j} dp_j + \frac{\partial Z}{\partial t} dt.$$

A similar assertion holds for the dX_i. Thus, condition (4.6) must be replaced by

$$P_i \, dX_i - dZ - \left(P_i \frac{\partial X_i}{\partial t} - \frac{\partial Z}{\partial t} \right) dt = \rho(p_i \, dx_i - dz). \tag{4.7}$$

By (4.1), $\partial X_i / \partial t = \xi_i(X, Z, P, t)$, $\partial Z / \partial t = \zeta(X, Z, P, t)$. Let us introduce the function

$$\mathcal{H} \equiv \mathcal{H}(X, Z, P, t) \equiv P_i \xi_i(X, Z, P, t) - \zeta(X, Z, P, t). \tag{4.8}$$

Then relation (4.7) takes the form

$$P \cdot dX - dZ = \rho(p \cdot dx - dz) + \mathcal{H} \, dt. \tag{4.9}$$

If $dt = 0$, equation (4.9) reduces to (4.6). (4.9) represents a system of $2n + 2$ equations relating the variables (X, Z, P, t) with those of (x, z, p, t), which is obtained by expanding the differentials and comparing coefficients. To obtain the conditions we seek, we shall rewrite these conditions in the (X, Z, P, t) variables. This is most simply done by working directly with (4.9). First differentiate (4.9) with respect to t and note that the differential operator, d, commutes with the differentiation d/dt. This leads to

$$\pi_j\, dX_j + P_j\, d\xi_j - d\zeta = \dot{\rho}(p_j\, dx_j - dz) + \dot{\mathcal{H}}\, dt, \tag{4.10}$$

where $\partial P_j/\partial t = \pi_j(X, Z, P, t)$, the dot, as usual, represents d/dt. From (4.9) and (4.10) we obtain

$$\pi_j\, dX_j + P_j\, d\xi_j - d\zeta - \dot{\mathcal{H}}\, dt = \frac{\dot{\rho}}{\rho}\left(P_j\, dX_j - dZ - \mathcal{H}\, dt\right). \tag{4.11}$$

From (4.8) we find $d\mathcal{H} = \xi_j\, dP_j + P_j\, d\xi_j - d\zeta$ so that (4.11) takes the form

$$d\mathcal{H} + \pi_j\, dX_j - \xi_j\, dP_j = \frac{\dot{\rho}}{\rho}\left(P_j\, dX_j - dZ\right) + \left(\dot{\mathcal{H}} - \frac{\dot{\rho}}{\rho}\mathcal{H}\right) dt. \tag{4.12}$$

Expand $d\mathcal{H}$ in the form

$$d\mathcal{H} = \frac{\partial \mathcal{H}}{\partial X_j}\, dX_j + \frac{\partial \mathcal{H}}{\partial Z}\, dZ + \frac{\partial \mathcal{H}}{\partial P_j}\, dP_j + \frac{\partial \mathcal{H}}{\partial t}\, dt,$$

insert the result into (4.12) and compare coefficients to obtain the following system

$$\begin{aligned} \frac{\partial \mathcal{H}}{\partial X_j} &= -\pi_j + \frac{\dot{\rho}}{\rho} P_j, & \frac{\partial \mathcal{H}}{\partial P_j} &= \xi_j, \\ \frac{\partial \mathcal{H}}{\partial Z} &= -\frac{\dot{\rho}}{\rho}, & \frac{\partial \mathcal{H}}{\partial t} &= \dot{\mathcal{H}} - \frac{\dot{\rho}}{\rho}\mathcal{H}. \end{aligned} \tag{4.13}$$

The ξ_j and π_j are obtained directly from (4.13) by eliminating the quotient $\dot{\rho}/\rho$ and solving. To obtain ζ combine (4.8) with (4.13). We find

$$\begin{aligned} \xi_j &= \frac{\partial \mathcal{H}}{\partial P_j}, \\ \zeta &= P_j \xi_j - \mathcal{H} = P_j \frac{\partial \mathcal{H}}{\partial P_j} - \mathcal{H}, \\ \pi_j &= -\frac{\partial \mathcal{H}}{\partial X_j} - P_j \frac{\partial \mathcal{H}}{\partial Z}, \qquad\qquad j = 1, \ldots, n, \end{aligned} \tag{4.14}$$

which is system (4.5). □

The converse of this theorem is also valid. We state and prove

Theorem 4.2. *The solution to the canonical equations (4.5), which satisfy the initial conditions (4.2), generates a one-parameter family of contact transformations, which for $t = 0$ contains the identity.*

Proof. We must show that every solution of (4.5) and (4.2) satisfies the strip condition (4.9). For notational purposes let us set $\Omega = \Omega(t) \equiv P_j\, dX_j - dZ - \mathcal{H}\, dt$ and $\Omega(0) \equiv \omega = p_j\, dx_j - dz$. Then the strip condition (4.9) takes the form $\Omega(t) = \rho\omega$. Set up a differential equation for Ω making use of (4.5). The proof is simply a calculation. We find $\dot{\Omega} = \dot{P}_j\, dX_j + P_j\, d\dot{X}_j - d\dot{Z} - \dot{\mathcal{H}}\, dt$. Since $\mathcal{H} = P_j\, d\dot{X}_j - d\dot{Z}$,

$$\begin{aligned}
\dot{\Omega} &= \dot{P}_j\, dX_j - \dot{X}_j P_j \\
&= -\left(\frac{\partial \mathcal{H}}{\partial X_j} + P_j \frac{\partial \mathcal{H}}{\partial Z}\right) dX_j - \frac{\partial \mathcal{H}}{\partial P_j}\, dP_j \\
&= -\left(\frac{\partial \mathcal{H}}{\partial X_j}\, dX_j + \frac{\partial \mathcal{H}}{\partial P_j}\, dP_j\right) - \frac{\partial \mathcal{H}}{\partial Z}\,(P_j\, dX_j) \\
&= -\left(\frac{\partial \mathcal{H}}{\partial X_j}\, dX_j + \frac{\partial \mathcal{H}}{\partial P_j}\, dP_j\right) + \frac{\partial \mathcal{H}}{\partial Z}\, dZ + \frac{\partial \mathcal{H}}{\partial t}\, dt \\
&\quad - \frac{\partial \mathcal{H}}{\partial Z}\,(P_j\, dX_j) + \frac{\partial \mathcal{H}}{\partial Z}\, dZ + \frac{\partial \mathcal{H}}{\partial t}\, dt \\
&= -\, d\mathcal{H} - \frac{\partial \mathcal{H}}{\partial Z}\,(P_j\, dX_j - dZ - \mathcal{H}\, dt) - \frac{\partial \mathcal{H}}{\partial Z}\mathcal{H}\, dt + \frac{\partial \mathcal{H}}{\partial t}\, dt \\
&= -\, d\mathcal{H} - \frac{\partial \mathcal{H}}{\partial Z}\Omega - \frac{\partial \mathcal{H}}{\partial Z}\mathcal{H}\, dt + \frac{\partial \mathcal{H}}{\partial t}\, dt.
\end{aligned}$$

Thus we obtain the ODE for Ω

$$\dot{\Omega} = -d\mathcal{H} - \frac{\partial \mathcal{H}}{\partial Z}\Omega - \frac{\partial \mathcal{H}}{\partial Z}\mathcal{H}\, dt + \frac{\partial \mathcal{H}}{\partial t}\, dt.$$

Next we calculate, using (4.5)

$$\begin{aligned}
\frac{d\mathcal{H}}{dt} &= \frac{\partial \mathcal{H}}{\partial X_j}\frac{dX_j}{dt} + \frac{\partial \mathcal{H}}{\partial P_j}\frac{dP_j}{dt} + \frac{\partial \mathcal{H}}{\partial Z}\frac{dZ}{dt} + \frac{\partial \mathcal{H}}{\partial t} \\
&= \frac{\partial \mathcal{H}}{\partial X_j}\frac{\partial \mathcal{H}}{\partial P_j} - \frac{\partial \mathcal{H}}{\partial P_j}\left(\frac{\partial \mathcal{H}}{\partial X_j} + P_j\frac{\partial \mathcal{H}}{\partial Z}\right) + \frac{\partial \mathcal{H}}{\partial Z}\left(P_j\frac{\partial \mathcal{H}}{\partial P_j} - \mathcal{H}\right) + \frac{\partial \mathcal{H}}{\partial t} \\
&= -\mathcal{H}\frac{\partial \mathcal{H}}{\partial Z} + \frac{\partial \mathcal{H}}{\partial t}.
\end{aligned}$$

Thus,

$$d\mathcal{H} = -\mathcal{H}\frac{\partial \mathcal{H}}{\partial Z}\, dt + \frac{\partial \mathcal{H}}{\partial t}\, dt$$

and so from the previous calculation

$$\dot{\Omega} = -\frac{\partial \mathcal{H}}{\partial Z}\Omega.$$

We integrate to obtain $\Omega = \rho\omega$, where

$$\rho = \exp\left(-\int_0^t \frac{\partial \mathcal{H}}{\partial Z}\, dt\right), \tag{4.15}$$

which proves the assertion. □

We close this section with a few remarks on the characteristic function $\mathcal{H} = \mathcal{H}(X, Z, P, t)$. From the forth equation in (4.13), we have

$$\rho \frac{\partial \mathcal{H}}{\partial t} = \rho \dot{\mathcal{H}} - \dot{\rho} \mathcal{H}.$$

Divide by ρ^2 to find

$$\frac{1}{\rho} \frac{\partial \mathcal{H}}{\partial t} = \frac{\rho \dot{\mathcal{H}} - \dot{\rho} \mathcal{H}}{\rho^2} = \frac{d}{dt} \left(\frac{\mathcal{H}}{\rho} \right).$$

Integrate with respect to t to find

$$\frac{\mathcal{H}}{\rho} - \frac{\mathcal{H}^0}{\rho^0} = \int_0^t \frac{1}{\rho} \frac{\partial \mathcal{H}}{\partial t}\, dt, \tag{4.16}$$

where the superscript indicates that the arguments of $\mathcal{H}$ and ρ are to be taken at $t = 0$: $\rho^0 = \rho(x, z, p, 0)$, $\mathcal{H}^0 = \mathcal{H}(x, z, p, 0)$. The fact that $\rho^0 = 1$ is a consequence of (4.15).

We consider two special cases.

Case 1. $\partial \mathcal{H}/\partial t = 0$ so that $\mathcal{H}$ does not depend explicitly on t.

Then the family $\{S_t\}$ represents a one-parameter group of contact transformations. (The proof can be found in [14].) Relation (4.16) implies that

$$\mathcal{H}(X, Z, P) = \mathcal{H}^0(x, z, p)\rho(x, z, p). \tag{4.17}$$

(4.17) has a geometric interpretation. Let us think of the parameter t as the time and the curve along which $(X, Z, P) = S_t(x, z, p)$ moves in $\mathbf{R}^{2n+1}$ as its orbit under the group of contact transformations. Along this orbit the function $\mathcal{H}(X, Z, P)$, up to the factor $\mathcal{H}^0$, coincides with $\rho(X, Z, P)$.

If in particular $\mathcal{H}^0 = 0$ at a point (x, z, p), then $\mathcal{H}(X, Z, P) = 0$ along the whole orbit through it. The strip condition is along the orbit. If we think of (X, Z, P) as an element in $\mathbf{R}^{n+1}$, then we refer to the orbit as an orbital strip of the group of contact transformations in $\mathbf{R}^{n+1}$. For points on the orbital strip, the second equation in (4.5) simplifies to

$$\frac{dZ}{dt} = P_j \frac{\partial \mathcal{H}}{\partial P_j}, \quad j = 1, \ldots, n.$$

Case 2. $\partial \mathcal{H}/\partial Z = 0$ so that $\mathcal{H}$ does not depend explicitly on Z and by (4.15) $\rho = \rho(X, Z, P, t) \equiv 1$.

The canonical equations (4.5) reduce to

$$\frac{dX_j}{dt} = \frac{\partial \mathcal{H}}{\partial P_j}, \quad \frac{dP_j}{dt} = -\frac{\partial \mathcal{H}}{\partial X_j} \tag{4.18}$$

together with the additional equation

$$\frac{dZ}{dt} = P_j \frac{\partial \mathcal{H}}{\partial P_j} - \mathcal{H}, \quad j = 1, \ldots, n \tag{4.19}$$

for the construction of Z.

The transformations determined by (4.18) are the special, or xp-transformations which commute with translations along the z-axis. Equation (4.9) in this case reads

$$P_j\,dX_j - p_j\,dx_j = d(Z - z) + \mathcal{H}\,dt.$$

If in addition, $\partial\mathcal{H}/\partial t = 0$, then $\mathcal{H} = \mathcal{H}^0$. The family determined by solutions to (4.18) is a group of contact transformations which on the orbit passing through (x, z, p) satisfies $\mathcal{H}(X, Z, P) = \mathcal{H}^0(x, z, p)$.

5. Transformations of canonical differential equations

In this section we show that the form of the generalized Hamiltonian system is preserved by contact transformations.

Consider the generalized Hamiltonian system

$$\begin{aligned} \dot{x}_j &= \frac{\partial\mathcal{H}}{\partial p_j}, \\ \dot{z} &= p_j\frac{\partial\mathcal{H}}{\partial p_j} - \mathcal{H}, \\ \dot{p}_j &= -\frac{\partial\mathcal{H}}{\partial x_j} - p_j\frac{\partial\mathcal{H}}{\partial z}, \quad j = 1, \dots, n, \end{aligned} \tag{5.1}$$

where $H = H(x, z, p, t)$, and with initial values

$$x_j(0) = x_j^0, \quad z(0) = z^0, \quad p_j(0) = p_j^0, \quad j = 1, \dots, n. \tag{5.2}$$

Theorem 5.1. *If system (5.1) with initial values (5.2) is transformed with the contact transformation*

$$T_t\colon \quad \begin{aligned} X_j &= X_j(x, z, p, t), \\ Z &= Z(x, z, p, t), \\ P_j &= P_j(x, z, p, t), \quad j = 1, \dots, n, \end{aligned} \tag{5.3}$$

then the transformed system is a generalized Hamiltonian system with characteristic function $\sigma H + K$, where $K(X, Z, P, t)$ and $\sigma(X, Z, P, t)$ are the characteristic function and the multiplier of the contact transformation T_t, i.e., $P_j\,dX_j - dZ - K\,dt = \sigma(p_j\,dx_j - dz)$.

Proof. Let S_t denote the contact transformation defined by the solution of the generalized Hamiltonian system (5.1)–(5.2), i.e., let $(x, z, p) = S_t(x^0, z^0, p^0)$. Now carry out the substitution indicated by (5.3). The initial values transform as follows

$$(X^0, Z^0, P^0) = T_0(x^0, z^0, p^0) = \big(X(x^0, z^0, p^0, 0), Z(x^0, z^0, p^0, 0), P(x^0, z^0, p^0, 0)\big)$$

and the solutions to (5.1)– (5.2) transform to functions of (X^0, Z^0, P^0, t) according to

$$(X, Z, P) = T_t S_t T_0^{-1}(X^0, Z^0, P^0). \tag{5.4}$$

Let $S_t^* \equiv T_t S_t T_0^{-1}$. $\{S_t^*\}$ is a one parameter family of contact transformations, so there exists a canonical system for it which is determined by a characteristic function $\mathcal{H}^* = \mathcal{H}^*(X, Z, P, t)$. We must determine $\mathcal{H}^*$ in terms of known quantities.

Since T_0 is a contact transformation, we find from (5.3)

$$P_j^0\, dX_j^0 - dZ^0 = \sigma^0 \big(p_j^0\, dx_j^0 - dz^0\big), \tag{5.5}$$

where $\sigma^0 = \sigma(X^0, Z^0, P^0, 0)$. Further,

$$p_j\, dx_j - dz = \rho\big(p_j^0\, dx_j^0 - dz^0\big) + \mathcal{H}\, dt. \tag{5.6}$$

Now, using (5.3), (5.6), and (5.5) we find

$$\begin{aligned} P_j\, dX_j - dZ &= \sigma\big(p_j\, dx_j - dz\big) + K\, dt \\ &= \sigma\left(\rho(p_j^0\, dx_j^0 - dz^0) + \mathcal{H}\, dt\right) + K\, dt \\ &= \sigma\left(\frac{\rho}{\sigma^0}(P_j^0\, dX_j^0 - dZ^0) + \mathcal{H}\, dt\right) + K\, dt \\ &= \frac{\sigma\rho}{\sigma^0}(P_j^0\, dX_j^0 - dZ^0) + (\sigma\mathcal{H} + K)\, dt. \end{aligned}$$

The coefficient of dt represents the desired characteristic function $\mathcal{H}^* = \mathcal{H}^*(X, Z, P, t) = (\sigma\mathcal{H} + K)$. Observe that σ and K are already evaluated at (X, Z, P, t). The function H, initially evaluated at (x, z, p, t) must simply be rewritten in terms of the new variables $(X, Z, P) = T_t^{-1}(x, z, p)$. Having determined the characteristic function $\mathcal{H}^*$ we can rewrite the system (5.1) immediately in terms of the new variables. □

We close this section with a final remark. Suppose $\mathcal{H}$ is independent of z so that the canonical equations are

$$\dot{x}_j = \frac{\partial \mathcal{H}}{\partial p_j}, \quad \dot{p}_j = -\frac{\partial \mathcal{H}}{\partial x_j}. \tag{5.7}$$

Now make the substitution

$$X = X(x, p), \quad P = P(x, p) \quad \text{with} \quad P \cdot dX = p \cdot dx. \tag{5.8}$$

This is a special contact transformation which is independent of the parameter t. Then $\sigma = 1$, $K = 0$ and $\mathcal{H}^*$ is obtained by evaluating $\mathcal{H}$ at $x = x(X, P)$, $p = p(X, P)$ and the canonical equations in the (X, P) variables are

$$\dot{X}_j = \frac{\partial \mathcal{H}}{\partial P_j}, \quad \dot{P}_j = -\frac{\partial \mathcal{H}}{\partial X_j}. \tag{5.9}$$

Since (5.7) transforms in (5.9) with $\mathcal{H}^*$ arising from $\mathcal{H}$ by means of (5.8), the special contact transformation is also called a canonical transformation.

6. Liouville-type integrability theorem

This section reviews two remarkable theorems. The first gives a necessary and sufficient condition for the integrability of the generalized Hamiltonian system in terms of a complete set of first integrals. The second provides a method of generating such first integrals.

Consider the generalized Hamiltonian system (5.1). The following theorem extends the classical theorem of Liouville which gives a necessary and sufficient condition for the integrability of the classical Hamiltonian system in terms of a complete set of first integrals which are in involution with respect to the Poisson bracket.

Theorem 6.1. *Suppose* $X_1(x,z,p,t),\dots,X_n(x,z,p,t)$, $Z(x,z,p,t)$ *are* $n+1$ *independent first integrals for* (5.1) *which are pairwise in involution with respect to the Mayer bracket:*

$$\{X_i, X_j\}_{xzp} = 0, \qquad i,j = 1,\dots,n,$$
$$\{X_i, Z\}_{xzp} = 0, \qquad i = 1,\dots,n.$$

Then the general solution to the system (5.1) *can be constructed by means of a quadrature.*

Proof. Construct the functions $P_1,\dots,P_n$ so that (X,Z,P) is a contact transformation (follow the procedure in the proof of Theorem 3.4). Let $H^*(X,Z,P)$ be the characteristic function of this contact transformation. Along a solution of the system (5.1), $X_i = c_i$, $Z = \gamma$, where c_i and γ are constants, so that $\dot{X}_i = \dot{Z} = 0$. From the proof of Theorem 5.1 we know that

$$\frac{\partial H^*}{\partial P_j} = \dot{X}_i = 0, \quad j = 1,\dots,n,$$

hence $H^* = H^*(c,\gamma,t)$. In the new variables

$$\dot{P}_j = -\frac{\partial H^*}{\partial c_j} - P_j\frac{\partial H^*}{\partial \gamma}, \quad j = 1,\dots,n,$$

which is immediately solvable. The complete solution is given by

$$X_i(x,z,p,t) = c_i, \quad i = 1\dots,n,$$
$$Z(x,z,p,t) = \gamma,$$
$$P_i(x,z,p,t) = -\big(\textstyle\int \exp\big(\int \frac{\partial H^*(c,\gamma,t)}{\partial\gamma}\,dt\big)\frac{\partial H^*}{\partial c_j}(c,\gamma,t)\,dt\big)/\exp\big(\int\frac{\partial H^*}{\partial\gamma}\,dt\big),\ i = 1,\dots,n.$$

We now solve this system for $x_1,\dots,x_n,z,p_1,\dots,p_n$. □

Theorem 6.2. $F(x,z,p,t) = \mathrm{const}$ *is a first integral for the generalized Hamiltonian system* (5.1) *if and only if it satisfies* $F_t + \{F,H\} - F_z H = 0$, *where* $\{F,H\}$ *is the Mayer bracket of* F *and the characteristic function* H *of system* (5.1). *The subscripts denote partial differentiation.*

Proof. We observe that if $F(x,z,p,t)$ is a first integral for system (6.1), then

$$\begin{aligned} 0 &= F_t + \frac{\partial F}{\partial x_j}\dot{x}_j + \frac{\partial F}{\partial z}\dot{z} + \frac{\partial F}{\partial p_j}\dot{p}_j \\ &= F_t + F_{x_j}H_{p_j} + F_z(p_jH_{p_j} - H) + F_{p_j}(-H_{x_j} - p_jH_z) \\ &= F_t + (F_{x_j} + p_jF_z)H_{p_j} - (H_{x_j} + p_jH_z)F_{p_j} - F_zH. \end{aligned}$$

We therefore obtain the equality $F_t + \{F,H\} - F_zH = 0$, which is a first order partial differential equation for F having (5.1) as its system of characteristic equations. □

The following theorem gives a method for generating first integrals for the generalized Hamiltonian system.

Theorem 6.3. *If $F(x,z,p,t) = \alpha$, $G(x,z,p,t) = \beta$, with α and β constants, are first integrals for the system (5.1), then $\rho\{F,G\}$ is also a first integral for the same system.*

Proof. Let $F(x,z,p,t) = \alpha$, $G(x,z,p,t) = \beta$, α and β constants, be two first integrals for system (5.1). The Jacobi identity for the Mayer bracket is

$$\{F,\{G,H\}\} + \{G,\{H,F\}\} + \{H,\{F,G\}\} + F_z\{G,H\} + G_z\{H,F\} + H_z\{F,G\} = 0.$$

Replace $\{F,H\}$ and $\{G,H\}$ using the identity provided by Theorem 6.2 and rearrange to obtain the identity

$$-\frac{\partial}{\partial t}\{F,G\} - \{\{F,G\},H\} + H\frac{\partial}{\partial z}\{F,G\} + \frac{\partial H}{\partial z}\{F,G\} = 0.$$

We can rewrite this identity as $d(\rho\{F,G\})/dt = 0$, where $\rho = \exp\big(-\int_0^t \frac{\partial H}{\partial z}\,d\tau\big)$ and conclude that along a solution, $\rho\{F,G\}$ is a constant. □

7. The connection with the variational principle of Herglotz

Let us denote by $\mathcal{L} \equiv \mathcal{L}(x,\dot{x},z,t) \equiv \mathcal{L}(x_1,\dots,x_n,\dot{x}_1,\dots,\dot{x}_n,z,t)$ the Lagrange function, or Lagrangian, of the variables $(x,\dot{x},z,t)$, $z = z(t)$ is a scalar valued function of t. The variable z is to be determined as the solution to the differential equation

$$\dot{z} = \mathcal{L}(x,\dot{x},z,t). \tag{7.1}$$

Observe that (7.1) represents a family of differential equations, since for each $x(t)$ a different differential equation arises, that is, given $x(t)$, $z(t)$ is determined by (7.1) so that $z(t)$ depends on $x(t)$. A fact which we make explicit by writing $z = z[x;t] = z(x,\dot{x},t)$. Problem (7.1) is a kind of control problem. The differential equation for z describes a process which depends on $(x,\dot{x})$ and which in turn can be chosen, that is they give us the opportunity to control or guide the process and are therefore referred to as controls.

Theorem 7.1. *The functions (x, z) for which the functional z has stationary values satisfies the following system of ordinary differential equations*

$$\begin{aligned} \dot{p}_j &= \mathcal{L}_j + \mathcal{L}_z p_j, \quad j = 1, \dots, n, \\ \dot{z} &= \mathcal{L} \quad \textit{with} \quad \mathcal{L}_j \equiv \frac{\partial \mathcal{L}}{\partial x_j}, \quad p_j \equiv \frac{\partial \mathcal{L}}{\partial \dot{x}_j}. \end{aligned} \tag{7.2}$$

The proof of this theorem can be found in [14].

Herglotz named equations (7.2) *generalized Euler-Lagrange equations.*

Theorem 7.2. *Let $\mathcal{L} = \mathcal{L}(x, \dot{x}, z, t)$ and suppose $\det(\partial^2\mathcal{L}/\partial\dot{x}_i\partial\dot{x}_j) \neq 0$. Then the solutions to (7.2) determine a family of contact transformations. If $\mathcal{L}$ is independent of t, the family is a one-parameter group.*

The proof can be found in [9].

Observe that system (7.2) is a generalized Hamiltonian system with

$$H(x, p, z, t) \equiv p_j\dot{x}_j - L(x, \dot{x}, z, t), \quad p_j \equiv \frac{\partial L}{\partial \dot{x}_j}.$$

We can summarize these considerations in the following general statement.

The following four kinds of problems are equivalent:

- *Variational problems for the functional z defined by the differential equation (7.1).*
- *Euler-Lagrange equations for the stationary values of the functional z defined by (7.1).*
- *The generalized Hamiltonian system.*
- *One parameter families of contact transformations.*

Example 7.1. Consider the Lagrangian function $L = m\dot{x}^2/2 - lx^2/2 - \alpha z$, where m, l, α are positive constants. Then $L_{\dot{x}} = m\dot{x} = p$, $L_x = -lx$, $L_z = -\alpha$. The Hamiltonian or characteristic function H is

$$H = H(x, p, z) = \frac{p^2}{2m} + \frac{lx^2}{2} + \alpha z.$$

The canonical system is

$$\begin{aligned} \dot{x} &= \frac{p}{m}, \\ \dot{z} &= \frac{p^2}{2m} - \frac{lx^2}{2} - \alpha z, \\ \dot{p} &= -(lx + \alpha p) \end{aligned}$$

and the Lagrange equation is $m\ddot{x} = -lx - \alpha m\dot{x}$ or

$$\ddot{x} + \alpha\dot{x} + \omega^2 x = 0, \quad \text{where} \quad \omega^2 = \frac{l}{m},$$

which is the equation of the damped harmonic oscillator.

Conclusion. Remarks for further research

Every reader of this paper will find suitable directions for his/her research, nevertheless, I like to mention a few. It will be valuable to see how these results extend to the variational principle of Herglots with several independent variables, which is so useful for the variational description of physical fields.

Another direction is to find methods for solving the generalized Euler-Lagrange equations obtained from Herglotz variational principle, perhaps using the results in this paper. In addition, to find further properties of the generalized Euler-Lagrange equations.

Is it possible to extend, in an appropriate sense, the variational principle of Herglotz to evolution equations? If so, then to investigate their relationship to the Hamiltonian evolution equations.

A more in-depth treatment of the theory of contact transformations, the generalized Hamiltonian system and the variational principle of Herglotz can be found in [14].

In the last 12 years or so about 200 new applications of the variational principle of Herglotz and the variational principle which generalizes it to one with several independent variables were published. They are in theoretical and applied physics, quantum mechanics, field theory, chemistry, mathematics, cosmology, dynamical systems, and many more branches of the exact sciences. I like to mention [1] and [25].

References

[1] R. Almeida and A. Malinowska, Fractional variational principle of Herglotz, Discrete and Continuous Dynamical Systems B, 19(8) (2014) 2367–2381.

[2] C. Caratheodory, Calculus of variations and partial differential equations of the first order, ed. 2, Chelsea Publishing, New York, 1989.

[3] V. Donchev, Variational symmetries, conserved quantities and identities for several equations of mathematical physics, J. Math. Phys. 55(3) (2014) 032901, 11 pp.

[4] K. Furta, A. Sano, and D. Atherton, State variable methods in automatic control, John Wiley, New York, 1988.

[5] B. Georgieva and R. Guenther, First Noether-type theorem for the generalized variational principle of Herglotz, Topol. Methods Nonlinear Anal. 20 (2002) 261–273.

[6] B. Georgieva, R. Guenther, and Th. Bodurov, Generalized variational principle of Herglotz for several independent variables. First Noether-type theorem, J. Math. Phys. 44(9) (2003) 3911–3927.

[7] B. Georgieva and R. Guenther, Second Noether-type theorem for the generalized variational principle of Herglotz, Topol. Methods Nonlinear Anal. 26 (2005) 307–314.

[8] B. Georgieva, Symmetries of the Herglotz variational principle in the case of one independent variable, Annuaire Univ. Sofia Fac. Math. Inf. 100 (2010) 113–122.

[9] B. Georgieva, The variational principle of Herglotz and related results, Geom. Integr. Quantization 12 (2011) 214–225.

[10] B. Georgieva, Symmetries of the generalized variational functional of Herglotz for several independent variables, J. Anal. Appl. 30 (2011) 253–268.

[11] B. Georgieva, Multiparameter contact transformations, Geom. Integr. Quantization 14 (2013) 87–102.

[12] B. Georgieva and Th. Bodurov, Identities from infinite-dimensional symmetries of Herglotz variational functional, J. Math. Phys. 54 (2013) 062901, 16 pp.
[13] H. Goldstein, Classical mechanics, ed. 2, Addison-Wesley Publishing, 1981.
[14] R. Guenther, A. Gottsch, and C. Guenther, The Herglotz lectures on contact transformations and hamiltonian systems, Juliusz Center for Nonlinear Studies, 1996.
[15] G. Herglotz, Gesammelte Schriften, ed. by H. Schwerdtfeger, Göttingen, Vandenhoeck & Ruprecht, 1979.
[16] G. Herglotz, Berührungstransformationen, Lectures at the University of Göttingen, Göttingen, 1930.
[17] S. Lie, Die Theorie der Integralinvarianten ist ein Korollar der Theorie der Differentialinvarianten, Leipz. Berich. 3 (1897) 342–357.
[18] S. Lie, Gesammelte Abhandlungen, Bd. 6, Teubner, Leipzig, 1927, 649–663.
[19] J. D. Logan, Invariant variational principles, Academic Press, New York, 1977.
[20] R. Mrugala, Contact transformations and brackets in classical thermodynamics, Acta Physica Polonica A58(1) (1980) 19–29.
[21] E. Noether, Invariante Variationsprobleme, Nachr. König. Gesell. Wissen. Göttingen, Matt.-Phys. Kl. (1918) 235–257 (English translation: Transport Theory and Stat. Phys. 1 (1971) 186–207).
[22] E. Noether, Invarianten beliebiger Differentialausdrücke, Nachr. König. Gesell. Wissen Göttingen, Matt.-Phys. Kl. (1918), 37–44.
[23] P. J. Olver, Applications of Lie groups to differential equations, ed. 2, Springer-Verlag, New York, 1993.
[24] P. Roman, Theory of elementary particles, ed. 2, North-Holland Publishing, Amsterdam, 1964.
[25] S. Santos, N. Martins, and D. Torres, Higher-order variational problems of Herglotz type, Vietnam J. Math. 42 (2014) 409–419.
[26] W. Strauss, Nonlinear invariant wave equations, in: Invariant wave equations, ed. by G. Velo and A. Wightman, Lecture Notes in Physics, No. 73, Springer-Verlag, New York, 1978, 197–249.

Received on March 19, 2024
Accepted on May 31, 2024

Bogdana A. Georgieva

Department of Mechatronics, Robotics and Mechanics
Faculty of Mathematics and Informatics
Sofia University "St. Kliment Ohridski"
5, James Bourchier Blvd.
1164 Sofia
BULGARIA

E-mail: georgieva@fmi.uni-sofia.bg

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“
ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Том 111

ANNUAL OF SOFIA UNIVERSITY “ST. KLIMENT OHRIDSKI”
FACULTY OF MATHEMATICS AND INFORMATICS
Volume 111

DIGITAL HOLOGRAM DENOISING BY FILTERING IN THE HOLOGRAM PLANE USING THE HILBERT-HUANG TRANSFORM

SIMEON KARPUZOV, ASSEN SHULEV AND GEORGE PETKOV

Image quality degradation is one of the main problems in Digital Holography. A novel method for noise reduction based on the Hilbert-Huang transform is introduced that reduces the normalized contrast and partially removes the DC-term. It is successfully applied in the hologram plane and requires a single hologram to function. The method causes no reduction in spatial resolution. Moreover, it can be combined with existing noise-filtering methods to improve image quality even further.

Keywords: digital holography, noise filtering, Hilbert-Huang transform, empirical mode decomposition, scattered data interpolation

2020 Mathematics Subject Classification: 68U10, 68W01

CCS Concepts:

- Applied Computing∼Physical sciences and engineering∼Physics

1. Introduction

Digital holography is an optical technique used to capture and display information on both light intensity and phase [16, 28]. It has applications in many different fields, such as microscopy [14, 33], cryptography [6], nondestructive testing [9, 12], art and entertainment displays [35] and others [15, 31]. The technique uses lasers as a light source, and due to their coherent nature, digital holograms exhibit a mixture of highly coherent speckles and additive incoherent noise. These effects severely degrade the quality of the recorded holograms, and significant effort is put into limiting the noise.

Generally speaking, two main classes of techniques tackle the hologram noise problem: optical and numerical. Optical methods involve some reconfiguration of the

DOI: 10.60063/GSU.FMI.111.77-90

standard recording holographic setups, such as multiple wavelength recording [13], white-light illumination [24], moving diffusers and apertures [17] and others [23]. They all show substantial noise reduction qualities, and some of the best available methods are part of this class. Still, additional work and optical elements must be introduced during the digital hologram's recording stage to benefit from them. The second group of methods, the numerical ones, utilizes standard holographic setups but processes numerically the recorded or reconstructed holograms. When operating on the recorded hologram, usually the goal is to produce multiple holograms in which the speckle patterns are de-correlated – the multi-look digital holography approach (MLDH). The best of these methods is demonstrated in [2]. The techniques that operate on the reconstructed image are various spatial filters, the best performing of which is BM3D [3, 4, 8]. Methods based on deep learning techniques also show promise, but they will not explore them here due to some practical considerations such as ease of applicability and result interpretability (especially for medical tasks) [36].

An area of digital hologram denoising left somewhat unexplored is filtering in the hologram plane without relying on the MLDH methods. This paper proposes a technique based on the Hilbert-Huang transform (HHT) [15] for two-dimensional data. The technique uses a single shot and a standard off-axis recording hologram setup. It is different from existing methods in the following two ways:

- The HHT is a global transformation method that works with the entirety of the signal.
- The method works with the recorded hologram, and the noise filtering occurs in the hologram plane.

The 2D-HHT technique can decompose the input hologram into several components, the sum of which is the original hologram. We hypothesize that some components are non-informative and only introduce noise or other unneeded low-frequency data. Such components can be discarded. The interaction between different components is going to be examined. We will omit and add new elements. Various combinations will be tested to find the best. When talking about a "*best one*", it is strictly meant in the conditions of the current experiment with its specific parameters.

2. Methods

The Hilbert-Huang transform is a novel way to decompose a signal into a data-dependent basis through the Empirical Mode Decomposition (EMD) [22] and obtain frequency information by using Hilbert Spectral analysis [19]. HHT works very well on nonstationary and nonlinear signals. This can be utilized in digital holography. The EMD process can be broken down into several steps:

1. Firstly, the local minima and maxima of the input signal $H_{i-1}(x, y)$ are identified;

2. Afterwards, cubic splines are used to connect those extrema, thus forming an upper and lower envelope – $E_u(x, y)$ and $E_l(x, y)$;

3. The mean of the two envelopes – $m(x, y) = \frac{1}{2}[E_u(x, y) + E_l(x, y)]$ is then subtracted from the original signal $H_i(x, y) = H_{i-1}(x, y) - m(x, y)$;

4. The process is repeated with the updated value $H_i(x, y)$ until an intrinsic mode function (IMF) with specific properties is found. Similarly to the one-dimensional case, the mean of the two envelope surfaces is zero and the number of local minima and maxima is approximately the same.

5. With the acquired IMF, steps 1–4 are repeated. The collection of IMFs represents the original data, and their sum is the starting signal – $H(x, y) = \sum_{i=1}^{C} \mathrm{IMF}_i(x, y) + r_c$, where r_c is the signal trend, and C is the total number of IMFs.

After the EMD components have been obtained, frequency analysis can be performed by applying the Hilbert transform to each IMF. Depending on the task, various manipulations and processing can be used for single or multiple components, and components may be dropped or added from the collection of IMFs.

The original method applies to 1D data, but since its introduction, many authors have shown the possibility of extending the ideas to 2D signals [21, 27]. Some take shortcuts regarding interpolation or extrema localization, but we have attempted to stay true to the original technique. The recorded hologram is the 2D data on which we apply the HHT method. Since the data is two-dimensional, upper and lower envelope surfaces are built correspondingly through the local maxima and minima. The mean surface between the upper and lower envelope surfaces is subtracted from the original hologram, and the process continues until a singular frequency component remains. Building the envelope surfaces is a scattered data interpolation problem for which several algorithms have been developed [1, 10]. In the current work, the primary interpolating method is the biharmonic spline [26]. It is combined with a 2D cubic interpolating technique using Delaunay triangulation [20]. We usually calculate the first three to four components with the latter and then finish with the former algorithm. The reason for this approach is that for large data sets, the biharmonic spline requires a large amount of memory – 83 GB for 8 Mpix holograms. Using only cubic interpolation with triangulation also leads to convergence, but more components are generated and have significant amplitude differences. Other methods [25, 29] were also considered. However, it was found that the combination of the methods mentioned above led to a very stable convergence to a single oscillatory IMF. The next part of this paper presents an investigation into the interpolation methods and convergence, as well as component selection.

3. Experimental and method setup

An Nd: YAG laser with a wavelength of 532 nm is used as a light source. The light sensor is a CMOS camera with a pixel size of 1.67×1.67 µm^2. The distance

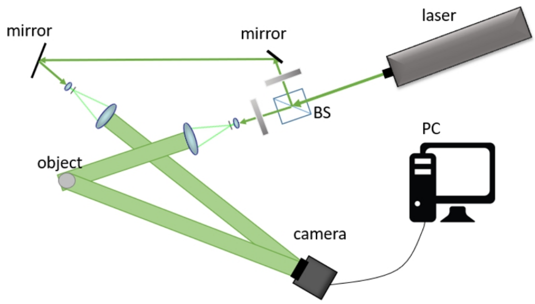


Figure 1. Standard off-axis holographic setup,
B.S. – beam splitter

between the object and the camera is 34 cm. The experimental setup used in the current work is shown in Figure 1.

In the current work, holograms were reconstructed using the Fresnel diffraction formula. Noise reduction is evaluated using the normalized contrast V, defined as

$$V = \frac{\sigma}{\mu}, \tag{3.1}$$

where σ and μ are the standard deviation and the mean of the pixel values of a homogenous area in the image. A lower value of the normalized contrast indicates less noise. Whenever "ground truth" information is available (as is in the case with computer generated holograms) mean squared error (MSE) can be used:

$$\mathrm{MSE} = \frac{\sum_{m,n=1}^{M,N} \left(I_{\mathrm{true}}(m,n) - I_{\mathrm{noisy}}(m,n)\right)^2}{M \times N} \tag{3.2}$$

where I_{true} is the ground truth image, I_{noisy} is the noisy image, $M \times N$ is their size.

The first step in preparing the algorithm that processes the holograms is to achieve convergence of the EMD stage. It is a challenging task as the local extrema are distributed in a non-structured manner throughout the two-dimensional image array (the recorded hologram). Our initial interpolating attempt was trying out the cubic interpolation based on Delaunay triangulation. This technique creates a piecewise C1 surface of the input scattered data. While the method is fast and could achieve convergence, it is not systematic and often produced over 25 components. This uncertain behaviour is unwanted. Purely linear interpolation and others, such as nearest and natural neighbour interpolation, do not achieve convergence. Those methods are also quite different from the spline interpolation used in the 1D case of Huang's original work. Biharmonic splines were also tried out, but due to the high number of extrema ($>$ 4E+04), they turned out to be very computationally and memory expensive. Due to this limitation, the biharmonic spline interpolation technique was chosen only for smaller holograms or EMD components with fewer local extrema ($<$4E+04). For smaller holograms, the method is relatively stable and robust. Shepard's interpolation method was also employed, but it achieved convergence slower than the biharmonic splines. The best result regarding systematic convergence and the number of produced components was achieved using a combination of cubic interpolation and biharmonic splines. The cubic interpolation algorithm calculates the first three or four components as they are the most computationally intensive. After that, biharmonic splines are used. It solves the problems with convergence in the first method and the computational costs of the second. With this combination, we can achieve hologram convergence in 8 or 9 steps (depending on the hologram). An example decomposition using this approach is presented in Figure 2.

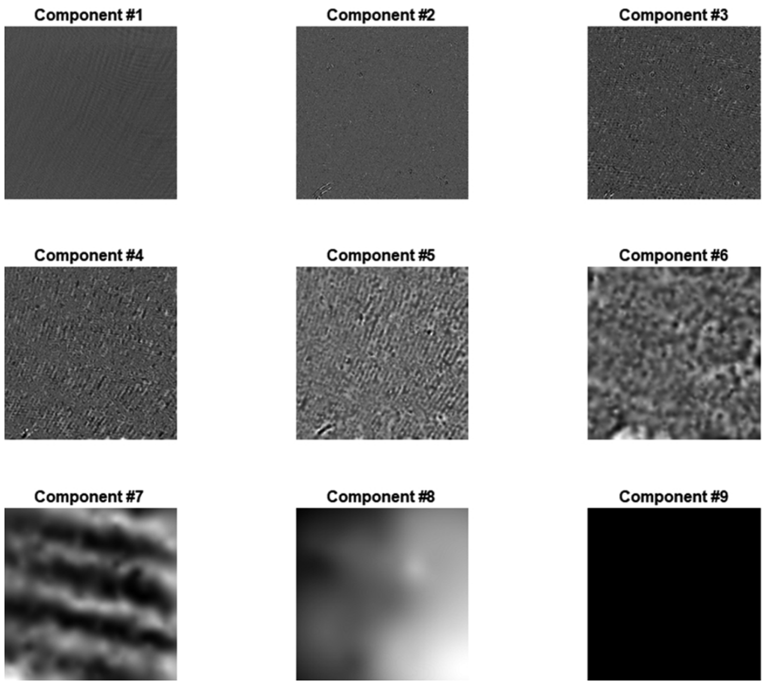


Figure 2. Decomposition of a hologram into 9 EMD components

Processing times as well as extrema count are provided in Table 1. This information provides a general quantitative outline of the computation process. We like to remind again that when only cubic interpolation is used, the process does not converge.

We used a Lenovo™ IdeaPad 130-15IKB laptop with an Intel® Core™ i5-8250 CPU and 20 GB of RAM to derive the processing times in Table 1. They may vary between different computers, development environments and code structure. We can see that the biharmonic spline interpolation for component #5 takes a considerable amount of time, but afterwards the number of extrema quickly goes to 1. Building envelope surfaces using cubic interpolation is quite fast and helps reduce the minima and maxima count so that we can use the biharmonic spline method.

Table 1

Processing time and number of extrema for each EMD component calculation

Interpolation method	Component #	Number of extrema	Processing time [s]
Cubic	1	167.01E+04	23.51
Cubic	2	28.91E+04	13.08
Cubic	3	9.14E+04	11.32
Cubic	4	4.70E+04	10.61
Biharmonic spline	5	3.94E+04	5395.45
Biharmonic spline	6	0.33E+04	423.66
Biharmonic spline	7	424	56.67
Biharmonic spline	8	44	7.41
Biharmonic spline	9	1	0.92

4. Results

After having found a robust way of decomposing an image into its empirical mode components, we investigated how to manipulate them to achieve noise reduction after reconstruction. Firstly, we examined and reconstructed the hologram from each component individually. For different recorded objects all the relevant information was contained in the first two decompositions. The first component contained most useful image information, the second one contained some lower frequency image information, and components 3–9 had other low-frequency signals and the DC-term. This result, regarding the distribution of low and high-frequency data throughout the decomposition components, matches previous findings [5]. The reconstructions of the components can be seen in Figure 3.

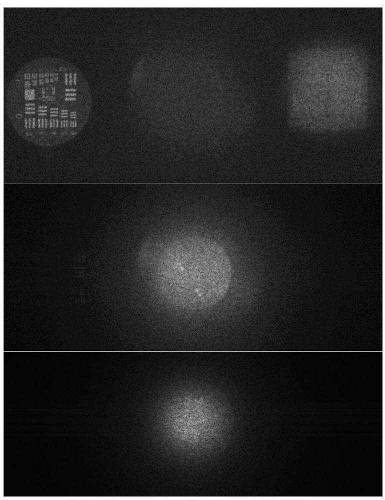


Figure 3. The first, second and third component reconstructions.
White stripes in the second image are visible to the right of the DC-term

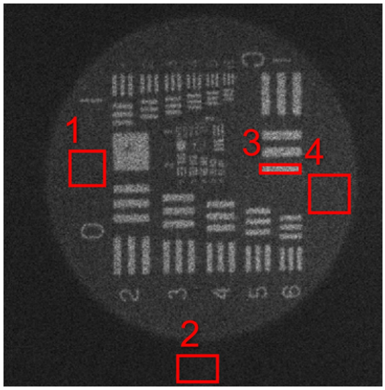


Figure 4. Areas for which V is calculated

This result pointed us to reconstructing the sum of the first two components only and examining the normalized contrast for the areas specified in Figure 4.

Finally, the contrast values for the filtered hologram are compared with the results from a standard Fresnel reconstruction where the hologram is not filtered. A slight but consistent improvement is present, which can be seen in Table 2.

Table 2

Contrast values for the different regions

V	Unfiltered	2D-HHT
Region 1	0.5201	0.5191
Region 2	0.5388	0.5378
Region 3	0.5299	0.5296
Region 4	0.5219	0.5212

We also tried adding additional noise components to the sum closely related to the specific speckle pattern. We tried adding different speckle patterns to various components, removing only selected ones, and adding multiple speckle patterns. However, these manipulations were not beneficial to noise reduction. We also tried a new decomposition of the sum of the first and last components, but it did not produce any significant change. With that, we have concluded that for holograms made with the setup shown in Figure 1, the best application of the 2D-HHT is to select only the first two components and discard the rest. Throughout the rest of the paper, we are going to use only the first two components from the empiric mode decomposition.

Further examination of the method shows its beneficial effect in reducing the zeroth order image (also known as the DC-term). It is a result of the imaging equations and off-axis holographic setup [18]. The DC-term appears as a very bright spot at the center of the reconstructed image and is of no practical use. Although more efficient methods [16] exist for its removal, the effect is considerable and should

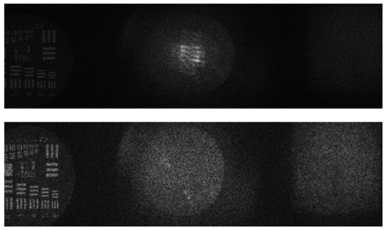


Figure 5. Top – nontreated hologram, bottom – hologram after application of the 2D-HHT algorithm. The images on the bottom are clearer as the bright DC-term is removed

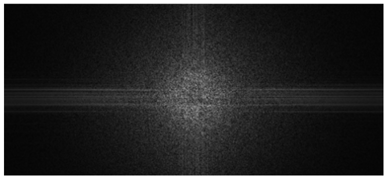


Figure 6. The sum of the seven unwanted HHT components

be mentioned. It can be observed in Figure 5. The sum of the seven discarded components can also be reconstructed and viewed. It consists of parts of the DC-term and background noise and lacks any information about the object. It was calculated that this sum of components has a mean value of 5% of the mean value of the untreated image, meaning that with its removal, 5% of the unwanted signal was discarded. The sum can be observed in Figure 6.

All these discoveries confirm our hypothesis that part of the unwanted signal or noise is contained in the decomposition components, and by discarding them, noise reduction is achieved. It means that 2D-HHT is a viable method for hologram filtering.

An investigation into the smallest resolvable details in both restored images shows no loss of spatial resolution when applying the 2D-HHT algorithm to a hologram (Figure 7).

5. Further investigation

2D-HHT can operate successfully on recorded holograms without a change in resolution so it can be combined with existing noise-filtering methods.

Here we have used 2D-HHT in combination with the Frost filter [11], BM3D [3, 4, 7] and NLM [32, 34] methods. We have chosen these techniques due to their performance and straight-forward application. The NLM filter was set up for a search window size of 11 pixels and a comparison window size of 5 pixels. BM3D was run with default parameters. The Frost filter had a damping parameter of 0.7 and a search window of 5 pixels. We evaluated the noise suppression at three spots in the reconstructed hologram – two inside and one outside the recorded object. It

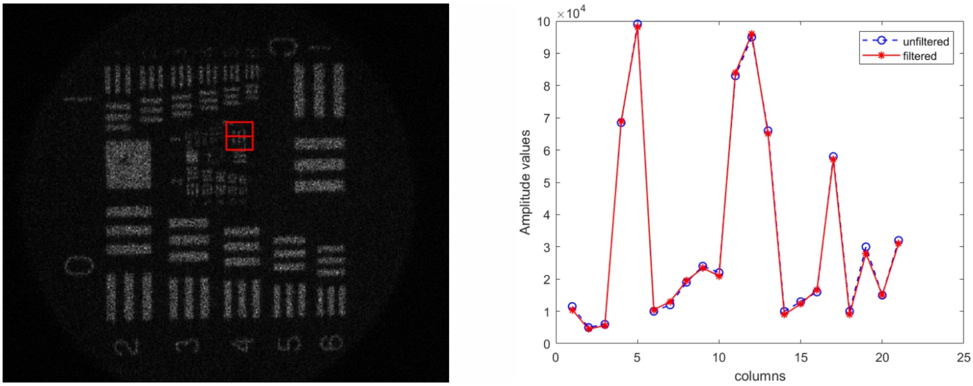


Figure 7. Inspection of the resolution. Left – selection of the finest detail to evaluate, right – a cross-section of the three white stripes, circles – 2D-HHT, stars – unfiltered

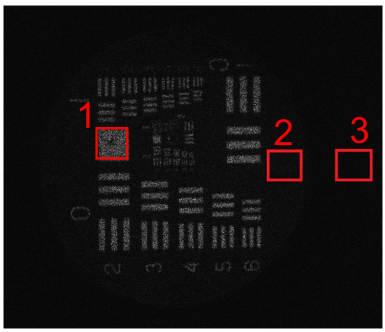


Figure 8. Regions for contrast estimation

gives us a good overview of the overall filtering performance. The spots are shown in Figure 8.

The contrasts for those areas are calculated and presented in Figures 9, 10 and 11.

Figures 9–11 show how contrast is affected when our method is applied to existing filtering techniques. For all areas of interest, an additional noise reduction can be observed for the holograms treated with the 2D-HHT algorithm.

An additional check into the method's usefulness was carried out on simulated holograms. Instead of a conventionally recorded hologram, we generated an interference pattern by a GPU algorithm [30] without adding thermal noise. It was done to show that the current method not only deals with camera noise but also has an overall beneficial effect on the hologram. Contrast values for several areas were calculated and displayed in Table 3.

Using a computer simulated hologram allowed us to measure the MSE value (Eq. (3.2)). Results for different speckle sizes are presented in Figure 13.

Similarly, a small but consistent improvement in the MSE values can be observed for the filtered hologram.

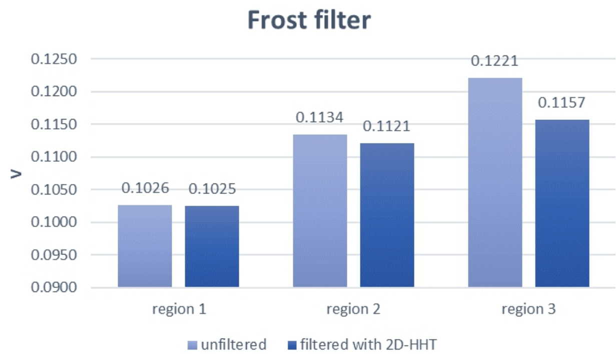


Figure 9. Contrast estimation for the Frost filter. The normalized contrast in on the y-axis. The blue bars on the left show results for application of the Frost filter only, while the orange bars to the right combine our method and the Frost filter

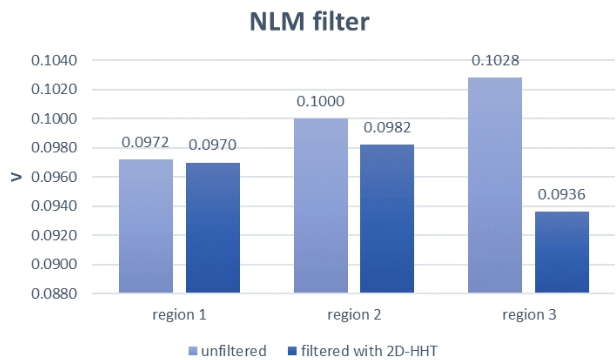


Figure 10. Contrast estimation for the NLM filter. The normalized contrast in on the y-axis. The blue bars on the left show results for application of the NLM filter only, while the orange bars to the right combine our method and the NLM filter

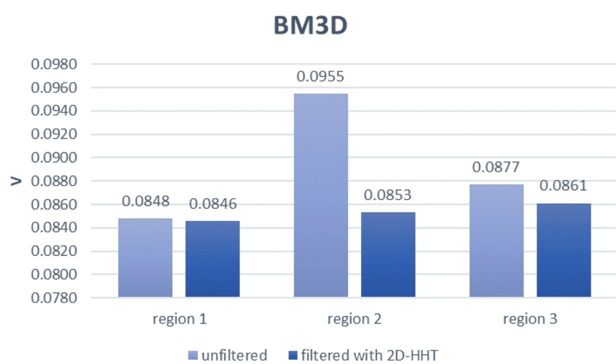


Figure 11. Contrast estimation for the BM3D filter. The normalized contrast in on the y-axis. The blue bars on the left show results for application of the BM3D filter only, while the orange bars to the right combine our method and the BM3D filter

Table 3

Contrast estimation for simulated hologram

V	Unfiltered	2D-HHT
Region 1	0.3603	0.3601
Region 2	0.5277	0.5214
Region 3	0.3706	0.3704

Figure 12. Contrast estimation areas on the simulated hologram

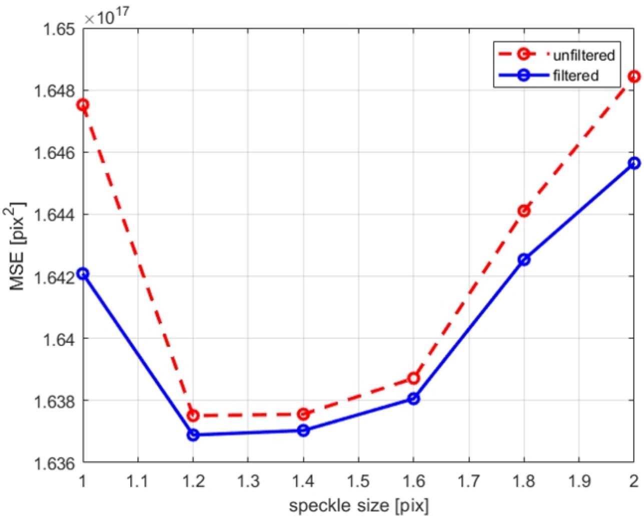


Figure 13. MSE value for different speckle sizes. The curve for the filtered holograms by our method is entirely below the curve for the untreated

6. Conclussion

This paper presented a novel method for noise reduction in digital holograms. It is based on the Hilbert-Huang transform and, unlike existing approaches, tackles the noise problem in the hologram plane. Various beneficial effects of the method were demonstrated, including reduced normalized contrast, DC-term removal and background noise reduction. Furthermore, the technique preserves image resolution, which allows it to be combined with other existing methods. A detailed evaluation of the noise reduction has been carried out. The proposed method was found to keep its beneficial effects when combined with existing techniques. It helps to further reduce the normalized contrast without impact on resolution. Future developments of the method include investigation of parallelization of different parts of the technique, such as the scattered data interpolation algorithms. Calculating the envelope surfaces on a GPU could help reduce processing times.

Acknowledgements

The result presented in this paper is part of the GATE project.

The GATE project has received funding from the European Union's Horizon 2020 WIDESPREAD-2018-2020 TEAMING Phase 2 programme under Grant Agreement No. 857155 and Operational Programme Science and Education for Smart Growth under Grant Agreement No. BG05M2OP001-1.003-0002-C01.

References

[1] P. Alfeld, Scattered data interpolation in three or more variables, in: Mathematical methods in computer aided geometric design, Elsevier, 1989, 1–33.

[2] V. Bianco et al., Random resampling masks: a non-Bayesian one-shot strategy for noise reduction in digital holography, Opt. Lett. 38(5) (2013) 619–621.

[3] V. Bianco et al., Quasi noise-free digital holography, Light Sci. Appl. 5 (2016) e16142.

[4] V. Bianco et al., Strategies for reducing speckle noise in digital holography, Light: Sci. Appl. 7(1) (2018) 48.

[5] A.-O. Boudraa and J.-C. Cexus, EMD-based signal filtering, IEEE Trans. Instrum. Meas. 56(6) (2007) 2196–2202.

[6] W. Chen, B. Javidi and X. Chen, Advances in optical security systems, Adv. Opt. Photonics 6(2) (2014) 120–155.

[7] K. Dabov et al., Image denoising by sparse 3-D transform-domain collaborative filtering, IEEE Trans. Image Process 16(8) (2007) 2080–2095.

[8] K. Dabov et al., Image restoration by sparse 3D transform-domain collaborative filtering, in: Image Processing: Algorithms and Systems VI, SPIE, 2008.

[9] R. Erf, Holographic nondestructive testing, Elsevier, 2012.

[10] R. Franke and G. M. Nielson, Scattered data interpolation and applications: A tutorial and survey, in: Geometric Modeling, ed. by H. Hagen and D. Roller, Computer Graphics – Systems and Applications, Springer, Berlin, Heidelberg, 1991, 131–160.

[11] V. Frost et al., An adaptive filter for smoothing noisy radar images, Proc. IEEE 69(1) (1981) 133–135.

[12] R. Grant and G. Brown, Holographic nondestructive testing (HNDT) in the Automotive Industry, SAE Transactions 78 (1969) 305–313.

[13] H. Jiang, J. Zhao and J. Di, Digital color holographic recording and reconstruction using synthetic aperture and multiple reference waves, Opt. Commun. 285(13–14) (2012) 3046–3049.

[14] M. K. Kim, Digital holographic microscopy, in: Digital Holographic Microscopy, Springer, 2011, 149–190.

[15] K. Kosma et al., Digital holographic interferometry for cultural heritage structural diagnostics: A coherent and a low-coherence optical set-up for the study of a marquetry sample, Strain 54(3) (2018) e12263.

[16] T. Kreis, Handbook of holographic interferometry: optical and digital methods, John Wiley & Sons, 2006.

[17] Y. Kuratomi et al., Speckle reduction mechanism in laser rear projection displays using a small moving diffuser, JOSA A 27(8) (2010) 1812–1817.

[18] E. N. Leith and J. Upatnieks, Holograms: their properties and uses, Opt. Eng. 4(1) (1965) 3–6.

[19] Y.-W. Liu, Hilbert transform and applications, in: Fourier Transform Applications, ed. by S. M. Salih, 2012, 291–300, https://doi.org/10.5772/37727.

[20] E. Quak and L. L. Schumaker, Cubic spline fitting using data dependent triangulations, Comput. Aided Geom. Des. 7(1–4) (1990) 293–301.

[21] N. Rehman and D. P. Mandic, Multivariate empirical mode decomposition, Proc. R. Soc. A: Math. Phys. Eng. Sci. 466(2117) (2010) 1291–1302.

[22] G. Rilling, P. Flandrin and P. Goncalves, On empirical mode decomposition and its algorithms, in: IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing, IEEER Grado, 2003.

[23] L. Rong et al., Speckle noise reduction in digital holography by use of multiple polarization holograms, Chin. Opt. Lett. 8(7) (2010) 653–655.

[24] J. Rosen and G. Brooker, Fluorescence incoherent color holography. Opt. Express 15(5) (2007) 2244–2250.

[25] O. Rukundo and H. Cao, Nearest neighbor value interpolation, arXiv preprint arXiv:1211.1768, 2012.

[26] D. T. Sandwell, Biharmonic spline interpolation of GEOS-3 and SEASAT altimeter data, Geophys. Res. Lett. 14(2) (1987) 139–142.

[27] J. Schmitt et al., 2D Hilbert-Huang transform, in: 2014 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2014.

[28] U. Schnars and W. Jüptner, Direct recording of holograms by a CCD target and numerical reconstruction, Appl. Opt. 33(2) (1994) 179–181.

[29] D. Shepard, A two-dimensional interpolation function for irregularly-spaced data, in: Proc. 1968 23rd ACM National Conf, 1968.

[30] T. Shimobaba et al., Fast calculation of computer-generated-hologram on AMD HD5000 series GPU and OpenCL, Opt. Express 18(10) (2010) 9955–9960.

[31] V. Tornari et al., Impact of relative humidity on wood sample: A climate chamber experimental simulation monitored by digital holographic speckle pattern interferometry, J. Imaging 5(7) (2019) 65.

[32] A. Uzan, Y. Rivenson and A. Stern, Speckle denoising in digital holography by nonlocal means filtering, Appl. Opt. 52(1) (2013) A195–A200.

[33] R. F. Vanligten and H. Osterberg, Holographic microscopy, Nature 211(5046) (1966) 282–283.

[34] J. Xiao et al., Nonlocal means filter based on cosine similarity applied in speckle reduction of digital holography, Appl. Opt. 61(25) (2022) 7474–7481.

[35] F. Yaraş, H. Kang and L. Onural, State of the art in holographic displays: a survey, J. Disp. Technol. 6(10) (2010) 443–454.

[36] T. Zeng, Y. Zhu and E. Y. Lam, Deep learning for digital holography: a review, Opt. Express 29(24) (2021) 40572–40593.

Received on April 9, 2024
Revised on May 21, 2024
Accepted on May 31, 2024

Simeon Karpuzov and George Petkov
Sofia University "St. Kliment Ohridski", GATE Institute
5, James Bourchier Blvd
1164 Sofia
BULGARIA
E-mails: simeon.karpuzov@gate-ai.eu
georgi.petkov@gate-ai.eu

Assen Shulev
Institute of Mechanics, Bulgarian Academy of Sciences
Acad. G. Bontchev St., Bl. 4
1113 Sofia
BULGARIA
E-mail: assen@imbm.bas.bg

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“
ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Том 111

ANNUAL OF SOFIA UNIVERSITY “ST. KLIMENT OHRIDSKI”
FACULTY OF MATHEMATICS AND INFORMATICS
Volume 111

TANGENT CODES

AZNIV K. KASPARIAN and EVGENIYA D. VELIKOVA

The present article studies the finite Zariski tangent spaces to an affine variety X as linear codes, in order to characterize their typical or exceptional properties by global geometric conditions on X. We provide procedures for optimizing one of the parameters length, dimension or minimum distance of a single code by families of tangent codes.

Keywords: Zariski tangent space, minimum distance of a tangent code, genus reduction of a tangent code

2020 Mathematics Subject Classification: Primary: 94B27, 14G50; Secondary: 14G17, 11T71

1. Introduction

Codes with additional structure are usually equipped with a priori properties, which facilitate their characterization and decoding. For instance, algebro-geometric Goppa codes allowed Tsfasman, Vlădut and Zink to improve the asymptotic Gilbert-Varshamov bound on the information rate for a fixed relative minimum distance (cf. [11]). Justesen, Larsen, Elbrønd, Jensen, Havemose, Høholdt, Skorobogatov, Vlădut, Krachkovskii, Porter, Duursma, Feng, Rao and others developed efficient algorithms for decoding Goppa codes after obtaining the error support of the received word (Pellikaan's [8] is a survey on these results). Duursma's considerations from [3] imply that the averaged homogeneous weight enumerator of Goppa codes, associated with a complete set of representatives of the linear equivalence classes of divisors of fixed degree is related to the ζ-polynomial of the underlying curve (cf. [6] for the exact formulation). The realizations of codes by points of a Grassmannian, a determinantal variety or a modification of an arc provide other examples for exploiting “an extra structure” on the objects under study.

DOI: 10.60063/GSU.FMI.111.91-114

The present article interprets the finite Zariski tangent spaces to an affine variety X, defined over a finite field $\mathbb{F}_q$ as linear codes, in order to control the length, the dimension and the minimum distance of these codes by the equations of X. A series of extremal problems from coding theory minimizes the genus $g := n+1-k-d > 0$ of a not MDS $\mathbb{F}_q$-linear $[n,k,d]$-code C. We "deform" any such C into infinite families of linear codes with parameters $[n-1,k,d]$, $[n,k+1,d]$, $[n,k,d+1]$, called respectively a length, a dimension or an weight reductions of C. All of these families decrease the genus by 1.

The parity check matrices of the tangent codes to an affine variety X are the values of the Jacobian matrix of a generating set of the absolute ideal of X. That suggests their possible applications to the theory of convolutional codes (cf. [2, Ch. 9]). Tangent codes to appropriate families of affine varieties seem suitable for studying optimization and asymptotic problems on linear codes, due to their "geometrically integrable dynamical nature".

Here is a synopsis of the paper. Section 2 comprises some preliminaries on the Zariski topology and the Zariski tangent spaces $T_a(X, \mathbb{F}_{q^m})$ to an affine variety X. Our research starts in Section 3 by studying the minimum distance $d(T_a(X, \mathbb{F}_{q^m}))$ of a finite Zariski tangent space $T_a(X, \mathbb{F}_{q^m})$ to an irreducible affine variety $X/\mathbb{F}_q \subset \overline{\mathbb{F}_q}^n$, defined over $\mathbb{F}_q$. Proposition 3.2 (i) establishes that if X has some tangent code of minimum distance $\geq d+1$ then "almost all" finite Zariski tangent spaces to X are of minimum distance $\geq d+1$. The existence of a non-finite puncturing $\Pi_\gamma : X \to \Pi_\gamma(X)$ at $|\gamma| = d$ coordinates prohibits tangent codes of minimum distance $\geq d+1$, according to Proposition 3.2 (ii). Proposition 3.2 (iii) provides two sufficient conditions for the presence of a lower bound $d+1$ on "almost all" tangent codes to X. For an arbitrary $\mathbb{F}_q$-linear $[n,k,d]$-code C, Corollary 3.3 from Subsection 3.1 designs such a "twisted embedding" of $\overline{\mathbb{F}_q^k}$ in $\overline{\mathbb{F}_q}^n$, tangent to $C = T_{0^n}(X, \mathbb{F}_q)$ at the origin 0^n, whose finite Zariski tangent spaces "reproduce" the parameters $[n,k,d]$ of at "almost all the points" of X. By Proposition 3.4, for any family $\pi : \mathcal{C} \to \mathbb{F}_q^n$ of linear codes $\pi^{-1}(a) = \mathcal{C}(a) \subset \mathbb{F}_q^n$ there is an explicit (not necessarily irreducible) affine variety $X \subset \overline{\mathbb{F}_q}^n$, whose Zariski tangent spaces $T_a(X, \mathbb{F}_q) \subseteq \mathcal{C}(a)$ are contained in the members of the family for $\forall a \in \mathbb{F}_q^n$.

Chapter 4 is devoted to the construction of families of genus reductions of an $\mathbb{F}_q$-linear $[n,k,d]$-code C of genus $g := n+1-k-d > 0$. These are parameterized by Zariski open, Zariski dense subsets of affine spaces and defined by explicit polynomial parity check matrices. The length reduction of C with parameters $[n-1,k,d]$ consists of "almost all" tangent codes to the image $\Pi_n(X)$ of the puncturing $\Pi_n : X \to \Pi_n(X)$ of a "twisted embedding" of $\overline{\mathbb{F}_q^k}$ in $\overline{\mathbb{F}_q^n}$, at the last coordinate. The dimension reductions of C with parameters $[n,k+1,d]$ are parameterized by "almost all the points" of $\overline{\mathbb{F}_q}^{2(n-k)}$. Their parity check matrices are obtained by projecting the columns of a parity check matrix $H \in M_{(n-k)\times n}(\mathbb{F}_q)$ of C on appropriate hyperplanes in $\overline{\mathbb{F}_q}^{n-k}$. The existence of a polynomial parity check matrix of weight reductions of C with parameters $[n,k,\geq d+1]$ is established by an induction on the columns of the corresponding parity check matrices.

A work in progress focuses on simultaneous decoding of tangent codes with fixed error support and on the duals of the tangent codes. It relates some standard operations on tangent codes with appropriate operations of the associated affine varieties and constructs morphisms of affine varieties, whose differentials are Hamming isometries of the corresponding tangent codes.

2. Algebraic geometry preliminaries

Let $\overline{\mathbb{F}_q} = \cup_{m=1}^{\infty} \mathbb{F}_{q^m}$ be the algebraic closure of the finite field $\mathbb{F}_q$ with q elements and $\overline{\mathbb{F}_q}^n$ be the n-dimensional affine space over $\overline{\mathbb{F}_q}$. An affine variety $X \subset \overline{\mathbb{F}_q}^n$ is the common zero set

$$X = V(f_1, \dots, f_m) = \{a \in \overline{\mathbb{F}_q}^n \mid f_1(a) = \cdots = f_m(a) = 0\}$$

of polynomials $f_1, \dots, f_m \in \overline{\mathbb{F}_q}[x_1, \dots, x_n]$. We say that $X \subset \overline{\mathbb{F}_q}^n$ is defined over $\mathbb{F}_q$ and denote $X/\mathbb{F}_q \subset \overline{\mathbb{F}_q}^n$ if the absolute ideal

$$I(X, \overline{\mathbb{F}_q}) := \{f \in \overline{\mathbb{F}_q}[x_1, \dots, x_n] \mid f(a) = 0, \forall a \in X\}$$

of X is generated by polynomials $f_1, \dots, f_m \in \mathbb{F}_q[x_1, \dots, x_n]$ with coefficients from $\mathbb{F}_q$.

The affine subvarieties of X form a family of closed subsets. The corresponding topology is referred to as the Zariski topology on X. The Zariski closure $\overline{M}$ of a subset $M \subseteq X$ is defined as the intersection of the Zariski closed subsets Z of X, containing M. It is easy to observe that $\overline{M} = VI(M, \overline{\mathbb{F}_q})$ is the affine variety of the absolute ideal $I(M, \overline{\mathbb{F}_q}) \triangleleft \overline{\mathbb{F}_q}[x_1, \dots, x_n]$ of M. A subset $M \subseteq X$ is Zariski dense if its Zariski closure $\overline{M} = X$ coincides with X. A property $\mathcal{P}(a)$, depending on a point $a \in \overline{\mathbb{F}_q}^n$ holds at a generic point of an affine variety $X \subset \overline{\mathbb{F}_q}^n$ if there is a Zariski dense subset $M \subseteq X$, such that $\mathcal{P}(a)$ is true for all $a \in M$.

An affine variety $X \subset \overline{\mathbb{F}_q}^n$ is irreducible if any decomposition $X = Z_1 \cup Z_2$ into a union of Zariski closed subsets $Z_j \subseteq X$ has $Z_1 = X$ or $Z_2 = X$. This holds exactly when the absolute ideal $I(X, \overline{\mathbb{F}_q}) \triangleleft \overline{\mathbb{F}_q}[x_1, \dots, x_n]$ of X is prime, i.e., $fg \in I(X, \overline{\mathbb{F}_q})$ for $f, g \in \overline{\mathbb{F}_q}[x_1, \dots, x_n]$ requires $f \in I(X, \overline{\mathbb{F}_q})$ or $g \in I(X, \overline{\mathbb{F}_q})$. A prominent property of the irreducible affine varieties X is the Zariski density of an arbitrary non-empty Zariski open subset $U \subseteq X$. This is equivalent to $U \cap W \neq \varnothing$ for any non-empty Zariski open subsets $U \subseteq X$ and $W \subseteq X$.

For an arbitrary irreducible affine variety $X/\mathbb{F}_q \subset \overline{\mathbb{F}_q}^n$, defined over $\mathbb{F}_q$ and an arbitrary constant field $\mathbb{F}_q \subseteq F \subseteq \overline{\mathbb{F}_q}$, the affine coordinate ring

$$F[X] := F[x_1, \dots, x_n]/I(X, F)$$

of X over F is an integral domain. The fraction field

$$F(X) := \left\{ \frac{\varphi_1}{\varphi_2} \,\middle|\, \varphi_1, \varphi_2 \in F[X],\ \varphi_2 \neq 0 \in F[X] \right\}$$

of $F[X]$ is called the function field of X over F. The points $a \in X$ correspond to the maximal ideals $I(a, \overline{\mathbb{F}_q}) \triangleleft \overline{\mathbb{F}_q}[x_1, \ldots, x_n]$, containing $I(X, \overline{\mathbb{F}_q})$. For any F-rational point $a \in X(F) := X \cap F^n$ the localization

$$\mathcal{O}_a(X,F) := \left\{ \frac{\varphi_1}{\varphi_2} \,\middle|\, \varphi_1, \varphi_2 \in F[X],\ \varphi_2(a) \neq 0 \right\}$$

of $F[X]$ at $F[X] \setminus (I(a,F)/I(X,F))$ is the local ring of a in X over F. An F-linear derivation $D_a \colon \mathcal{O}_a(X,F) \to F$ at $a \in X(F)$ is an F-linear map, subject to Leibnitz-Newton rule $D_a(\psi_1\psi_2) = D_a(\psi_1)\psi_2(a) + \psi_1(a)D_a(\psi_2)$, $\forall \psi_1, \psi_2 \in \mathcal{O}_a(X,F)$. The F-linear space

$$T_a(X,F) := \mathrm{Der}_a(\mathcal{O}_a(X,F),F)$$

of the F-linear derivations $D_a \colon \mathcal{O}_a(X,F) \to F$ at $a \in X(F)$ is called the Zariski tangent space to X at a over F.

In order to derive a coordinate description of $T_a(X,F)$, note that any F-linear derivation $D_a \colon \mathcal{O}_a(X,F) \to F$ at $a \in X(F)$ restricts to an F-linear derivation $D_a \colon F[X] \to F$ at a. According to

$$D_a(\varphi_1) = D_a\left(\frac{\varphi_1}{\varphi_2}\right)\varphi_2(a) + \frac{\varphi_1(a)}{\varphi_2(a)} D_a(\varphi_2)$$

for all $\varphi_1, \varphi_2 \in F[X]$ with $\varphi_2(a) \neq 0$, any F-linear derivation $D_a \colon F[X] \to F$ at $a \in X(F)$ has unique extension to an F-linear derivation $D_a \colon \mathcal{O}_a(X,F) \to F$ at a. In such a way, there arises an F-linear isomorphism

$$T_a(X,F) \simeq \mathrm{Der}_a(F[X],F).$$

Any F-linear derivation $D_a \colon F[X] \to F$ of the affine ring $F[X]$ of X at $a \in X(F)$ lifts to an F-linear derivation $D_a \colon F[x_1, \ldots, x_n] \to F$ of the polynomial ring at a, vanishing on the ideal $I(X,F)$ of X over F. If $I(X,F) = \langle f_1, \ldots, f_m \rangle_F \triangleleft F[x_1, \ldots, x_n]$ is generated by $f_1, \ldots, f_m \in F[x_1, \ldots, x_n]$, then for arbitrary $g_1, \ldots, g_m \in F[x_1, \ldots, x_n]$ one has

$$D_a\left(\sum_{i=1}^m f_i g_i\right) = \sum_{i=1}^m D_a(f_i) g_i(a)$$

and the Zariski tangent space

$$T_a(X,F) \simeq \{ D_a \in \mathrm{Der}_a(F[x_1, \ldots, x_n], F) \mid D_a(f_1) = \cdots = D_a(f_m) = 0 \}$$

to X at a consists of the derivations $D_a \colon F[x_1, \ldots, x_n] \to F$ at a, vanishing on $f_1, \ldots, f_m$. In such a way, the coordinate description of $T_a(X,F)$ reduces to the coordinate description of

$$\mathrm{Der}_a(F[x_1, \ldots, x_n], F) = \mathrm{Der}_a(F[\overline{\mathbb{F}_q}^n], F) = T_a(\overline{\mathbb{F}_q}^n, F).$$

In order to endow $T_a(\overline{\mathbb{F}_q}^n, F)$ with a basis over F, let us note that the polynomial ring

$$F[x_1, \ldots, x_n] = F[x_1 - a_1, \ldots, x_n - a_n] = \oplus_{i=0}^\infty F[x_1 - a_1, \ldots, x_n - a_n]^{(i)}$$

has a natural grading by the F-linear spaces $F[x_1 - a_1, \dots, x_n - a_n]^{(i)}$ of the homogeneous polynomials on $x_1 - a_1, \dots, x_n - a_n$ of degree $i \geq 0$. An arbitrary F-linear derivation $D_a \colon F[x_1, \dots, x_n] \to F$ at $a \in F^n$ vanishes on $F[x_1 - a_1, \dots, x_n - a_n]^{(0)} = F$ and on the homogeneous polynomials $F[x_1 - a_1, \dots, x_n - a_n]^{(i)}$ of degree $i \geq 2$. Thus, D_a is uniquely determined by its restriction to the n-dimensional space

$$F[x_1 - a_1, \dots, x_n - a_n]^{(1)} = \operatorname{Span}_F(x_1 - a_1, \dots, x_n - a_n)$$

over F. That enables to identify the Zariski tangent space

$$T_a(\overline{\mathbb{F}_q}^n, F) \simeq \operatorname{Der}_a(F[x_1, \dots, x_n], F) \simeq \operatorname{Hom}_F(F[x_1 - a_1, \dots, x_n - a_n]^{(1)}, F)$$

to $\overline{\mathbb{F}_q}^n$ at a with the space of the F-linear functionals on the homogeneous linear polynomials $F[x_1 - a_1, \dots, x_n - a_n]^{(1)}$. Note that $x_1 - a_1, \dots, x_n - a_n$ is a basis of $F[x_1 - a_1, \dots, x_n - a_n]^{(1)}$ over F and denote by $\left(\frac{\partial}{\partial x_1}\right)_a, \dots, \left(\frac{\partial}{\partial x_n}\right)_a$ its dual basis. In other words, $\left(\frac{\partial}{\partial x_j}\right)_a \in T_a(\overline{\mathbb{F}_q}^n, F)$ are the uniquely determined F-linear functionals on $F[x_1 - a_1, \dots, x_n - a_n]^{(1)}$ with

$$\left(\frac{\partial}{\partial x_j}\right)_a (x_i - a_i) = \delta_{ij} = \begin{cases} 1 & \text{for } 1 \leq i = j \leq n, \\ 0 & \text{for } 1 \leq i \neq j \leq n. \end{cases}$$

As a result, the Zariski tangent space to X at $a \in X(F)$ over F can be described as the linear subspace

$$T_a(X, F) = \left\{ v = \sum_{j=1}^n v_j \left(\frac{\partial}{\partial x_j}\right)_a \;\Big|\; \sum_{j=1}^n v_j \frac{\partial f_i}{\partial x_j}(a) = 0, \ 1 \leq i \leq m \right\}$$

of F^n for any generating set $f_1, \dots, f_m$ of $I(X, F) = \langle f_1, \dots, f_m \rangle_F$.

Definition 2.1. If $F = \mathbb{F}_{q^s}$ is a finite field and $X/\mathbb{F}_q \subset \overline{\mathbb{F}_q}^n$ is an arbitrary irreducible affine variety defined over $\mathbb{F}_q$ then the linear space $T_a(X, \mathbb{F}_{q^s}) \subset \mathbb{F}_{q^s}^n$ over $\mathbb{F}_{q^s}$ is called a tangent code. The parity check matrix of that code is the Jacobian matrix

$$\frac{\partial f}{\partial x} = \frac{\partial(f_1, \dots, f_m)}{\partial(x_1, \dots, x_n)} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \dots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_m}{\partial x_1} & \dots & \frac{\partial f_m}{\partial x_n} \end{pmatrix}$$

of a generating set $f_1, \dots, f_m$ of $I(X, \mathbb{F}_{q^s}) \triangleleft \mathbb{F}_{q^s}[x_1, \dots, x_n]$.

Let $X/\mathbb{F}_q \subset \overline{\mathbb{F}_q}^n$ be an irreducible affine variety, defined over $\mathbb{F}_q$ and $a = (a_1, \dots, a_n) \in X$. The minimal extension $\mathbb{F}_{q^{\delta(a)}} := \mathbb{F}_q(a_1, \dots, a_n)$ of the basic field $\mathbb{F}_q$, which contains the components of a is called the definition field of a. If $\mathbb{F}_{q^{\delta(a_i)}} = \mathbb{F}_q(a_i)$ are the definition fields of $a_i \in \overline{\mathbb{F}_q}$ over $\mathbb{F}_q$, then $\delta(a)$ is the least common multiple of $\delta(a_1), \dots, \delta(a_n)$. Note that $a \in X(\mathbb{F}_{q^m}) := X \cap \mathbb{F}_{q^m}^n$ is an $\mathbb{F}_{q^m}$-rational point if and only if $\delta(a)$ divides m. For all $l \in \mathbb{N}$ the Zariski tangent spaces $T_a(X, \mathbb{F}_{q^{l\delta(a)}})$ have one and a same parity check matrix

$$\frac{\partial f}{\partial x}(a) := \frac{\partial(f_1, \dots, f_m)}{\partial(x_1, \dots, x_n)}(a) \in M_{m \times n}(\mathbb{F}_{q^{\delta(a)}})$$

and are uniquely determined by $T_a(X, \mathbb{F}_{q^{\delta(a)}})$ as the tensor products

$$T_a(X, \mathbb{F}_{q^{l\delta(a)}}) = T_a(X, \mathbb{F}_{q^{\delta(a)}}) \otimes_{\mathbb{F}_{q^{\delta(a)}}} \mathbb{F}_{q^{l\delta(a)}}.$$

In particular, $T_a(X, \mathbb{F}_{q^{\delta(a)}})$ and $T_a(X, \mathbb{F}_{q^{l\delta(a)}})$ have one and a same dimension $n - \mathrm{rk}_{\mathbb{F}_{q^{\delta(a)}}} \frac{\partial f}{\partial x}(a)$ over $\mathbb{F}_{q^{\delta(a)}}$, respectively, over $\mathbb{F}_{q^{l\delta(a)}}$. The minimum distances of $T_a(X, \mathbb{F}_{q^{\delta(a)}})$ and $T_a(X, \mathbb{F}_{q^{l\delta(a)}})$ coincide, as far as they equal the minimal natural number d for which $\frac{\partial f}{\partial x}(a)$ has d linearly dependent columns. From now on, we write $\dim T_a(X, \mathbb{F}_{q^{\delta(a)}})$ for the dimension of $T_a(X, \mathbb{F}_{q^{\delta(a)}})$ over $\mathbb{F}_{q^{\delta(a)}}$.

Let $X = X_1 \cup \dots \cup X_s$ be a reducible affine variety and $a \in X_{i_1} \cap \dots \cap X_{i_r}$ with $1 \leq i_1 < \dots < i_r \leq s$ be a common point of $r \geq 2$ irreducible components X_{i_j} of X. In general, X_{i_j} have different Zariski tangent spaces at a and the union $T_a(X_{i_1}, \mathbb{F}_{q^{\delta(a)}}) \cup \dots \cup T_a(X_{i_r}, \mathbb{F}_{q^{\delta(a)}})$ is not an $\mathbb{F}_{q^{\delta(a)}}$-linear subspace of $\mathbb{F}^n_{q^{\delta(a)}}$. That is why we give the following definition of a tangent code to a reducible variety.

Definition 2.2. If $X/\mathbb{F}_q \subset \overline{\mathbb{F}_q}^n$ is a reducible affine variety, defined over $\mathbb{F}_q$, then the tangent code $T_a(X, \mathbb{F}_{q^{\delta(a)}})$ to X at $a \in X$ is the $\mathbb{F}_{q^{\delta(a)}}$-linear code of length n with parity check matrix

$$\frac{\partial f}{\partial x}(a) = \frac{\partial(f_1, \dots, f_m)}{\partial(x_1, \dots, x_n)}(a) \in M_{m \times n}(\mathbb{F}_{q^{\delta(a)}}),$$

for some generators $f_1, \dots, f_m \in \mathbb{F}_q[x_1, \dots, x_n]$ of $I(X, \overline{\mathbb{F}_q}) = \langle f_1, \dots, f_m \rangle_{\overline{\mathbb{F}_q}}$.

For a systematic study of the Zariski tangent spaces to an affine variety see [1, 7, 9, 10] or [4].

3. Immediate properties of tangent codes construction

3.1. Typical minimum distance of a tangent code

Let us recall that the Hamming weight $w(x)$ of vector $x = (x_1, \dots, x_n) \in \mathbb{F}_q^n$ is the number of the non-zero components and $w(x) \in \{0, 1, \dots, n\}$. The Hamming distance $d(x, y)$ between vectors $x, y \in \mathbb{F}_q^n$ is the number of the different components $x_i \neq y_i$ and $d \colon \mathbb{F}_q^n \times \mathbb{F}_q^n \to \{0, 1, \dots, n\}$, where $d(x, y) := w(x - y)$.

For an arbitrary finite set S and an arbitrary natural number $t \leq |S|$ let us denote by $\binom{S}{t}$ the collection of the t-sets of S, i.e., the family of the unordered subsets of S of cardinality t. In the case of $S = \{1, \dots, n\}$, we write $\binom{1,\dots,n}{t}$ instead of $\binom{\{1,\dots,n\}}{t}$. For an arbitrary subset $\gamma \in \binom{1,\dots,n}{d}$ of $\{1, \dots, n\}$ of cardinality d, the erasing

$$\Pi_\gamma \colon \overline{\mathbb{F}_q}^n \longrightarrow \overline{\mathbb{F}_q}^{n-d}$$

of the components $x_\gamma = (x_{\gamma_1}, \dots, x_{\gamma_d})$, labeled by $\gamma = \{\gamma_1, \dots, \gamma_d\}$ is called the puncturing at γ. If $\neg\gamma = \{1, \dots, n\} \setminus \gamma = \{\delta_1, \dots, \delta_{n-d}\}$ is the complement of γ, then

$$\Pi_\gamma(x_1, \dots, x_n) = x_{\neg\gamma} = (x_{\delta_1}, \dots, x_{\delta_{n-d}}).$$

Any codeword of a linear code $C \subset \mathbb{F}_q^n$, whose weight equals the minimum Hamming distance is in the kernel of some puncturing Π_γ of C at $\gamma \in \binom{1,\dots,n}{d}$. Note that the puncturing

$$\Pi_\gamma \colon T_a(X, \mathbb{F}_{q^{\delta(a)}}) \longrightarrow \Pi_\gamma T_a(X, \mathbb{F}_{q^{\delta(a)}}) \subseteq \mathbb{F}_{q^{\delta(a)}}^{n-|\gamma|}$$

of a finite Zariski tangent space to X coincides with the differential

$$\Pi_\gamma = (d\Pi_\gamma)_a \colon T_a(X, \mathbb{F}_{q^{\delta(a)}}) \longrightarrow T_{\Pi_\gamma(a)}(\Pi_\gamma(X), \mathbb{F}_{q^{\delta(a)}})$$

of the puncturing

$$\Pi_\gamma \colon X \longrightarrow \Pi_\gamma(X)$$

of the corresponding irreducible affine variety X. That allows to study the minimum distance of $T_a(X, \mathbb{F}_{q^{\delta(a)}})$ by the global properties of the puncturing $\Pi_\gamma \colon X \to \Pi_\gamma(X)$ of X.

In order to formulate precisely, let us recall that a finite morphism $\varphi \colon X \to \varphi(X)$ is called separable if the finite extension $\overline{\mathbb{F}_q}(\varphi(X)) \subseteq \overline{\mathbb{F}_q}(X)$ of the corresponding function fields is separable. This means that the minimal polynomial $g_\xi(t) \in \overline{\mathbb{F}_q}(\varphi(X))[t]$ of an arbitrary element $\xi \in \overline{\mathbb{F}_q}(X)$ over $\overline{\mathbb{F}_q}(\varphi(X))$ has no multiple roots.

A morphism $\varphi \colon X \to \varphi(X)$ is infinitesimally injective at some point $a \in X$, if the differential $(d\varphi)_a \colon T_a(X, \mathbb{F}_{q^{\delta(a)}}) \to T_{\varphi(a)}(\varphi(X), \mathbb{F}_{q^{\delta(a)}})$ of φ at a is an $\mathbb{F}_{q^{\delta(a)}}$-linear embedding. Let us denote by $\operatorname{Inf}\operatorname{Inj}(\varphi)$ the set of the points $a \in X$, at which the morphism $\varphi \colon X \to \varphi(X)$ is infinitesimally injective.

Lemma 3.1. *Let us suppose that* $X/\mathbb{F}_q \subset \overline{\mathbb{F}_q}^n$ *is an irreducible affine variety, defined over* $\mathbb{F}_q$ *and*

$$\Pi_\gamma \colon X \longrightarrow \Pi_\gamma(X) \subseteq \overline{\mathbb{F}_q}^{n-d}$$

is its puncturing at $\gamma \in \binom{1,\dots,n}{d}$.

(i) *The infinitesimally injective locus*

$$\operatorname{Inf}\operatorname{Inj}(\Pi_\gamma) = X \setminus V\left(\det \frac{\partial f_\delta}{\partial x_\gamma} \,\Big|\, \delta \in \binom{1,\dots,m}{d}\right) \tag{3.1}$$

is a Zariski open subset of X.

(ii) *If the set* $\operatorname{Inf}\operatorname{Inj}(\Pi_\gamma) \cap \Pi_\gamma^{-1}(\Pi_\gamma(X)^{\text{smooth}}) \neq \varnothing$ *is non-empty, then the puncturing* $\Pi_\gamma \colon X \to \Pi_\gamma(X)$ *is a finite morphism,*

$$\operatorname{Inf}\operatorname{Inj}(\Pi_\gamma) \cap \Pi_\gamma^{-1}(\Pi_\gamma(X)^{\text{smooth}}) \subseteq X^{\text{smooth}}$$

and the differentials

$$(d\Pi_\gamma)_a \colon T_a(X, \mathbb{F}_{q^{\delta(a)}}) \longrightarrow T_{\Pi_\gamma(a)}(\Pi_\gamma(X), \mathbb{F}_{q^{\delta(a)}})$$

are surjective at all the points $a \in \operatorname{Inf}\operatorname{Inj}(\Pi_\gamma) \cap \Pi_\gamma^{-1}(\Pi_\gamma(X)^{\text{smooth}})$.

(iii) *If the puncturing* $\Pi_\gamma \colon X \to \Pi_\gamma(X)$ *is a finite separable morphism then the intersection* $\mathrm{Inf\,Inj}(\Pi_\gamma) \cap \Pi_\gamma^{-1}(\Pi_\gamma(X)^{\mathrm{smooth}}) \neq \varnothing$ *is a Zariski dense subset of* X*. In particular, for a finite* $\Pi_\gamma \colon X \to \Pi_\gamma(X)$*, whose degree* $\deg \Pi_\gamma := [\overline{\mathbb{F}_q}(X) \colon \overline{\mathbb{F}_q}(\Pi_\gamma(X))]$ *is relatively prime to* $p = \mathrm{char}\mathbb{F}_q$*, the subset* $\varnothing \neq \mathrm{Inf\,Inj}(\Pi_\gamma) \cap \Pi_\gamma^{-1}(\Pi_\gamma(X)^{\mathrm{smooth}}) \subseteq X$ *is Zariski dense.*

Proof. (i) The kernel of the differential

$$(d\Pi_\gamma)_a \colon T_a(X, \mathbb{F}_{q^{\delta(a)}}) \longrightarrow T_{\Pi_\gamma(a)}(\Pi_\gamma(X), \mathbb{F}_{q^{\delta(a)}})$$

consists of the tangent vectors $v(a) \in T_a(X, \mathbb{F}_{q^{\delta(a)}})$ with $\mathrm{Supp}(v(a)) \subseteq \gamma$. Thus, $\ker(d\Pi_\gamma) \neq \{0^n\}$ exactly when $\mathrm{rk}\, \frac{\partial f}{\partial x_\gamma}(a) < d$. That justifies

$$X \setminus \mathrm{Inf\,Inj}(\Pi_\gamma) = X \cap V\left(\det \frac{\partial f_\delta}{\partial x_\gamma} \,\Big|\, \delta \in \binom{1, \ldots, m}{d}\right),$$

whereas (3.1).

(ii) Let us recall that $\dim T_a(X, \mathbb{F}_{q^{\delta(a)}}) \geq \dim X = k$ at all the points $a \in X$. If $a \in \mathrm{Inf\,Inj}(\Pi_\gamma) \cap \Pi_\gamma^{-1}(\Pi_\gamma(X)^{\mathrm{smooth}})$, then

$$(d\Pi_\gamma)_a \colon T_a(X, \mathbb{F}_{q^{\delta(a)}}) \longrightarrow T_{\Pi_\gamma(a)}(\Pi_\gamma(X), \mathbb{F}_{q^{\delta(a)}})$$

is injective and $\dim T_{\Pi_\gamma(a)}(\Pi_\gamma(X), \mathbb{F}_{q^{\delta(a)}}) = \dim \Pi_\gamma(X)$. Combining with the inequality $\dim \Pi_\gamma(X) \leq \dim X$, one obtains

$$\begin{aligned}\dim X \leq \dim T_a(X, \mathbb{F}_{q^{\delta(a)}}) &= \dim (d\Pi_\gamma)_a T_a(X, \mathbb{F}_{q^{\delta(a)}}) \\ &\leq \dim T_{\Pi_\gamma(a)}(\Pi_\gamma(X), \mathbb{F}_{q^{\delta(a)}}) = \dim \Pi_\gamma(X) \leq \dim X.\end{aligned}$$

Therefore $(d\Pi_\gamma)_a T_a(X, \mathbb{F}_{q^{\delta(a)}}) = T_{\Pi_\gamma(a)}(\Pi_\gamma(X), \mathbb{F}_{q^{\delta(a)}})$, $\dim X = \dim T_a(X, \overline{\mathbb{F}_q})$ and the dimensions $\dim \Pi_\gamma(X) = \dim X$ coincide. In other words, the differential $(d\Pi_\gamma)_a \colon T_a(X, \mathbb{F}_{q^{\delta(a)}}) \longrightarrow T_{\Pi_\gamma(a)}(\Pi_\gamma(X), \mathbb{F}_{q^{\delta(a)}})$ is surjective, $a \in X^{\mathrm{smooth}}$ is a smooth point and $\Pi_\gamma \colon X \to \Pi_\gamma(X)$ is a finite morphism.

(iii) Without loss of generality, assume that $\gamma = \{1, \ldots, d\}$, whereas $\neg\gamma := \{1, \ldots, n\} \setminus \gamma = \{d+1, \ldots, n\}$. Note that the puncturing $\Pi_\gamma \colon X \to \Pi_\gamma(X)$ is a finite morphism if and only if $\overline{x_s} := x_s + I(X, \overline{\mathbb{F}_q}) \in \overline{\mathbb{F}_q}(X)$ are algebraic over $\overline{\mathbb{F}_q}(\Pi_\gamma(X)) = \overline{\mathbb{F}_q}(\overline{x_{\neg\gamma}})$ for all $1 \leq s \leq d$. Let $g_s(x_s) \in \overline{\mathbb{F}_q}(\Pi_\gamma(X))[x_s]$ be the minimal polynomial of $\overline{x_s}$ over $\overline{\mathbb{F}_q}(\Pi_\gamma(X))$ and $f_s(x_s, x_{\neg\gamma}) \in \overline{\mathbb{F}_q}[x_s, x_{\neg\gamma}]$ be the product of g_s with the least common multiple of the denominators of the coefficients of g_s. Then $f_s(x_s, x_{\neg\gamma})$ is irreducible in $\overline{\mathbb{F}_a}[x_s, x_{\neg\gamma}]$ and defined up to a multiple from $\overline{\mathbb{F}_q}^*$. Moreover, $f_s(x_s, x_{\neg\gamma}) \in I(X, \overline{\mathbb{F}_q})$ is of minimal degree $\deg_{x_s} f_s(x_s, x_{\neg\gamma}) = \deg g_s(x_s) = \deg_{\overline{\mathbb{F}_q}(\Pi_\gamma(X))} \overline{x_s}$ with respect to x_s. According to $f_1, \ldots, f_d \in I(X, \overline{\mathbb{F}_q})$, the Zariski tangent space $T_a(X, \mathbb{F}_{q^{\delta(a)}})$ at an arbitrary point $a \in X$ is contained in the $\mathbb{F}_{q^{\delta(a)}}$-linear code $C(a)$ with parity check matrix

$$\frac{\partial(f_1, \ldots, f_d)}{\partial(x_1, \ldots, x_n)}(a) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(a) & \cdots & 0 & \frac{\partial f_1}{\partial x_{d+1}}(a) & \cdots & \frac{\partial f_1}{\partial x_n}(a) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{\partial f_d}{\partial x_d}(a) & \frac{\partial f_d}{\partial x_{d+1}}(a) & \cdots & \frac{\partial f_d}{\partial x_n}(a) \end{pmatrix}.$$

Note that $\Pi_\gamma^{-1}(\Pi_\gamma(X)^{\rm smooth})$ is a non-empty, Zariski open, Zariski dense subset of the irreducible affine variety X and $\operatorname{Inf}\operatorname{Inj}(\Pi_\gamma) \subseteq X$ is Zariski open by (i), so that the intersection $\operatorname{Inf}\operatorname{Inj}(\Pi_\gamma) \cap \Pi_\gamma^{-1}(\Pi_\gamma(X)^{\rm smooth}) = \varnothing$ only when $\operatorname{Inf}\operatorname{Inj}(\Pi_\gamma) = \varnothing$. We claim that $\operatorname{Inf}\operatorname{Inj}(\Pi_\gamma) = \varnothing$ requires the inseparability of $\overline{x_s} := x_s + I(X, \overline{\mathbb{F}_q}) \in \overline{\mathbb{F}_q}(X)$ over $\overline{\mathbb{F}_q}(\Pi_\gamma)$ for some $1 \leq s \leq d$. This suffices for $\operatorname{Inf}\operatorname{Inj}(\Pi_\gamma) \cap \Pi_\gamma^{-1}(\Pi_\gamma(X)^{\rm smooth}) \neq \varnothing$ in the case of a finite separable morphism $\Pi_\gamma \colon X \to \Pi_\gamma(X)$. The inseparability of $\overline{x_s} := x_s + \in I(X, \overline{\mathbb{F}_q}) \in \overline{\mathbb{F}_q}(X)$ over $\overline{\mathbb{F}_q}(\Pi_\gamma)$ holds only when $p = \operatorname{char}\mathbb{F}_q$ divides the degree

$$\deg_{\overline{\mathbb{F}_q}(\Pi_\gamma(X))} \overline{x_s} := [\overline{\mathbb{F}_q}(\Pi_\gamma(X))(\overline{x_s}) \colon \overline{\mathbb{F}_q}(\Pi_\gamma(X))]$$

of $\overline{x_s}$ over $\overline{\mathbb{F}_q}(\Pi_\gamma(X))$. Bearing in mind that the degree $\deg_{\overline{\mathbb{F}_q}(\Pi_\gamma(X))} \overline{x_s}$ of $\overline{x_s}$ divides the degree $\deg \Pi_\gamma = [\overline{\mathbb{F}_q}(X) \colon \overline{\mathbb{F}_q}(\Pi_\gamma(X))]$ of Π_γ, one concludes that the intersection $\operatorname{Inf}\operatorname{Inj}(\Pi_\gamma) \cap \Pi_\gamma^{-1}(\Pi_\gamma(X)^{\rm smooth}) \neq \varnothing$ is non-empty in the case of $\gcd(\deg \Pi_\gamma, p) = 1$.

By the very definition of an etale morphism, $\operatorname{Inf}\operatorname{Inj}(\Pi_\gamma) = \varnothing$ amounts to the existence of a nowhere vanishing vector field $v \colon X \to \coprod_{a \in X} T_a(X, \mathbb{F}_{q^{\delta(a)}})$ with $\operatorname{Supp} v(a) \subseteq \gamma$ for all $a \in X$. Then $v(a) \in C(a)$ for all $a \in X$ and $\operatorname{rk} \frac{\partial(f_1, \dots, f_d)}{\partial(x_1, \dots, x_d)}(a) < d$. Thus,

$$\det \frac{\partial(f_1, \dots, f_d)}{\partial(x_1, \dots, x_d)}(a) = \prod_{s=1}^{d} \frac{\partial f_s}{\partial x_s}(a) = 0 \quad \text{for} \quad \forall a \in X$$

and $\prod_{s=1}^{d} \frac{\partial f_s}{\partial x_s} \in I(X, \overline{\mathbb{F}_q})$. The absolute ideal $I(X, \overline{\mathbb{F}_q}) \lhd \overline{\mathbb{F}_q}[x_1, \dots, x_n]$ of the irreducible affine variety X is prime, so that there is $1 \leq s \leq d$ with $\frac{\partial f_s}{\partial x_s} \in I(X, \overline{\mathbb{F}_q})$. Since $f_s(x_s, x_{\neg\gamma}) \in I(X, \overline{\mathbb{F}_q})$ is of minimal $\deg_{x_s} f_s(x_s, x_{\neg\gamma})$ and $\deg_{x_s} \frac{\partial f_s(x_s, x_{\neg\gamma})}{\partial x_s} < \deg_{x_s} f_s(x_s, x_{\neg\gamma})$, there follows $\frac{\partial f_s(x_s, x_{\neg\gamma})}{\partial x_s} \equiv 0_{\overline{\mathbb{F}_q}} \in \overline{\mathbb{F}_q}[x_s, x_{\neg\gamma}]$. As a result, $\frac{\partial g_s(x_s)}{\partial x_s} \equiv 0$ and $\overline{x_s}$ is inseparable over $\overline{\mathbb{F}_q}(\Pi_\gamma(X))$. □

Note that Lemma 3.1 (ii) is a sort of a generalization of the Implicit Function Theorem, according to which a puncturing $\Pi_\gamma \colon X \to \Pi_\gamma(X)$ with an injective differential at some $a \in \Pi_\gamma^{-1}(\Pi_\gamma(X))^{\rm smooth}$ is a finite morphism.

For an arbitrary irreducible affine variety $X/\mathbb{F}_q \subset \overline{\mathbb{F}_q}^n$, defined over $\mathbb{F}_q$, let us denote by

$$X^{(\leq d)} := \{a \in X \mid d(T_a(X, \mathbb{F}_{q^{\delta(a)}})) \leq d\}$$

the set of the points $a \in X$, at which the finite Zariski tangent spaces are of minimum distance $\leq d$. Similarly, put

$$X^{(d)} := \{a \in X \mid d(T_a(X, \mathbb{F}_{q^{\delta(a)}})) = d\} \text{ and } X^{(\geq d)} := \{a \in X \mid d(T_a(X, \mathbb{F}_{q^{\delta(a)}})) \geq d\}.$$

The next proposition establishes that if an irreducible affine variety X admits a tangent code $T_a(X, \mathbb{F}_{q^{\delta(a)}})$ of minimum distance $\geq d + 1$ then "almost all" finite Zariski tangent spaces to X are of minimum distance $\geq d+1$. If there is a non-finite puncturing $\Pi_\gamma \colon X \to \Pi_\gamma(X)$ at $|\gamma| = d$ variables, we show that all the tangent codes to X are of minimum distance $\leq d$. When all the puncturings $\Pi_\gamma \colon X \to \Pi_\gamma(X)$ at

$|\gamma| = d$ variables are finite and separable, the minimum distance of a finite Zariski tangent space to X is bounded below by $d+1$ at "almost all" points of X.

Proposition 3.2. *Let $X/\mathbb{F}_q \subset \overline{\mathbb{F}_q}^n$ be an irreducible affine variety of dimension $k \in \mathbb{N}$, defined over $\mathbb{F}_q$.*

(i) *For an arbitrary natural number $d \leq n-k+1$ the locus*

$$X^{(\geq d+1)} = \cap_{\gamma \in \binom{1,\dots,n}{d}} \operatorname{Inf}\operatorname{Inj}(\Pi_\gamma)$$
$$= X \setminus V\left(\prod_{i \in \binom{1,\dots,n}{d}} \det \frac{\partial f_{\varphi(i)}}{\partial x_i} \,\middle|\, \varphi\colon \binom{1,\dots,n}{d} \to \binom{1,\dots,m}{d} \right)$$

is a Zariski open subset of X.

(ii) *If there is a non-finite puncturing $\Pi_\gamma\colon X \to \Pi_\gamma(X)$ at $|\gamma| = d$ coordinates, then $X = X^{(\leq d)}$. Moreover, in the case of $X^{(d)} \neq \varnothing$ the locus $X^{(d)} = X^{(\geq d)}$ is a Zariski dense, Zariski open subset of X.*

(iii) *If for any $\gamma \in \binom{1,\dots,n}{d}$ the puncturing $\Pi_\gamma\colon X \to \Pi_\gamma(X)$ is finite and separable, then the subset $X^{(\geq d+1)} \subseteq X$ is Zariski dense. In particular, if for any $\gamma \in \binom{1,\dots,n}{d}$ the puncturing $\Pi_\gamma\colon X \to \Pi_\gamma(X)$ is a finite morphism with $\gcd(\deg \Pi_\gamma, \operatorname{char}\mathbb{F}_q) = 1$ for $\deg \Pi_\gamma := [\overline{\mathbb{F}_q}(X)\colon \overline{\mathbb{F}_q}(\Pi_\gamma(X))]$, then $X^{(\geq d+1)}$ is a Zariski dense subset of X.*

Proof. (i) Let us observe that $a \in X^{(\geq d+1)}$ if and only if there is no tangent vector $v \in T_a(X, \mathbb{F}_{q^{\delta(a)}}) \setminus \{0^n\}$ with $\operatorname{Supp}(v) \subseteq \gamma$ for some $\gamma \in \binom{1,\dots,n}{d}$. That amounts to

$$\ker(d\Pi_\gamma)_a = \{v \in T_a(X, \mathbb{F}_{q^{\delta(a)}}) \mid \operatorname{Supp}(v) \subseteq \gamma\} = \{0^n\}$$

and holds exactly when $a \in \operatorname{Inf}\operatorname{Inj}(\Pi_\gamma)$ for all $\gamma \in \binom{1,\dots,n}{d}$.

Let $I(X, \overline{\mathbb{F}_q}) = \langle f_1, \dots, f_m \rangle \lhd \overline{\mathbb{F}_q}[x_1, \dots, x_n]$ be generated by some polynomials $f_1, \dots, f_m \in \mathbb{F}_q[x_1, \dots, x_n]$. Then $a \in X^{(\geq d+1)}$ exactly when any d-tuple of columns of $\frac{\partial f}{\partial x}(a)$ is linearly independent. In other words, $\operatorname{rk} \frac{\partial f}{\partial x_i}(a) = \operatorname{rk} \frac{\partial(f_1,\dots,f_m)}{\partial(x_{i_1},\dots,x_{i_d})}(a) = d$ for all $i \in \binom{1,\dots,n}{d}$. By $k = \dim X \geq n-m$ there follows $m \geq n-k \geq d$ and $\operatorname{rk} \frac{\partial f}{\partial x_i}(a) = d$ is equivalent to $\det \frac{\partial f_\gamma}{\partial x_i}(a) \neq 0$ for some $\gamma \in \binom{1,\dots,m}{d}$. Thus,

$$\begin{aligned} X^{(\geq d+1)} &= \cap_{i \in \binom{1,\dots,n}{d}} \left[\cup_{\gamma \in \binom{1,\dots,m}{d}} \left(X \setminus V\left(\det \frac{\partial f_\gamma}{\partial x_i} \right) \right) \right] \\ &= \cap_{i \in \binom{1,\dots,n}{d}} \left[X \setminus V\left(\det \frac{\partial f_\gamma}{\partial x_i} \,\middle|\, \gamma \in \binom{1,\dots,m}{d} \right) \right] \\ &= X \setminus \cup_{i \in \binom{1,\dots,n}{d}} V\left(\det \frac{\partial f_\gamma}{\partial x_i} \,\middle|\, \gamma \in \binom{1,\dots,m}{d} \right) \\ &= X \setminus V\left(\prod_{i \in \binom{1,\dots,n}{d}} \det \frac{\partial f_{\varphi(i)}}{\partial x_i} \,\middle|\, \varphi\colon \binom{1,\dots,n}{d} \to \binom{1,\dots,m}{d} \right), \end{aligned} \tag{3.2}$$

where $\varphi\colon \binom{1,\dots,n}{d} \to \binom{1,\dots,m}{d}$ vary over all the maps of the collection of the subsets of $\{1,\dots,n\}$ of cardinality d in the family of the subsets of $\{1,\dots,m\}$ of cardinality d. The last equality in (3.2) follows from

$$\cup_{i\in\binom{1,\dots,n}{d}} V(S_i) = V\left(\prod_{i\in\binom{1,\dots,n}{d}} S_i\right)$$

for

$$\prod_{i\in\binom{1,\dots,n}{d}} S_i := \left\{\prod_{i\in\binom{1,\dots,n}{d}} g_i \mid g_i \in S_i\right\}, \quad S_i := \left\{\det \frac{\partial f_\gamma}{\partial x_i}\middle| \gamma \in \binom{1,\dots,m}{d}\right\}.$$

(ii) We claim that at any point $a \in \Pi_\gamma^{-1}(\Pi_\gamma(X)^{\mathrm{smooth}})$ the Zariski tangent space $T_a(X, \mathbb{F}_{q^{\delta(a)}})$ contains a non-zero word, supported by γ. To this end, it suffices to establish that the differential

$$(d\Pi_\gamma)_a\colon T_a(X, \mathbb{F}_{q^{\delta(a)}}) \longrightarrow T_{\Pi_\gamma(a)}(\Pi_\gamma(X), \mathbb{F}_{q^{\delta(a)}})$$

of Π_γ at a is non-injective. Assume the opposite, i.e., that $\ker(d\Pi_\gamma)_a = 0$. Then

$$k \leq \dim T_a(X, \mathbb{F}_{q^{\delta(a)}}) \leq \dim T_{\Pi_\gamma(a)}(\Pi_\gamma(X), \mathbb{F}_{q^{\delta(a)}}) = \dim \Pi_\gamma(X).$$

The morphism $\Pi_\gamma\colon X \to \Pi_\gamma(X)$ is not finite, so that $\dim \Pi_\gamma(X) < \dim X = k$. That leads to a contradiction and implies that $\ker(d\Pi_\gamma)_a \neq 0$ at any point $a \in \Pi_\gamma^{-1}(\Pi_\gamma(X)^{\mathrm{smooth}})$. As a result, $\Pi_\gamma^{-1}(\Pi_\gamma(X)^{\mathrm{smooth}}) \subseteq X^{(\leq d)}$. According to (i), $X^{(\leq d)}$ is a Zariski closed subset of X. The non-empty, Zariski open, Zariski dense subset $\Pi_\gamma^{-1}(\Pi_\gamma(X)^{\mathrm{smooth}})$ of X is Zariski dense, so that

$$X = \overline{\Pi_\gamma^{-1}(\Pi_\gamma(X)^{\mathrm{smooth}})} \subseteq \overline{X^{(\leq d)}} = X^{(\leq d)},$$

whereas $X = X^{(\leq d)}$. Now, $X^{(d)} = X^{(\leq d)} \cap X^{(\geq d)} = X \cap X^{(\geq d)} = X^{(\geq d)}$ is a Zariski open subset of X, whereas Zariski dense for $X^{(d)} \neq \varnothing$.

(iii) According to Lemma 3.1 (iii), if $\Pi_\gamma\colon X \to \Pi_\gamma(X)$ is a finite separable morphism or a finite morphism with $\gcd(\deg \Pi_\gamma, \mathrm{char}\mathbb{F}_q) = 1$, then $\mathrm{Inf\,Inj}(\Pi_\gamma) \cap \Pi_\gamma^{-1}(\Pi_\gamma(X)^{\mathrm{smooth}}) \neq \varnothing$. In particular, $\mathrm{Inf\,Inj}(\Pi_\gamma) \neq \varnothing$. Since $\mathrm{Inf\,Inj}(\Pi_\gamma)$ is Zariski open by Lemma 3.1 (i), the finite intersection $X^{(\geq d+1)} = \cap_{\gamma\in\binom{1,\dots,n}{d}} \mathrm{Inf\,Inj}(\Pi_\gamma)$ of the non-empty, Zariski open subsets $\mathrm{Inf\,Inj}(\Pi_\gamma) \subseteq X$ is a non-empty, Zariski open, Zariski dense subset of the irreducible affine variety X. □

The above proposition reveals that for any point $a \in X^{(d)}$ there exists a d-tuple of indices $\gamma \in \binom{1,\dots,n}{d}$, such that $\Pi_\gamma\colon X \to \Pi_\gamma(X)$ is not infinitesimally injective at a.

3.2. Reproducing the dimension and the minimum distance of a code

For an arbitrary $\mathbb{F}_q$-linear $[n,k,d]$-code C we provide explicit equations of a twisted embedding of $\overline{\mathbb{F}_q}^k$ in $\overline{\mathbb{F}_q}^n$, whose tangent codes $T_a(X, \mathbb{F}_{q^{\delta(a)}})$ at a generic point reproduce the length n, the dimension k and the minimum distance d of C.

If not specified otherwise, $H = (H_1 \dots H_n)$ is a parity check matrix of the linear code under consideration. For any $\lambda \in \binom{1,\dots,n}{t}$ we denote by H_λ the columns of H, labeled by λ. If $\mu \in \binom{1,\dots,m}{s}$, then $H_{\mu,\lambda}$ is the collection of the rows of H_λ, labeled by μ.

Corollary 3.3. *Let C be an $\mathbb{F}_q$-linear $[n,k,d]$-code and $\sigma \in \binom{1,\dots,n}{d}$ be the support of a non-zero word $c \in C \setminus \{0^n\}$. Then there is a smooth irreducible k-dimensional affine variety $X/\mathbb{F}_q \subset \overline{\mathbb{F}_q}^n$, isomorphic to $\overline{\mathbb{F}_q}^k$, such that $0^n \in X$, $T_{0^n}(X, \mathbb{F}_q) = C$ and $c \in T_a(X, \mathbb{F}_{q^{\delta(a)}})$ for all $a \in X$.*

In particular, $X = X^{(\leq d)}$, so that $X^{(d)} = X^{(\geq d)} \neq \varnothing$ is a Zariski open, Zariski dense subset of X and $T_a(X, \mathbb{F}_{q^{\delta(a)}})$ are $[n,k,d]$-codes for all $a \in X^{(d)}$.

Proof. Let $H \in M_{(n-k)\times n}(\mathbb{F}_q)$ be a parity check matrix of the code C with columns $H_s \in M_{(n-k)\times 1}(\mathbb{F}_q)$ and $\sigma' = \sigma \setminus \{\sigma_d\}$ for some $\sigma_d \in \sigma$. Since C is of minimum distance d, the columns of H, labeled by σ' are linearly independent. Bearing in mind that H is of $\mathrm{rk}(H) = n-k$, one concludes the existence of $\tau \in \binom{\{1,\dots,n\}\setminus\sigma}{n-k-d+1}$, such that the square matrix $H_{\sigma'\cup\tau} = (H_{\sigma'} H_\tau) \in M_{(n-k)\times(n-k)}(\mathbb{F}_q)$ is non-singular. If $s \in \sigma \cup \tau$ and $1 \leq i \leq n-k$, then let $f_{i,s}(x_s) := H_{i,s}x_s$. For $s \in \{1,\dots,n\} \setminus (\sigma\cup\tau)$ and $1 \leq i \leq n-k$ take

$$f_{i,s}(x_s) := H_{i,s}x_s + \sum_{r=2}^{m_{i,s}} b_{i,s,r}x_s^r \in \mathbb{F}_q[x_s]$$

for some $m_{i,s} \in \mathbb{N} \setminus \{1\}$ and $b_{i,s,r} \in \mathbb{F}_q$, $\forall 2 \leq r \leq m_{i,s}$. Consider

$$f_i(x_1,\dots,x_n) := \sum_{s=1}^n f_{i,s}(x_s) = \sum_{s=1}^n H_{i,s}x_s + \sum_{s\in\{1,\dots,n\}\setminus(\sigma\cup\tau)} \sum_{r=2}^{m_{i,s}} b_{i,s,r}x_s^r$$

for all $1 \leq i \leq n-k$ and the affine variety $X := V(f_1,\dots,f_{n-k}) \subset \overline{\mathbb{F}_q}^n$, defined over $\mathbb{F}_q$. Let us denote $\rho := \{1,\dots,n\} \setminus (\sigma'\cup\tau)$ and observe that $f_i(x_1,\dots,x_n) = 0$ are equivalent to

$$\sum_{s\in\sigma'\cup\tau} H_{i,s}x_s = g_i(x_\rho) \quad \text{for some} \quad g_i(x_\rho) \in \mathbb{F}_q[x_\rho] \quad \text{and all } 1 \leq i \leq n-k.$$

Viewing $x_{\sigma'\cup\tau}$ as a column of variables, labeled by $\sigma'\cup\tau \in \binom{1,\dots,n}{n-k}$, one can write the equations of X in the form

$$H_{\sigma'\cup\tau}\, x_{\sigma'\cup\tau} = \begin{pmatrix} g_1(x_\rho) \\ \vdots \\ g_{n-k}(x_\rho) \end{pmatrix}.$$

The invertibility of $H_{\sigma'\cup\tau}$ allows to represent the equations of X in the form

$$x_{\sigma'\cup\tau} = (H_{\sigma'\cup\tau})^{-1}\begin{pmatrix} g_1(x_\rho) \\ \vdots \\ g_{n-k}(x_\rho)\end{pmatrix}.$$

Thus, the puncturing $\Pi_{\sigma'\cup\tau}\colon X \to \overline{\mathbb{F}_q}^k$ at $\sigma'\cup\tau \in \binom{1,\dots,n}{n-k}$ is biregular, with inverse

$$(\Pi_{\sigma'\cup\tau})^{-1}(x_\rho) = \left((H_{\sigma'\cup\tau})^{-1}\begin{pmatrix} g_1(x_\rho) \\ \vdots \\ g_{n-k}(x_\rho)\end{pmatrix}, x_\rho\right).$$

In particular, X is a smooth irreducible affine variety of dimension $\dim X = k$.

The tangent spaces $T_a(X, \mathbb{F}_{q^{\delta(a)}})$ at all the points $a \in X$ are linear codes of length n and dimension k, whose parity check matrices $\frac{\partial(f_1,\dots,f_{n-k})}{\partial(x_1,\dots,x_n)}(a)$ have columns $H_{\sigma\cup\tau}$, labeled by $\sigma\cup\tau \in \binom{1,\dots,n}{n-k+1}$. That is why $c \in C$ with $\mathrm{Supp}(c) = \sigma$ belongs to $T_a(X, \mathbb{F}_{q^{\delta(a)}})$ for $\forall a \in X$ and the minimum distance $d(T_a(X, \mathbb{F}_{q^{\delta(a)}})) \leq d$ at $\forall a \in X$. In other words, $X = X^{(\leq d)}$. By the very construction of $f_i(x_1,\dots,x_n)$ one has $0^n \in X$ and $\frac{\partial(f_1,\dots,f_{n-k})}{\partial(x_1,\dots,x_n)}(0^n) = H$, whereas $T_{0^n}(X, \mathbb{F}_q) = C$. As a result, $0^n \in X^{(d)} = X^{(\geq d)}$ is non-empty and the Zariski tangent space $T_a(X, \mathbb{F}_{q^{\delta(a)}})$ at a generic $a \in X$ is an $[n,k,d]$-code. □

The above proposition reveals that a single linear code C does not reflect global properties of the affine varieties X, tangent to C at some point $a \in X$. It illustrates how the equations of X govern the behavior of a generic tangent code to X.

3.3. Inscription of Zariski tangent spaces in families of linear codes

Proposition 3.4. *Let $\mathcal{C} \to S$ be a family of $\mathbb{F}_q$-linear codes $\mathcal{C}(a) \subset \mathbb{F}_q^n$, $a \in S$ of arbitrary dimension and minimum distance, parameterized by a subset $S \subseteq \mathbb{F}_q^n$. Then there exists a (not necessarily irreducible) affine variety $X \subseteq \overline{\mathbb{F}_q}^n$, containing all the $\mathbb{F}_q$-rational points $\mathbb{F}_q^n$ of $\overline{\mathbb{F}_q}^n$ and such that $T_a(X, \mathbb{F}_q) \subseteq \mathcal{C}(a)$ at $\forall a \in S$.*

Proof. Let $\mathcal{H} \to S$ be a family of parity-check matrices $\mathcal{H}(a) \in M_{(n-k)\times n}(\mathbb{F}_q)$ of $\mathcal{C}(a) \subset \mathbb{F}_q^n$ for all $a \in S$ and denote by $\mathcal{H}(a)_{ij} \in \mathbb{F}_q$ the entries of these matrices. For an arbitrary $\beta \in \mathbb{F}_q$, consider the Lagrange basis polynomial

$$L^\beta_{\mathbb{F}_q}(t) := \prod_{\alpha\in\mathbb{F}_q\setminus\{\beta\}} \frac{t-\alpha}{\beta-\alpha}$$

with $L^\beta_{\mathbb{F}_q}(t)(\beta) = 1$ and $L^\beta_{\mathbb{F}_q}(t)\big|_{\mathbb{F}_q\setminus\{\beta\}} = 0$. Straightforwardly,

$$L^0_{\mathbb{F}_q}(t) = -t^{q-1} + 1 \quad \text{and} \quad L^\beta_{\mathbb{F}_q}(t) = -t^{q-1} - \sum_{s=1}^{q-2} \beta^{-s}t^s, \quad \forall \beta \in \mathbb{F}_q^*.$$

Let us denote by

$$\Phi_p \colon \overline{\mathbb{F}_q}^n \longrightarrow \overline{\mathbb{F}_q}^n, \quad \Phi_p(a_1,\dots,a_n) = (a_1^p,\dots,a_n^p), \quad \forall a = (a_1,\dots,a_n) \in \overline{\mathbb{F}_q}^n$$

the Frobenius automorphism of degree $p = \mathrm{char}\mathbb{F}_q$ and consider

$$f_i(x_1,\dots,x_n) := \sum_{b\in\Phi_p(S)} \left[\sum_{j=1}^n \mathcal{H}(\Phi_p^{-1}(b))_{ij}(x_j - x_j^q) \right] L_{\mathbb{F}_q}^{b_1}(x_1^p)\dots L_{\mathbb{F}_q}^{b_n}(x_n^p) \in \mathbb{F}_q[x_1,\dots,x_n]$$

for $1 \leq i \leq n-k$. The affine algebraic set $X := V(f_1,\dots,f_{n-k}) \subset \overline{\mathbb{F}_q}^n$ is claimed to satisfy the announced conditions. First of all, X passes through all the $\mathbb{F}_q$-rational points $\mathbb{F}_q^n$ of the affine space $\overline{\mathbb{F}_q}^n$, since $\forall a = (a_1,\dots,a_n) \in \mathbb{F}_q^n$ has components $a_j = a_j^q$ and $f_i(a_1,\dots,a_n) = 0$ for $\forall 1 \leq i \leq n-k$. The partial derivatives of f_i are $\frac{\partial f_i}{\partial x_j} = \sum_{b\in\Phi_p(S)} \mathcal{H}(\Phi_p^{-1}(b))_{ij} L_{\mathbb{F}_q}^{b_1}(x_1^p)\dots L_{\mathbb{F}_q}^{b_n}(x_n^p)$ and their values at $a \in S \subseteq \mathbb{F}_q^n$ equal $\frac{\partial f_i}{\partial x_j}(a) = \mathcal{H}(\Phi_p^{-1}\Phi_p(a))_{ij} = \mathcal{H}(a)_{ij}$. Note that the composition of Lagrange interpolation polynomials with the Frobenius automorphism Φ_p is designed in such a way that to adjust

$$\frac{\partial(f_1,\dots,f_{n-k})}{\partial(x_1,\dots,x_n)}(a) = \mathcal{H}(a)$$

at all the points $a \in S$. By $f_1,\dots,f_{n-k} \in I(X,\overline{\mathbb{F}_q}) = r(\langle f_1,\dots,f_{n-k}\rangle)$ for the radical $r(\langle f_1,\dots,f_{n-k}\rangle) \triangleleft \overline{\mathbb{F}_q}[x_1,\dots,x_n]$ of $\langle f_1,\dots,f_{n-k}\rangle \triangleleft \overline{\mathbb{F}_q}[x_1,\dots,x_n]$, the Zariski tangent space $T_a(X,\mathbb{F}_q) \subseteq \mathcal{C}(a)$ to X at an arbitrary point $a \in S$ is contained in the linear code $\mathcal{C}(a)$ with parity check matrix $\frac{\partial(f_1,\dots,f_{n-k})}{\partial(x_1,\dots,x_n)}(a)$. □

4. Families of genus reductions of a code

The genus of an $\mathbb{F}_q$-linear $[n,k,d]$-code C is defined as the deviation $g := n+1-k-d$ of its parameters from the equality in the Singleton Bound $n+1-k-d \geq 0$. One of the problems in coding theory is to obtain a linear code C' of genus $g' = g-1 \geq 0$ from the given linear code C of genus $g \geq 1$. We say that C' is a genus reduction of C. There are three standard ways for construction of a genus reduction C'. These are, respectively, the length, the dimension and the weight reductions of C with parameters $[n-1,k,d]$, $[n,k+1,d]$, $[n,k,d+1]$. In the next three subsections we use the set up of tangent codes, in order to construct families of length, dimension and weight reductions of C, parameterized by appropriate Zariski dense subsets of affine spaces over $\overline{\mathbb{F}_q}$.

4.1. A family of length reductions of a linear code

Here is a simple lemma from coding theory, which will be used for the construction of a family of length reductions of a linear code.

Lemma 4.1. *Let C be an $\mathbb{F}_q$-linear code of genus $g = n+1-k-d > 0$ with a parity check matrix $H = (H_1 \dots H_n) \in M_{(n-k)\times n}(\mathbb{F}_q)$. If*

$$H_n \notin \cup_{\lambda \in \binom{1,\dots,n-1}{d-1}} \operatorname{Span}_{\mathbb{F}_q}(H_\lambda), \tag{4.1}$$

then the image $\Pi_n(C) \subset \mathbb{F}_q^{n-1}$ of the puncturing $\Pi_n : C \to \Pi_n(C)$ of the last component is an $\mathbb{F}_q$-linear $[n-1,k,d]$-code.

Proof. If $H_n \notin \cup_{\lambda \in \binom{1,\dots,n-1}{d-1}} \operatorname{Span}_{\mathbb{F}_q}(H_\lambda)$ and $c = (c_1,\dots,c_n) \in C$ is a word of weight $\operatorname{wt}(c) = d$, then $c_n = 0$ and $\operatorname{Supp}(c) \in \binom{1,\dots,n-1}{d}$. As a result, $\operatorname{wt}(\Pi_n(c)) = \operatorname{wt}(c) = d$ and $\Pi_n(C) \subset \mathbb{F}_q^{n-1}$ is of minimum distance d. Let us suppose that there is a non-zero $c \in \ker(\Pi_n) \cap C = \{(0^{n-1}, c_n) \in C\}$. Then $H_n = 0^n \in \cap_{\lambda \in \binom{1,\dots,n-1}{d-1}} \operatorname{Span}_{\mathbb{F}_q}(H_\lambda)$. The contradiction with the assumption (4.1) reveals that $\ker(\Pi_n) \cap C = \{0^n\}$ and $\dim_{\mathbb{F}_q} \Pi_n(C) = \dim_{\mathbb{F}_q}(C) = k$. □

Recall that a linear code $C \subset \mathbb{F}_q^n$ is non-degenerate if it is not contained in a coordinate hyperplane $V(x_i) = \{a \in \mathbb{F}_q^n \mid a_i = 0\}$ for some $1 \le i \le n$.

Proposition 4.2. *Let C be a non-degenerate $\mathbb{F}_q$-linear $[n,k,d]$-code of genus $g = n+1-k-d > 0$. Then there exist a finite extension $\mathbb{F}_{q^m} \supseteq \mathbb{F}_q$, a smooth irreducible affine variety $X/\mathbb{F}_{q^m} \subset \overline{\mathbb{F}_q}^n$, isomorphic to $\overline{\mathbb{F}_q}^k$ and a Zariski dense subset $S \subseteq X$, such that $0^n \in S$, $T_{0^n}(X, \mathbb{F}_{q^m}) = C \otimes_{\mathbb{F}_q} \mathbb{F}_{q^m}$, the puncturing $\Pi_n : X \to \Pi_n(X)$ at x_n is a finite morphism and the images*

$$(d\Pi_n)_a T_a(X, \mathbb{F}_{q^{\delta(a)}}) = T_{\Pi_n(a)}(\Pi_n(X), \mathbb{F}_{q^{\delta(a)}})$$

of the puncturings of $T_a(X, \mathbb{F}_{q^{\delta(a)}})$ at all the points $a \in S$ are $[n-1,k,d]$-codes.

Proof. Let $H' \in M_{(n-k)\times n}(\mathbb{F}_q)$ be a parity check matrix of C with columns H'_j for all $1 \le j \le n$. There is no loss in assuming that $H'_{k+1}, \dots, H'_n$ are linearly independent and form the identity matrix I_{n-k}. Any finite union of proper $\overline{\mathbb{F}_q}$-linear subspaces of the linear space $M_{(n-k)\times 1}(\overline{\mathbb{F}_q})$ over the infinite field $\overline{\mathbb{F}_q}$ has non-empty complement and there exists

$$c = \begin{pmatrix} c_1 \\ \vdots \\ c_{n-k} \end{pmatrix} \in M_{(n-k)\times 1}(\overline{\mathbb{F}_q}) \setminus \left\{ \left[\cup_{\lambda \in \binom{1,\dots,n-1}{d-1}} \operatorname{Span}_{\overline{\mathbb{F}_q}}(H'_\lambda) \right] \cup V(y_{n-k}) \right\}.$$

Let us denote by $\mathbb{F}_{q^m} := \mathbb{F}_q(c_1,\dots,c_{n-k})$ the definition field of c, put $p := \operatorname{char}\mathbb{F}_q$ for the characteristic of $\mathbb{F}_q$ and consider the affine variety $X := V(f_1,\dots,f_{n-k}) \subset \overline{\mathbb{F}_q}^n$, cut by the polynomials

$$f_i(x_1,\dots,x_k,x_{k+i},x_n) := \sum_{s=1}^{k} H'_{i,s} x_s + x_{k+i} + c_i x_n^{p+1} \quad \text{for } \forall 1 \le i \le n-k.$$

In order to construct a biregular morphism $X \to \overline{\mathbb{F}_q}^k$, note that $c_{n-k} \neq 0$ by the very choice of c and

$$X = V\left(f_i - \frac{c_i}{c_{n-k}} f_{n-k}, f_{n-k} \,\middle|\, 1 \leq i \leq n-k-1 \right).$$

The equations

$$f_i(x_1, \ldots, x_k, x_{k+i}, x_n) - \frac{c_i}{c_{n-k}} f_{n-k}(x_1, \ldots, x_k, x_n)$$
$$= \sum_{s=1}^{k} \left(H'_{i,s} - \frac{c_i}{c_{n-k}} H'_{n-k,s} \right) x_s + x_{k+i} - \frac{c_i}{c_{n-k}} x_n = 0$$

for $\forall 1 \leq i \leq n-k-1$ are equivalent to $x_{k+i} = \psi_{k+i}(x_1, \ldots, x_k, x_n)$ for

$$\psi_{k+i}(x_1, \ldots, x_k, x_n) := \sum_{s=1}^{k} \left(\frac{c_i}{c_{n-k}} H'_{n-k,s} - H'_{i,s} \right) x_s + \frac{c_i}{c_{n-k}} x_n,$$

$\forall 1 \leq i \leq n-k-1$. We claim the existence of $1 \leq s \leq k$ with $H'_{n-k,s} \neq 0$, since otherwise the last row of the parity check matrix H' of C is $(0^{n-1}, 1)$ and the non-degenerate code C is contained in the coordinate hyperplane with equation $x_n = 0$. Up to a permutation of the first k components of $\overline{\mathbb{F}_q}^n$, we assume that $H'_{n-k,k} \neq 0$. Then $f_{n-k}(x_1, \ldots, x_k, x_n) = 0$ is equivalent to $x_k = \psi_k(x_1, \ldots, x_{k-1}, x_n)$ for

$$\psi_k(x_1, \ldots, x_{k-1}, x_n) := -(H'_{n-k,k})^{-1} \left(\sum_{s=1}^{k-1} H'_{n-k,s} x_s + x_n + c_{n-k} x_n^{p+1} \right).$$

Thus, $X \subset \overline{\mathbb{F}_q}^n$ is cut by the equations

$$x_k - \psi_k(x_1, \ldots, x_{k-1}, x_n) = 0,$$
$$x_{k+i} - \psi_{k+i}(x_1, \ldots, x_{k-1}, \psi_k(x_1, \ldots, x_{k-1}, x_n), x_n) = 0 \quad \text{for} \quad \forall 1 \leq i \leq n-k-1$$

and the puncturing Π_α at $\alpha = \{k, k+1, \ldots, n-1\} \in \binom{1,\ldots,n}{n-k}$ provides a biregular morphism $\Pi_\alpha \colon X \to \overline{\mathbb{F}_q}^k$. In particular, X is a smooth irreducible affine variety, defined over $\mathbb{F}_{q^m}$. Note that the puncturing $\Pi_n \colon X \to \Pi_n(X)$ at x_n is a finite morphism, as far as the equation

$$f_{n-k}(x_1, \ldots, x_k, x_n) = \sum_{s=1}^{k} H'_{n-k,s} x_s + x_n + c_{n-k} x_n^{p+1} = 0$$

implies the algebraic dependence of the element $x_n + I(X, \overline{\mathbb{F}_q}) \in \overline{\mathbb{F}_q}(X)$ over the function field $\overline{\mathbb{F}_q}(\Pi_n(X)) = \overline{\mathbb{F}_q}(x_1 + I(X, \overline{\mathbb{F}_q}), \ldots, x_{n-1} + I(X, \overline{\mathbb{F}_q}))$.

For the rest of the proof, $T_a(X, \mathbb{F}_{q^{\delta(a)}})$, respectively, $T_{\Pi_n(a)}(\Pi_n(X), \mathbb{F}_{q^{\delta(a)}})$ are the Zariski tangent spaces over the definition fields $\mathbb{F}_{q^{\delta(a)}} := \mathbb{F}_{q^m}(a_1, \ldots, a_n)$ of $a \in X$ over $\mathbb{F}_{q^m}$. Note that

$$\frac{\partial(f_1, \ldots, f_{n-k})}{\partial x}(x_1, \ldots, x_n) = (H'_1 \ldots H'_{n-1} H_n(x_n)) = \frac{\partial f}{\partial x}(x_n) \qquad (4.2)$$

with $H_n(x_n) = H'_n + x_n^p c$ depends only on x_n. The columns of the Jacobian matrix $\frac{\partial(f_1,\dots,f_{n-k})}{\partial x}(x_n)$, labeled by $\beta = \{k+1,\dots,n\} \in \binom{1,\dots,n}{n-k}$ form the matrix

$$\frac{\partial(f_1,\dots,f_{n-k})}{\partial x_\beta}(x_n) = \begin{pmatrix} 1 & 0 & \cdots & 0 & c_1 x_n^p \\ 0 & 1 & \cdots & 0 & c_2 x_n^p \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & c_{n-k-1} x_n^p \\ 0 & 0 & \cdots & 0 & 1 + c_{n-k} x_n^p \end{pmatrix}$$

with determinant $\det \frac{\partial(f_1,\dots,f_{n-k})}{\partial x_\beta}(x_n) = 1 + c_{n-k}x_n^p$. Thus, at any point $a \in X \setminus V(c_{n-k}x_n^p + 1)$, the matrix $\frac{\partial(f_1,\dots,f_{n-k})}{\partial x}(a_n) \in M_{(n-k)\times n}(\mathbb{F}_{q^{\delta(a)}})$ is of rank $\operatorname{rk} \frac{\partial f}{\partial x}(a_n) = n-k$. According to

$$f_1,\dots,f_{n-k} \in I(X,\overline{\mathbb{F}_q}) = IV(f_1,\dots,f_{n-k}) = r(\langle f_1,\dots,f_{n-k}\rangle) \triangleleft \overline{\mathbb{F}_q}[x_1,\dots,x_n],$$

the Zariski tangent space $T_a(X,\mathbb{F}_{q^{\delta(a)}})$ at $a \in X$ is contained in the linear code $\mathcal{C}(a)$ with parity check matrix $\frac{\partial(f_1,\dots,f_{n-k})}{\partial x}(a)$. Since X is smooth, $\dim T_a(X,\mathbb{F}_{q^{\delta(a)}}) = \dim X = k$ at $\forall a \in X$ and $\frac{\partial(f_1,\dots,f_{n-k})}{\partial x}(a)$ is a parity check matrix of $T_a(X,\mathbb{F}_{q^{\delta(a)}})$ if and only if $\operatorname{rk}\frac{\partial(f_1,\dots,f_{n-k})}{\partial x}(a) = n-k$. In particular, $T_a(X,\mathbb{F}_{q^{\delta(a)}})$ has parity check matrix $\frac{\partial(f_1,\dots,f_{n-k})}{\partial x}(a)$ at all the points a of the non-empty, Zariski open, Zariski dense subset $X \setminus V(c_{n-k}x_n^p+1)$ of X. Note that $0^n \in X = V(f_1,\dots,f_{n-k})$ and $0^n \notin V(c_{n-k}x_n^p+1)$, so that $T_{0^n}(X,\mathbb{F}_{q^m})$ has parity check matrix $\frac{\partial(f_1,\dots,f_{n-k})}{\partial x}(0) = H'$ and $T_{0^n}(X,\mathbb{F}_{q^m}) = C \otimes_{\mathbb{F}_q} \mathbb{F}_{q^m}$.

Let $\Pi_n \colon \mathcal{C}(a) \to \Pi_n\mathcal{C}(a)$ be the puncturing at n and S_o be the set of those $a \in X$, at which $\Pi_n\mathcal{C}(a)$ is an $[n-1,k,d]$-code. By Lemma 4.1,

$$S_o \supseteq \left\{ a \in X \mid H_n(a_n) \notin \cup_{\lambda \in \binom{1,\dots,n-1}{d-1}} \operatorname{Span}_{\mathbb{F}_{q^{\delta(a)}}}(H'_\lambda) \right\},$$

whereas

$$Y := X \setminus S_o \subseteq Z := \left\{ a \in X \mid H_n(a_n) \in \cup_{\lambda \in \binom{1,\dots,n-1}{d-1}} \operatorname{Span}_{\overline{\mathbb{F}_q}}(H'_\lambda) \right\}.$$

We claim that Z is a proper Zariski closed subset of X. If so, then $X \setminus Z$ is a non-empty, Zariski open, Zariski dense subset of X and has non-empty, Zariski open, Zariski dense intersection $U := (X \setminus Z) \cap [X \setminus V(c_{n-k}x_n^p+1)]$ with the Zariski open subset $X \setminus V(c_{n-k}x_n^p+1) \neq \varnothing$ of the irreducible affine variety X. That suffices for $S' := S_o \cap [X \setminus V(c_{n-k}x_n^p+1)] \supseteq U$ to be a Zariski dense subset of X, containing 0^n and such that $(d\Pi_n)_a T_a(X,\mathbb{F}_{q^{\delta(a)}})$ are $[n-1,k,d]$-codes at all $a \in S'$.

Towards the study of Z, let

$$Z_\lambda := \{a \in X \mid H_n(a_n) \in \operatorname{Span}_{\overline{\mathbb{F}_q}}(H'_\lambda)\} = \{a \in X \mid \operatorname{rk}(H'_\lambda H_n(a_n)) < d\}$$

for $\lambda \in \binom{1,\dots,n-1}{d-1}$ and represent $Z = \cup_{\lambda \in \binom{1,\dots,n-1}{d-1}} Z_\lambda$. If $\mu \in \binom{1,\dots,n-k}{d}$ and

$$g_{\mu,\lambda}(x_n) := \det \frac{\partial f_\mu}{\partial(x_\lambda, x_n)}(x_n) \in \mathbb{F}_{q^m}[x_n]$$

is the determinant of the matrix

$$\frac{\partial f_\mu}{\partial(x_\lambda, x_n)}(x_n) = (H'_{\mu,\lambda} H'_{\mu,n}(x_n)) = (H'_{\mu,\lambda} H'_{\mu,n} + x_n^p c_{\mu,n})$$

formed by the rows of $(H'_\lambda H_n(x_n))$, labeled by $\mu \in \binom{1,\ldots,n-k}{d}$, then

$$Z_\lambda = X \cap V\left(g_{\mu,\lambda}(x_n) \,\middle|\, \forall \mu \in \binom{1,\ldots,n-k}{d} \right)$$

is a Zariski closed subset of X and, therefore, $Z = \cup_{\lambda \in \binom{1,\ldots,n-1}{d-1}} Z_\lambda$ is Zariski closed in X. The assumption

$$\cup_{\lambda \in \binom{1,\ldots,n-1}{d-1}} Z_\lambda = Z = X$$

for the irreducible affine variety X requires

$$X = Z_\lambda \subseteq V\left(g_{\mu,\lambda}(x_n) \,\middle|\, \forall \mu \in \binom{1,\ldots,n-k}{d} \right)$$

for some $\lambda \in \binom{1,\ldots,n-1}{d-1}$. Recall that the puncturing $\Pi_\alpha : X \longrightarrow \overline{\mathbb{F}_q}^k$ at the $(n-k)$-tuple $\alpha = \{k, k+1, \ldots, n-1\}$ is biregular and consider the sequence of affine varieties

$$\Pi_\alpha^{-1}(0^{k-1} \times \overline{\mathbb{F}_q}) \subseteq X \subseteq V\left(g_{\mu,\lambda}(x_n) \,\middle|\, \forall \mu \in \binom{1,\ldots,n-k}{d} \right),$$

where $0^{k-1} \times \overline{\mathbb{F}_q} = V(x_1, \ldots, x_{k-1}) \subset \overline{\mathbb{F}_q}^k$. Then $g_{\mu,\lambda}(x_n) \equiv 0$ for all $\mu \in \binom{1,\ldots,n-k}{d}$, which holds exactly when $\det(H'_{\mu,\lambda} H'_{\mu,n}) = 0$ and $\det(H'_{\mu,\lambda} c_\mu) = 0$ for all $\mu \in \binom{1,\ldots,n-k}{d}$. As a result, $\mathrm{rk}(H'_\lambda c) < d$ for $H'_\lambda \in M_{(n-k)\times(d-1)}(\mathbb{F}_q)$ of $\mathrm{rk}\, H'_\lambda = d-1$ and $c \in \mathrm{Span}_{\overline{\mathbb{F}_q}}(H'_\lambda)$. That contradicts the choice of c and shows that $Z \subsetneq X$ is a proper Zariski closed subset of X.

Note that the puncturing $\Pi_n \colon X \to \Pi_n(X)$ has injective differentials

$$(d\Pi_n)_a \colon T_a(X, \mathbb{F}_{q^{\delta(a)}}) \longrightarrow T_{\Pi_n(a)}(\Pi_n(X), \mathbb{F}_{q^{\delta(a)}}) \quad \text{at} \quad \forall a \in U,$$

so that the non-empty, Zariski open, Zariski dense subset $U \subseteq X$ is contained in the infinitesimally injective locus of Π_n, i.e., $U \subseteq \mathrm{Inf\,Inj}(\Pi_n)$. Intersecting U with the non-empty, Zariski open subset $\Pi_n^{-1}(\Pi_n(X)^{\mathrm{smooth}})$ of the irreducible affine variety X, one obtains a non-empty, Zariski open, Zariski dense subset $W := U \cap \Pi_n^{-1}(\Pi_n(X)^{\mathrm{smooth}}) \subseteq X$. Then

$$S := S' \cap \Pi_n^{-1}(\Pi_n(X)^{\mathrm{smooth}}) = S_o \cap [X \setminus V(c_{n-k} x_n^p + 1)] \cap \Pi_n^{-1}(\Pi_n(X)^{\mathrm{smooth}}) \supseteq W$$

is such a Zariski dense subset of X that

$$(d\Pi_n)_a T_a(X, \mathbb{F}_{q^{\delta(a)}}) = T_{\Pi_n(a)}(\Pi_n(X), \mathbb{F}_{q^{\delta(a)}})$$

are $[n-1, k, d]$-codes for all $a \in S$ according to Lemma 3.1 (ii). □

4.2. A family of dimension reductions of a linear code

The next proposition provides a family of dimension reductions of an $\mathbb{F}_q$-linear $[n,k,d]$-code C of genus $g = n+1-k-d > 0$, which is parameterized by a non-empty, Zariski open, Zariski dense subset of $\overline{\mathbb{F}_q}^{2(n-k)}$. The codes from the family are not tangent to a specific affine variety. We choose a parity check matrix of the original $[n,k,d]$-code C and project it on various hyperplanes in $\overline{\mathbb{F}_q}^{n-k}$, in order to obtain parity check matrices of $[n,k+1,d]$-codes over finite extensions of $\mathbb{F}_q$.

Proposition 4.3. *Let us suppose that C is an $\mathbb{F}_q$-linear $[n,k,d]$-code of genus $g = n+1-k-d > 0$. Then there exist a Zariski open, Zariski dense subset $\mathcal{W} \subset \overline{\mathbb{F}_q}^{2(n-k)}$ and a family $\mathcal{C} \to \mathcal{W}$ of $\mathbb{F}_{q^{\delta(u,v)}}$-linear $[n,k+1,d]$-codes $\mathcal{C}(u,v)$, containing C for any $(u,v) \in \mathcal{W}$, $u,v \in \overline{\mathbb{F}_q}^{n-k}$.*

Proof. Let $H = (H_1 \dots H_n) \in M_{(n-k)\times n}(\mathbb{F}_q)$ be a parity check matrix of C with columns $H_1,\dots,H_n \in \mathbb{F}_q^{n-k}$. For any $\lambda \in \binom{1,\dots,n}{d-1}$ let us consider $Z_\lambda := \mathrm{Span}_{\overline{\mathbb{F}_q}}(H_\lambda) \simeq \overline{\mathbb{F}_q}^{d-1}$ as an irreducible affine subvariety of $M_{(n-k)\times 1}(\overline{\mathbb{F}_q}) \simeq \overline{\mathbb{F}_q}^{n-k}$ and put

$$V(Q) := \left\{ (u,v) \in \overline{\mathbb{F}_q}^{n-k} \times \overline{\mathbb{F}_q}^{n-k} \mid Q(u,v) = \langle u,v \rangle = \sum_{s=1}^{n-k} u_s v_s = 0 \right\}$$

for the quadric in $\overline{\mathbb{F}_q}^{2(n-k)}$, given by the inner product in $\overline{\mathbb{F}_q}^{n-k}$. Observe that $Z_\lambda \times \overline{\mathbb{F}_q}^{n-k}$, $V(Q)$ and, therefore,

$$Z := V(Q) \cup \left(\cup_{\lambda \in \binom{1,\dots,n}{d-1}} Z_\lambda \times \overline{\mathbb{F}_q}^{n-k} \right)$$

are proper affine subvarieties of $\overline{\mathbb{F}_q}^{2(n-k)}$, due to the irreducibility of the affine space $\overline{\mathbb{F}_q}^{2(n-k)}$ and the assumption $g > 0$. Thus, $\mathcal{W} := \overline{\mathbb{F}_q}^{2(n-k)} \setminus Z$ is a non-empty, Zariski open, Zariski dense subset of $\overline{\mathbb{F}_q}^{2(n-k)}$. For any $(u,v) \in \mathcal{W}$ with $u,v \in \overline{\mathbb{F}_q}^{n-k}$, note that $u \notin \cup_{\lambda \in \binom{1,\dots,n}{d-1}} Z_\lambda = \cup_{\lambda \in \binom{1,\dots,n}{d-1}} \mathrm{Span}_{\overline{\mathbb{F}_q}}(H_\lambda)$ and

$$u \notin \mathcal{H}_v := \{ z \in \overline{\mathbb{F}_q}^{n-k} \mid \langle z,v \rangle = 0 \}$$

for the hyperplane $\mathcal{H}_v \subset \overline{\mathbb{F}_q}^{n-k}$ with gradient vector v. That allows to consider the $\overline{\mathbb{F}_q}$-linear maps

$$\mathcal{L}_{u,v} \colon \overline{\mathbb{F}_q}^{n-k} \longrightarrow \mathcal{H}_v,$$
$$\mathcal{L}_{u,v}(y) := y - \frac{\langle y,v \rangle}{\langle u,v \rangle} u \quad \text{for} \quad \forall y \in \overline{\mathbb{F}_q}^{n-k}, \quad \forall (u,v) \in \mathcal{W},$$

which project $\overline{\mathbb{F}_q}^{n-k}$ on $\mathcal{H}_v$, parallel to $\ker \mathcal{L}_{u,v} = \mathrm{Span}_{\overline{\mathbb{F}_q}}(u)$. Here we use that for any $z \in \mathcal{H}_v$ one has $\mathcal{L}_{u,v}(z) = z$. Let us consider the definition field $\mathbb{F}_{q^{\delta(u,v)}} =$

$\mathbb{F}_q(u_1,\ldots,u_{n-k},v_1,\ldots,v_{n-k})$ of $(u,v)\in\mathcal{W}$ over $\mathbb{F}_q$ and the matrix $H(u,v):=(\mathcal{L}_{u,v}(H_1)\ldots\mathcal{L}_{u,v}(H_n))\in M_{(n-k)\times n}(\mathbb{F}_{q^{\delta(u,v)}})$. The linear code $\mathcal{C}(u,v)$ with parity check matrix $H(u,v)$ contains C, as far as the $\overline{\mathbb{F}_q}$-linear map $\mathcal{L}_{u,v}$ transforms any non-trivial linear dependence $\sum\limits_{s=1}^{n}c_sH_s=0^{n-k}$ of the columns of H into a non-trivial linear dependence relation $\sum\limits_{s=1}^{n}c_s\mathcal{L}_{u,v}(H_s)=0^{n-k}$ of the columns of $H(u,v)$. In particular, $\mathcal{C}(u,v)$ contains words of weight d and the minimum distance $d(\mathcal{C}(u,v))\leq d$. If there is a non-zero word $a\in\mathcal{C}(u,v)\backslash\{0^n\}$ with $\mathrm{Supp}(a)\subseteq\lambda=\{\lambda_1,\ldots,\lambda_{d-1}\}\in\binom{1,\ldots,n}{d-1}$, then $0^n=\sum\limits_{s=1}^{d-1}a_{\lambda_s}\mathcal{L}_{u,v}(H_{\lambda_s})=\mathcal{L}_{u,v}\left(\sum\limits_{s=1}^{d-1}a_{\lambda_s}H_{\lambda_s}\right)$, whereas $\sum\limits_{s=1}^{d-1}a_{\lambda_s}H_{\lambda_s}\in\ker\mathcal{L}_{u,v}=\mathrm{Span}_{\overline{\mathbb{F}_q}}(u)$ and $\sum\limits_{s=1}^{d-1}a_{\lambda_s}H_{\lambda_s}=\lambda_0u$ for some $\lambda_0\in\overline{\mathbb{F}_q}$. According to $u\notin\mathrm{Span}_{\overline{\mathbb{F}_q}}(H_\lambda)$, there follow $\lambda_0=0$ and $\mathrm{rk}\,H_\lambda=\mathrm{rk}(H_{\lambda_1},\ldots,H_{\lambda_{d-1}})<d-1$. That contradicts the fact that C is of minimum distance d and shows that $\mathcal{C}(u,v)$ is of minimum distance $d(\mathcal{C}(u,v))=d$ for $\forall(u,v)\in\mathcal{W}$.

There remains to be checked that $\mathrm{rk}\,H(u,v)=n-k-1$ for all $(u,v)\in\mathcal{W}$, in order to derive that $\dim\mathcal{C}(u,v)=k+1$ and to conclude the proof of the proposition. To this end, note that $\mathcal{L}_{u,v}(H_s)\in\mathcal{H}_v$ for $\forall 1\leq s\leq n$, whereas $\mathrm{Span}_{\overline{\mathbb{F}_q}}(\mathcal{L}_{u,v}(H_1),\ldots,\mathcal{L}_{u,v}(H_n))\subseteq\mathcal{H}_v$ and $\mathrm{rk}\,H(u,v)\leq\dim_{\overline{\mathbb{F}_q}}\mathcal{H}_v=n-k-1$. On the other hand, $H_s=\mathcal{L}_{u,v}(H_s)+\frac{\langle H_s,v\rangle}{\langle u,v\rangle}u$ for all $1\leq s\leq n$ imply that

$$\overline{\mathbb{F}_q}^{n-k}=\mathrm{Span}_{\overline{\mathbb{F}_q}}(H_1,\ldots,H_n)\subseteq\mathrm{Span}_{\overline{\mathbb{F}_q}}(\mathcal{L}_{u,v}(H_1),\ldots,\mathcal{L}_{u,v}(H_n),u).$$

If $\mathrm{rk}\,H(u,v)\leq n-k-2$, then

$$n-k\leq\dim_{\overline{\mathbb{F}_q}}\mathrm{Span}_{\overline{\mathbb{F}_q}}(\mathcal{L}_{u,v}(H_1),\ldots,\mathcal{L}_{u,v}(H_n),u)\leq\mathrm{rk}\,H(u,v)+1\leq n-k-1$$

is an absurd, justifying $\mathrm{rk}\,H(u,v)=n-k-1$ and $\dim\mathcal{C}(u,v)=k+1$ for all $(u,v)\in\mathcal{W}:=\overline{\mathbb{F}_q}^{2(n-k)}\setminus\left[V(Q)\cup\left(\cup_{\lambda\in\binom{1,\ldots,n}{d-1}}Z_\lambda\times\overline{\mathbb{F}_q}^{n-k}\right)\right]$, $u,v\in\overline{\mathbb{F}_q}^{n-k}$. □

4.3. A family of weight reductions of a linear code

Let C be a linear $[n,k,d]$-code, which is not MDS. The next proposition establishes the existence of a family $\mathcal{C}\to U$ of $[n,k]$-codes $\mathcal{C}(a)$, $a\in U$ of minimum distance $\geq d+1$, parameterized by a non-empty, Zariski open, Zariski dense subset $U\subseteq\overline{\mathbb{F}_q}^n$. The codes from $\mathcal{C}$ are defined by a polynomial parity check matrix in n variables, but are not tangent to a specific affine subvariety of $\overline{\mathbb{F}_q}^n$.

Proposition 4.4. *Let us suppose that C is an $\mathbb{F}_q$-linear $[n,k,d]$-code of genus $g=n+1-k-d>0$. Then there exist a finite extension $\mathbb{F}_{q^m}\supseteq\mathbb{F}_q$, a non-empty, Zariski open, Zariski dense subset $U\subseteq\overline{\mathbb{F}_q}^n$ and a family $\mathcal{C}\to\overline{\mathbb{F}_q}^n$ of linear codes $\mathcal{C}(a)\subset\mathbb{F}^n_{q^{\delta(a)}}$ over the definition fields $\mathbb{F}_{q^{\delta(a)}}$ of $a\in\overline{\mathbb{F}_q}^n$ over $\mathbb{F}_{q^m}$, such that $\mathcal{C}(0^n)=C\otimes_{\mathbb{F}_q}\mathbb{F}_{q^m}$ and $\mathcal{C}(a)$ are of length n, dimension k and minimum distance $\geq d+1$ at all the points $a\in U$.*

Proof. Let $H' = (H'_1 \dots H'_n) \in M_{(n-k)\times n}(\mathbb{F}_q)$ be a parity check matrix of $C \subseteq \mathbb{F}_q^n$, whose first $n-k$ columns form a non-singular square matrix $(H'_1, \dots, H'_{n-k}) \in \mathrm{GL}(n-k, \mathbb{F}_q)$. By an induction on $d \leq j \leq n$, we choose appropriate $c_d, \dots, c_n \in M_{(n-k)\times 1}(\overline{\mathbb{F}_q})$, in order to set

$$H_j := H'_j \quad \text{for} \quad 1 \leq j \leq d-1,$$
$$H_j(x_j) := H'_j + x_j c_j \quad \text{for} \quad d \leq j \leq n$$

and to obtain a polynomial matrix

$$H(x_d, \dots, x_n) = (H'_1 \dots H'_{d-1} H_d(x_d) \dots H_n(x_n)) \in M_{(n-k)\times n}(\overline{\mathbb{F}_q}[x_d, \dots, x_n]).$$

Let $\mathbb{F}_{q^m} = \mathbb{F}_q(c_{ij} \mid 1 \leq i \leq n-k,\ d \leq j \leq n)$ be the common definition field of all the entries of $c_d, \dots, c_n$ over $\mathbb{F}_q$. At any point $a \in \overline{\mathbb{F}_q}^n$, define $\mathcal{C}(a)$ to be the linear code over the definition field $\mathbb{F}_{q^{\delta(a)}} = \mathbb{F}_{q^m}(a_1, \dots, a_n)$ of a over $\mathbb{F}_{q^m}$, with parity check matrix $H(a) = H(a_d, \dots, a_n) \in M_{(n-k)\times n}(\mathbb{F}_{q^{\delta(a)}})$. Our choice of $H(x_d, \dots, x_n)$ is such that $H(0^n) = H'$, whereas $\mathcal{C}(0^n) = C \times_{\mathbb{F}_q} \mathbb{F}_{q^m}$. It suffices to show the existence of non-empty, Zariski open, Zariski dense subsets $U' \subseteq \overline{\mathbb{F}_q}^n$, $U'' \subseteq \overline{\mathbb{F}_q}^n$, such that $\mathcal{C}(a)$ are of minimum distance $\geq d+1$ at all $a \in U'$ and $\mathcal{C}(b)$ are of dimension k at all $b \in U''$. Then $U := U' \cap U'' \subseteq \overline{\mathbb{F}_q}^n$ is a non-empty, Zariski open, Zariski dense subset, over which the codes $\mathcal{C}(a)$, $a \in U$ are of length n, dimension k and minimum distance $\geq d+1$. Regardless of the choice of $c_d, \dots, c_n \in M_{(n-k)\times 1}(\overline{\mathbb{F}_q})$, let $\gamma := \{1, \dots, n-k\}$ and note that

$$U'' := \overline{\mathbb{F}_q}^n \setminus V\left(\det H_\gamma(x_d, \dots, x_{n-k})\right)$$

is a Zariski open subset of $\overline{\mathbb{F}_q}^n$ with $\dim \mathcal{C}(b) = k$ at all $b \in U''$. Since $0^n \in U''$, the set U'' is non-empty and, therefore, Zariski dense in $\overline{\mathbb{F}_q}^n$.

By an induction on $d \leq j \leq n$, we choose $c_j \in M_{(n-k)\times 1}(\overline{\mathbb{F}_q})$ and show the existence of a non-empty, Zariski open, Zariski dense subset $U_j \subseteq \overline{\mathbb{F}_q}^j$ with $\mathrm{rk}\, H_\beta(u) = d$ for all $\beta \in \binom{1,\dots,j}{d}$ and all $u \in U_j$. Then $U' := U_n$ will be a non-empty, Zariski open, Zariski dense subset of $\overline{\mathbb{F}_q}^n$, such that $\mathcal{C}(a)$ is of minimum distance $\geq d+1$ at all $a \in U'$. To this end, let $j = d$, $\lambda := \{1, \dots, d-1\}$ and note that $\mathrm{Span}_{\overline{\mathbb{F}_q}}(H'_\lambda) \simeq \overline{\mathbb{F}_q}^{d-1}$ is a proper subspace of $M_{(n-k)\times 1}(\overline{\mathbb{F}_q}) \simeq \overline{\mathbb{F}_q}^{n-k}$, according to $g > 0$. That allows to choose

$$c_d \in M_{(n-k)\times 1}(\overline{\mathbb{F}_q}) \setminus \mathrm{Span}_{\overline{\mathbb{F}_q}}(H'_\lambda)$$

and to put $H_d(x_d) := H'_d + x_d c_d$. The family $\{H_d(a_d)\}_{a_d \in \overline{\mathbb{F}_q}}$ of columns is claimed to have at most one common entry $H_d(\kappa_d)$ with $\mathrm{Span}_{\overline{\mathbb{F}_q}}(H'_\lambda)$, so that $\mathrm{rk}\, H_{\lambda \cup \{d\}}(x_d) = d$ at all the points of the non-empty, Zariski open, Zariski dense subset $U_d := \overline{\mathbb{F}_q}^{d-1} \times (\overline{\mathbb{F}_q} \setminus \{\kappa_d\})$ of $\overline{\mathbb{F}_q}^d$. Indeed, if $H_d(x_d) \notin \mathrm{Span}_{\overline{\mathbb{F}_q}}(H'_\lambda)$ for all $x_d \in \overline{\mathbb{F}_q}$, there is nothing to be proved. In the case of $H_d(\kappa_d) \in \mathrm{Span}_{\overline{\mathbb{F}_q}}(H'_\lambda)$ for some $\kappa_d \in \overline{\mathbb{F}_q}$, let us move the origin of $M_{(n-k)\times 1}(\overline{\mathbb{F}_q})$ at $H_d(\kappa_d) \in M_{(n-k)\times 1}(\overline{\mathbb{F}_q})$. The 1-dimensional linear subspace $H_d(x_d)$ of the $(n-k)$-dimensional space $M_{(n-k)\times 1}(\overline{\mathbb{F}_q})$ intersects the $(d-1)$-dimensional linear subspace $\mathrm{Span}_{\overline{\mathbb{F}_q}}(H'_\lambda)$ in more than one point if and only if it is contained in $\mathrm{Span}_{\overline{\mathbb{F}_q}}(H'_\lambda)$. Then for arbitrary $x_d \neq y_d$ from $\overline{\mathbb{F}_q}$, one has $(x_d - y_d)c_d \in \mathrm{Span}_{\overline{\mathbb{F}_q}}(H'_\lambda)$, contrary to the choice of $c_d \notin \mathrm{Span}_{\overline{\mathbb{F}_q}}(H'_\lambda)$. That provides the base of the induction.

Suppose that $d+1 \leq j \leq n$ and $c_d, \dots, c_{j-1} \in M_{(n-k)\times 1}(\overline{\mathbb{F}_q})$ have been chosen in such a way that there exists a non-empty, Zariski open, Zariski dense subset $U_{j-1} \subseteq \overline{\mathbb{F}_q}^{j-1}$ with $\mathrm{rk}\, H_\beta(u) = d$ for all $\beta \in \binom{1,\dots,j-1}{d}$ and all $u \in U_{j-1}$. Fix an arbitrary $u \in U_{j-1}$ and choose

$$c_j \in M_{(n-k)\times 1}(\overline{\mathbb{F}_q}) \setminus \left[\cup_{\lambda \in \binom{1,\dots,j-1}{d-1}} \mathrm{Span}_{\overline{\mathbb{F}_q}}(H_\lambda(u))\right]. \tag{4.3}$$

The existence of c_j is due to the fact that $\cup_{\lambda \in \binom{1,\dots,j-1}{d-1}} \mathrm{Span}_{\overline{\mathbb{F}_q}}(H_\lambda(u))$ is a finite union of proper subspaces $\mathrm{Span}_{\overline{\mathbb{F}_q}}(H_\lambda(u)) \simeq \overline{\mathbb{F}_q}^{d-1}$ of the linear space $M_{(n-k)\times 1}(\overline{\mathbb{F}_q}) \simeq \overline{\mathbb{F}_q}^{n-k}$ over the infinite field $\overline{\mathbb{F}_q}$. We claim that

$$W_{j-1} := \{w \in U_{j-1} \mid c_j \notin \cup_{\lambda \in \binom{1,\dots,j-1}{d-1}} \mathrm{Span}_{\overline{\mathbb{F}_q}}(H_\lambda(w))\}$$

is a Zariski open subset of U_{j-1}. Indeed,

$$\begin{aligned} U_{j-1} \setminus W_{j-1} &= \cup_{\lambda \in \binom{1,\dots,j-1}{d-1}} \{t \in U_{j-1} \mid c_j \in \mathrm{Span}_{\overline{\mathbb{F}_q}}(H_\lambda(t))\} \\ &= \cup_{\lambda \in \binom{1,\dots,j-1}{d-1}} \{t \in U_{j-1} \mid \mathrm{rk}(H_\lambda(t)c_j) = d-1\}, \end{aligned}$$

as far as $\mathrm{rk}\, H_\beta(u) = d$ for all $\beta \in \binom{1,\dots,j-1}{d}$ and all $u \in U_{j-1}$ implies $\mathrm{rk}\, H_\lambda(t) = d-1$ for all $\lambda \in \binom{1,\dots,j-1}{d-1}$ and all $t \in U_{j-1}$. Denoting by Σ_d^{d-1} the set of the maps $\mu\colon \binom{1,\dots,j-1}{d-1} \to \binom{1,\dots,n-k}{d}$ and putting

$$Y_j := V\left(\prod_{\lambda \in \binom{1,\dots,j-1}{d-1}} \det(H_{\mu(\lambda),\lambda} c_{\mu(\lambda),j}) \,\middle|\, \forall \mu \in \Sigma_d^{d-1}\right),$$

one concludes that

$$\begin{aligned} U_{j-1} \setminus W_{j-1} &= \cup_{\lambda \in \binom{1,\dots,j-1}{d-1}} \left\{t \in U_{j-1} \mid \det(H_{\nu,\lambda}(t)c_{\nu,j}) = 0,\ \forall \nu \in \binom{1,\dots,n-k}{d}\right\} \\ &= \cup_{\lambda \in \binom{1,\dots,j-1}{d-1}} \left[U_{j-1} \cap V\left(\det(H_{\nu,\lambda}c_{\nu,j}) \mid \forall \nu \in \binom{1,\dots,n-k}{d}\right)\right] \\ &= U_{j-1} \cap Y_j \end{aligned}$$

is a Zariski closed subset of U_{j-1}, so that $W_{j-1} = U_{j-1} \setminus Y_j$ is Zariski open in U_{j-1}. According to $u \in W_{j-1}$ for the point $u \in U_{j-1}$, used in the choice (4.3) of c_j, $W_{j-1} \neq \varnothing$ is non-empty and, therefore, Zariski dense in $\overline{\mathbb{F}_q}^{j-1}$. Note that

$$\begin{aligned} U_j &:= \left\{(w, w_j) \in \overline{\mathbb{F}_q}^{j-1} \times \overline{\mathbb{F}_q} \mid \mathrm{rk}\, H_\beta(w, w_j) = d, \quad \forall \beta \in \binom{1,\dots,j}{d}\right\} \\ &= \left\{(w, w_j) \in W_{j-1} \times \overline{\mathbb{F}_q} \mid \mathrm{rk}(H_\lambda(w)H_j(w_j)) = d, \quad \forall \lambda \in \binom{1,\dots,j-1}{d-1}\right\} \end{aligned}$$

has complement

$$\begin{aligned}(W_{j-1}\times\overline{\mathbb{F}_q})\setminus U_j &= \cup_{\lambda\in\binom{1,\dots,j-1}{d-1}}\left\{(w,w_j)\in W_{j-1}\times\overline{\mathbb{F}_q}\mid \mathrm{rk}(H_\lambda(w)H_j(w_j))<d\right\}\\ &= \cup_{\lambda\in\binom{1,\dots,j-1}{d-1}}\left\{(w,w_j)\in W_{j-1}\times\overline{\mathbb{F}_q}\mid h_{\nu,\lambda}(w,w_j)=0,\ \forall\nu\in\binom{1,\dots,n-k}{d}\right\},\end{aligned}$$

where $h_{\nu,\lambda}(x_d,\dots,x_j):=\det(H_{\nu,\lambda}(x_d,\dots,x_{j-1})H_{\nu,j}(x_j))\in\overline{\mathbb{F}_q}[x_d,\dots,x_j]$. If

$$Z_j := V\left(\prod_{\lambda\in\binom{1,\dots,j-1}{d-1}} h_{\mu(\lambda),\lambda}(x_d,\dots,x_j)\,\middle|\,\forall\mu\in\Sigma_d^{d-1}\right),$$

then

$$\begin{aligned}(W_{j-1}\times\overline{\mathbb{F}_q})\setminus U_j &= (W_{j-1}\times\overline{\mathbb{F}_q})\cap\left[\cup_{\lambda\in\binom{1,\dots,j-1}{d-1}}V\left(h_{\nu,\lambda}\mid\forall\nu\in\binom{1,\dots,n-k}{d}\right)\right]\\ &= (W_{j-1}\times\overline{\mathbb{F}_q})\cap Z_j\end{aligned}$$

is Zariski closed in $W_{j-1}\times\overline{\mathbb{F}_q}$, so that $U_j=(W_{j-1}\times\overline{\mathbb{F}_q})\setminus Z_j$ is Zariski open in $W_{j-1}\times\overline{\mathbb{F}_q}$ and in $\overline{\mathbb{F}_q}^j$. The assumption $U_j=\varnothing$ implies $W_{j-1}\times\overline{\mathbb{F}_q}\subseteq Z_j$ and holds exactly when

$$\begin{aligned}h_{\mu(\lambda),\lambda} &= \det(H_{\mu(\lambda),\lambda}(x_d,\dots,x_{j-1})H'_{\mu(\lambda)j}+x_jc_{\mu(\lambda)j})\\ &= \det(H_{\mu(\lambda),\lambda}(x_d,\dots,x_{j-1})H'_{\mu(\lambda)j})+x_j\det(H_{\mu(\lambda),\lambda}(x_d,\dots,x_{j-1})c_{\mu(\lambda)j})\end{aligned}$$

is independent of x_j for all $\lambda\in\binom{1,\dots,j-1}{d-1}$ and all $\mu\colon\binom{1,\dots,j-1}{d-1}\to\binom{1,\dots,n-k}{d}$. That, in turn, is equivalent to $\det(H_{\nu,\lambda}(x_d,\dots,x_{j-1})c_{\nu j})=0$ for all $\nu\in\binom{1,\dots,n-k}{d}$ and all $\lambda\in\binom{1,\dots,j-1}{d-1}$ and specializes to $\det(H_{\nu,\lambda}(u)c_{\nu j})=0$ at the point $u\in U_{j-1}$, used in the choice (4.3) of c_j. As a result, $\mathrm{rk}(H_\lambda(u)c_j)<d$ for all $\lambda\in\binom{1,\dots,j-1}{d-1}$. The inductive hypothesis $\mathrm{rk}\,H_\beta(u)=d$ for all $\beta\in\binom{1,\dots,j-1}{d}$ requires $\mathrm{rk}\,H_\lambda(u)=d-1$ for all $\lambda\in\binom{1,\dots,j-1}{d-1}$ and $\mathrm{rk}(H_\lambda(u)c_j)<d$ is equivalent to $c_j\in\mathrm{Span}_{\overline{\mathbb{F}_q}}(H_\lambda(u))$ for all $\lambda\in\binom{1,\dots,j-1}{d-1}$. That contradicts the choice (4.3) of c_j and shows that $U_j\neq\varnothing$ is a non-empty, Zariski open, Zariski dense subset of $\overline{\mathbb{F}_q}^j$. □

Acknowledgements. The first author is partially supported by Contract 80-10-99/26.04.2023 of the Science Foundation of Sofia University "St. Kliment Ohridski".

References

[1] M. C. Beltrametti, E. Carletti, D. Gallarati, and G. M. Bragadin, Lectures on curves, surfaces and projective varieties (A classical view of algebraic geometry), European Mathematical Society Textbooks, Zürich, 2009.

[2] R. Blahut, Algebraic codes for data transmission, Cambridge University Press, 2003.

[3] I. Duursma, Weight distribution of geometric Goppa codes, Trans. Amer. Math. Soc. 351 (1999) 3609–3639.

[4] J. Harris, Algebraic geometry – A first course, Graduate Texts in Mathematics, Springer, 1992.

[5] W. C. Huffman, V. Pless, Fundamentals of error correcting codes, Cambridge University Press, 2003.

[6] A. Kasparian and I. Marinov, Duursma's reduced polynomial, Adv. Math. Commun. 11(4) (2017) 647–669.

[7] K. O'Grady, A first course in algebraic geometry, 2012.

[8] R. Pellikaan, On the efficient decoding of algebraic-geometric codes, in: Eurocode 92, ed. by P. Camion, P. Charpin and S. Harari, Udine, CISM Courses and Lectures 339, Springer, Wien, 1993, 231–253.

[9] M. Reid, Undergraduate Algebraic Geometry, London Mathematical Society Student Texts, 1989.

[10] I. R. Shafarevich, Basic Algebraic Geometry, v.1, 2, Nauka, Moscow, 1988 (in Russian).

[11] M. A. Tsfasman, S. G. Vlădut, and T. Zink, Modular curves, Shimura curves, and Goppa codes, better than Varshamov-Gilbert bound, Math. Nachr. 109 (1982) 21–28.

Received on September 18, 2023
Accepted on February 13, 2024

Azniv K. Kasparian and Evgeniya D. Velikova

Faculty of Mathematics and Informatics
Sofia University "St. Kliment Ohridski"
5, James Bourchier Blvd.
1164 Sofia
BULGARIA

E-mails: `kasparia@fmi.uni-fofia.bg`
`velikova@fmi.uni-sofia.bg`

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“
ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Том 111

ANNUAL OF SOFIA UNIVERSITY “ST. KLIMENT OHRIDSKI”
FACULTY OF MATHEMATICS AND INFORMATICS
Volume 111

SOCIAL MEDIA DATA MINING FOR EXPLORING READERS' LITERARY INTERESTS

INESSA KEVLER, KARINA AGAFONOVA, KRASSIMIRA IVANOVA and MILENA DOBREVA

Digital humanities have become a booming field of study thanks to the application of computer and information science methods to the humanities and the accumulation of large-scale digital resources. The use of data mining techniques and quantitative methods is expanding and starting to dominate the domain of literary studies, focusing not only on the study of the text itself but also on the study of social influences and preferences of the readers. By analysing social network data from VK, the most popular Russian platform, the study investigates which authors and literary works resonate most with users. To analyse the quotations, we created a framework which combines tools for extracting and cleaning data, named entity recognition and finally, corpus analysis; this framework can be applied to other datasets to compare and expand our findings. We also noted that within the information behaviour research domain, such studies of sharing among communities are not yet very popular. The study showcases the importance of employing advanced computational tools in humanities research. The obtained quantitative results are subjected to a critical analysis, which can serve as a basis for humanitarian understanding based on data. The compelling findings of this pilot study confirm opportunities for further research using more advanced quantitative methodologies, as well as broadening the scope of criteria that potentially influence the formation of reading preferences, such as level of education, pop-culture trends, social environment, etc. Additionally, it sheds light on the underexplored domain of community building in social media networks, calling for broader research into information behaviour and online community dynamics.

Keywords: text analysis, data mining, quantitative analysis, digital humanities, literary studies

2020 Mathematics Subject Classification: 91D30

CCS Concepts:

• Information systems∼Information systems applications∼Data mining∼Clustering

DOI: 10.60063/GSU.FMI.111.115-127

1. Introduction

The synergy of sciences and humanities enlarged over the last decades within the new field of "digital humanities", which initially started as "humanities computing". Digital humanities are combining methodologies from humanities and the sciences, opening up new possibilities for data mining, analysis, and visualization of large-scale primary and secondary data. As Kathleen Fitzpatrick, a Director of Digital Humanities and Professor of English at Michigan State University argued, digital humanities could be understood as "a nexus of fields within which scholars use computing technologies to investigate the kinds of questions that are traditional to the humanities, or, as is more true of my own work, ask traditional kinds of humanities-oriented questions about computing technologies" [4].

Data mining quickly became popular within the field of digital humanities due to several key factors. On the one hand, the digital age has created an unprecedented volume of data in various forms and sources. This deluge of unstructured but content-rich data requires advanced techniques such as data mining to make sense of it and extract meaningful insights. In parallel, advances in computer hardware and processing capabilities have enabled more efficient handling and analysis of massive data sets, thus making it possible to apply data mining techniques to digital humanities research. Sifting through this wealth of data through advanced knowledge-mining techniques has created opportunities to uncover hidden patterns and gain new perspectives on historical and cultural phenomena [12]. The advances in natural language processing (NLP) became particularly important alongside machine learning techniques that have made it possible to analyse and interpret textual data at scale, expanding the scope and depth of digital humanities research on textual sources. It can be argued that the interdisciplinary efforts of computer scientists, data analysts, and humanities researchers are critical to the rapid adoption of data mining in the digital humanities. They also contribute to the growing recognition of the value of data-driven insights for understanding and preserving human cultural and historical heritage.

An important branch in this regard is data mining from the vast amount of data generated on social media platforms. The process involves using specialised algorithms and techniques to sift through user-generated content, interactions, and behaviour on platforms, such as Facebook, Twitter, Instagram, etc., and extracting hidden patterns and insights employed for market research, sentiment analysis, trend identification, and personalised content recommendations. As such, social media data mining is an extremely valuable tool for analysis.

Within our study, we focus on the intersection between data mining, literary studies, and social media research. Literary studies are not limited to close reading and comprehension of a literary text. Literature as a cultural artefact is surrounded by a much larger range of research issues from the positions of various social and humanitarian sciences. For instance, the sociology of literature deals with the issues of the social influence of literature, and the philosophy of literature focuses on the ontological, epistemological, axiological, and ethical properties of literature. From the second half of the 19th century, literary studies began to interact closely with the sciences, and mathematics in particular. Researchers began to use quantitative methods to determine the frequency of words to attribute authorship to anonymous or disputed texts. Since then, the possibilities of quantitative methods have become wider, and they have more and more supporters. Nowadays, the application of quantitative methods is expanding and conquering literary studies, focusing not only on the study of the text but also on the study of social, philosophical, and other aspects of literature.

Our study focuses on a combination of the following research questions:

RQ1. Is the geographic coverage of a social media community consistent with a preference for authors from the same geographic region?

RQ2. Can we assume that posts containing poems will be the most popular among quotations from different literary genres? (we can assume poetry will be popular due to the ease of remembering poetry quotations and the popularity of the genre).

The first research questions are related to literary studies. The intention of exploring these questions is to investigate to what extent social media data can provide insights into literary perceptions and preferences. Applying the data collection and analysis methods while seeking the answers to these questions is also aimed at answering the following questions:

RQ3. Can a common framework for analysing social media data be developed to support various literary studies?

RQ4. Can this type of analysis be supported with data from other types of studies?

We have done a pilot study of the trends of readers' interests based on analysis of posts on social networks, which was presented and defended as the final project on the course "Cultural Data Mining" within the master's program in Digital Humanities at ITMO University in the 2023 year.

Within the study, we analysed social media data containing literary quotations by world-famous authors. This analysis allows us to determine which of the authors are most popular among the audience of the social network, as well as which of their works are most popular among the readers. We have chosen VK for our research, which is the most popular social media platform in Russia, according to Statista [11]. VK has multiple public pages, including numerous pages dedicated to literature. One of the most popular types of content within the literature pages on the social network is quotations from a literary text, usually mentioning the author and the title of the text. These quotations can be extracted into data sets which are unique in capturing trends in the readers' preferences.

The article is organised as follows. Section 2 provides a brief overview of recent relevant work. Section 3 presents the methodology of the study. Section 4 presents the dataset extraction and preprocessing steps. Section 5 shows the achieved results of statistical analysis based on social network data and text analysis of the quotations. Finally, Section 6 offers some conclusions and discusses potential further research.

2. Related work

Social media can play multiple functions. Aichner et al. reviewed the definitions of social media in their 25 years of existence, exploring some 60,000 articles including definitions of social media [1]. While there is no agreement on the actual definition of social media, the various roles they play are summarised and include socialising with friends and family, romance and flirting, job seeking, interacting with companies and brands, and doing business. We can argue that today's social media grew as place for identity building (be it of users, especially influencers, brands, institutions or communities), for *marketing and promotion*, for finding new contacts and *growing communities*, and for *keeping in touch*. In the last years, particularly extensive research has explored a negative side of social media, the spread of disinformation and misinformation.

Within this article, our main interest is in the community function of social media combined with the identity-building one, especially in relation to communities which bring together users interested in literature who frequently share quotations from literature works.

Surprisingly, there is very little research on the information behaviour of communities with a literary interest. For example, Vlieghe et al. [15] explored in 2016 the reader engagement in social media environments. They proposed the following classification of the type of engagement: a place to share reading experiences, a place to meet enthusiasts, a place to create identity, a place to acknowledge and encourage participation. They also introduced the concept of "affinity spaces to develop an understanding of how readers engage in a variety of literary practices in social media environments that focus on literature and reading". Their study used the ethnographic method combined with interviews of a number of people engaged in the social media explored.

Mas-Baeda et al. [9] studied the social media posts of Spanish publishers on Twitter (the article precedes the rebranding of the platform as X). This research does not focus on community building but rather explores the marketing use of social media by book publishers.

There are a number of studies which explored social media use for specific domains and groups of users outside of our domain of focus but may be of interest as a methodological approach and for comparison of issues within other communities. For example, Bocala-Wiedemann [2] explored the target group of young adults within the context of evangelistic social media content and identified the most effective platforms, namely Instagram, TikTok, and YouTube (2022). Vanherle et al. [13] explored an interesting aspect of social media – how information shared in a particular group may answer specific needs or topics (e.g., adolescents sharing alcohol references).

A more in-depth study of quoting practices in social media related to journalists' use has been offered in [7]. Studies focusing on specific professional use of social media are helpful and there is a clear need for similar studies of other groups.

It is also worth noting, that while sharing and community building on social media are still being researched, a recent publication argues that this role of social media is fading away. John [8] analysed the home pages of 61 social networking sites in the decade 2011 and 2020, and argues that "'sharing' has lost its central place in the terminology employed by social media platforms in their self-presentation".

In conclusion, we see a growing number of research publications reporting on the behaviour of specific communities on social media. However, in relation to literature, the studies are limited. There is interest in the publishers' use of social media and some studies of specific communities based on ethnographic and survey methods. Generally, quoting practices on social media have not yet been studied in depth, and our study aims to bring additional evidence to the scholarly discourse.

3. Methodology

Within our study, we used quantitative methodologies and natural language processing techniques [6] on datasets obtained from specialised VK social media pages. Our approach adhered to the seminal framework elucidated by Fayyad et al. [3] during the nascent stages of knowledge discovery.

As a starting point of our research, we put forward two hypotheses that we will test in this study. First, *we assume that native readers will prefer native authors* who describe the culture most understandable to readers. Secondly, *we assume that posts containing poems will be the most popular* since the preparation of such content requires less effort from administrators of niche pages in the social network.

It is worth noting that our research does not set itself the task of making any conclusions on the most popular authors in Russia. We pursued the following goals:

- To determine whose texts of world literature authors are most often quoted by niche pages, as well as which of them are gaining more reactions from users;
- To term the topics of the most quoted texts;
- To detect tendencies in acquired data using statistical analysis.

Before starting the study, we had to scope the best sources of data. After a manual analysis of content across different social media, we decided to use VK and identified three groups which discuss literature-related topics. We are aware that there are potentially multiple sources for such content also on other social media, but our aim was to provide a proof of concept on the viability of our methodology rather than produce a comprehensive analysis of all relevant social media content.

At first, data was collected from the niche pages in social media using VK API (Application Programming Interface): the Implicit Flow method was applied to obtain the user's access key, after which the wall.get method was applied to parse data from niche pages. Although the VK API is open, it has limitations, in particular in the amount of data available for uploading, but we managed to resolve this with the help of cyclic code.

Then, the named-entity recognition was made using Natasha[1] (a Python library for the Russian language) and the data was cleaned from irrelevant names. The data was cleaned up, primarily because we used the Natasha library to upload all proper names, and some of these names did not belong to the authors, but to characters. Such data needed additional manual processing.

We detected the most mentioned and shared literary authors. Moreover, we determined the topics of the literary texts and quotations from which users shared. To do this, we used Voyant Tools[2], having previously excluded prepositions, conjunctions, pronouns, interjections, and other noise from the textual data.

Finally, we have come to conclusions about the importance of world literature for Russian-native readers.

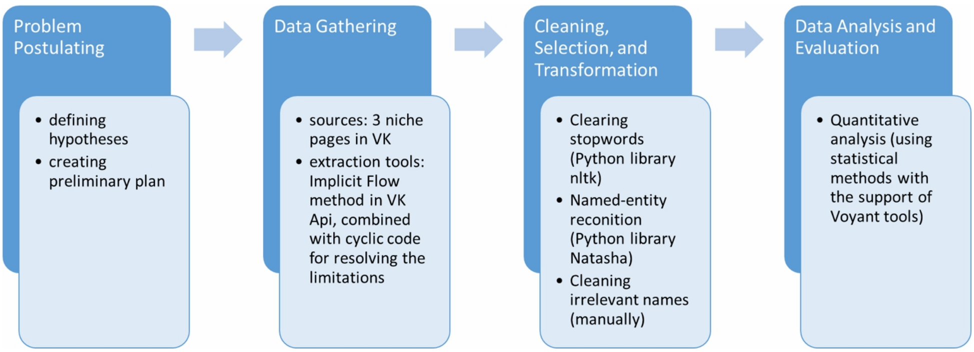


Figure 1. The process framework

4. Data sets and preprocessing

We decided to look for sources among niche pages in VK due to its open VK API [14]. When choosing sources, we preferred popular niche pages with the most subscribers, also meeting the following criteria:

- Published quotations are accompanied, at least, by a mention of the author (at most, by a mention of a precedent text);
- The content is published primarily in text format (we did not choose niche pages that published quotations exclusively as images, because it would require the use of other methods).

Finally, as a source of data, we have used the following resources:

[1] GitHub – natasha/natasha: Solves basic Russian NLP tasks, API for lower level Natasha projects, https://github.com/natasha/natasha

[2] Voyant Tools, https://voyant-tools.org/.

- "Стихи давно забытого поэта" (Poems of a long-forgotten poet, https://vk.com/memorypoets)
- "Литература" (Literature, https://vk.com/litratre)
- "Сочная литература" (Juicy literature, https://vk.com/juicylit)

It is worth noting that even though the name of the first source hints at exclusively poetic content, it contains a large number of quotations from prose texts, therefore it is relevant to our research goals.

VK API has technical limitations and makes it possible to collect only the last 100 posts. However, we created a code loop that allowed us to parse 10,000 posts from the first two resources and 1000 posts from the third resource (because this resource is smaller than the others) by the following categories: text, likes, comments, reposts.

From the textual data were recognized names of literary authors. The final data sets included authors mentioned more than once and were not subscribers of the source (who published their texts in the group on the rights of advertising).

5. Data analysis results

We performed two types of quantitative analyses on the already pre-processed data set, aiming to interpret the obtained results from the point of view of the historical events in the creation of the works and the present. The first is aimed at highlighting the most frequently cited authors, and the second one analyses which passages from which works are most often the object of interest on the part of the virtual community.

5.1. Analysis of data extracted from the social media posts

According to the data we have collected, the most frequently mentioned author, that is, the most cited on niche pages, is Erich Maria Remarque, the German-born writer – 323 posts are dedicated to him. Thus, both our hypotheses, are not supported by the data evidence.

Nevertheless, the second and third most popular authors in terms of the number of posts with quotations from their texts are just Russian poets: 243 posts contain full text or fragments from poems by Sergei Esenin, the Russian lyric poet, and 239 posts – from poems by Joseph Brodsky, the Russian-American poet. The most published poem by Sergei Esenin is "Tips of blue flame are dancing about . . . ": it was posted 14 times. Brodsky's poems do not fundamentally differ in the frequency of publications, which is somewhat astonishing since the legendary poem "Don't leave the room, don't make a mistake" has already found a focal embodiment in memes and has become an independent cultural artefact.

Furthermore, as a result of the data analysis, we found a quotation that is mistakenly attributed to Brodsky, as is commonly believed: "When you are tired of endless introspection, call me. Let's dance." According to an alternative opinion, it belongs to a certain poetess Henrietta Lardan, but there is no reliable evidence of this. However, this is the most reposted line, signed in social media with Brodsky's name: 817 people shared this post.

Other significant authors in terms of the number of mentions are Marina Tsvetaeva, one of the most influential Russian women poets (153 quotations in posts), along with Anna Akhmatova (85 quotations in posts).

All the Russian authors mentioned at the moment worked in the 20th century and belong to the Silver Age of Russian Poetry.

Turning to the prose writers, Mikhail Bulgakov is the most mentioned, and the most cited novel is "The Master and Margarita": among 101 quotations from Bulgakov's texts, 71 are quotations from this novel.

Returning to the conversation about Remarque, we have analysed the most cited novels. According to the data, we can distinguish two of the most frequently cited: "Arch of Triumph" and "Three Comrades", 70 and 66 citations, respectively (Figure 2).

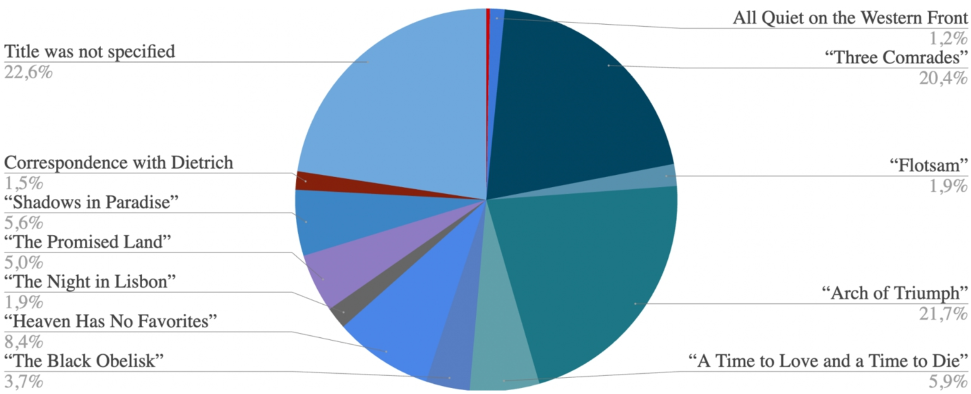


Figure 2. Relation of quotations from Remarque's texts in sources

Although Remarque is one of the main representatives of the Lost Generation and is known because of their texts about broken people in war and interwar periods, he won over many readers with his language of love. Therefore, "Arch of Triumph" is perceived by many not as a story about the broken destinies of people familiar with the war, but as a tragic love story. This explains why the most shared quotation is from his correspondence with Marlene Dietrich: 704 people shared this.

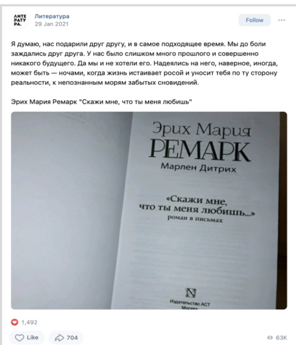
Such attention of Russian-native readers to German literature strengthens the position of German literature. In addition, we have seen increased attention to the military and emigrant texts of the Remark in recent years, this can be interpreted through the current political situation in which average citizens, like the characters of Remarque's novels, find themselves in the epicentre of a historical catastrophe.

Similarly to the popularity of poets, both Erich Maria Remarque and Mikhail Bulgakov also worked in the 20th century (Figure 3).

Table 1

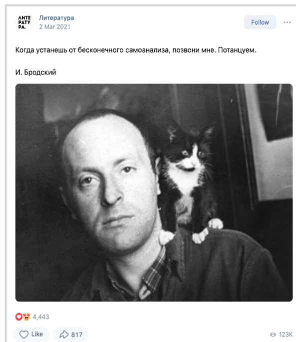
Statistical data on the most shared quotations of the most mentioned literary authors

Quoted text	Author	Statistics of quotations in social media		
		Likes	Comments	Reposts
Correspondence with Marlene Dietrich	Erich Maria Remarque	1492	10	704
“The Master and Margarita”	Mikhail Bulgakov	3186	13	550
Unknown	Attributed to Joseph Brodsky	4443	11	817
“Tips of blue flame are dancing about. . . ”	Sergei Esenin	3470	43	1029
“Please take care of me!” (has not been translated into English)	Marina Tsvetaeva	543	11	307
“Hands clasped, under the dark veil. . . ”	Anna Akhmatova	1543	9	366

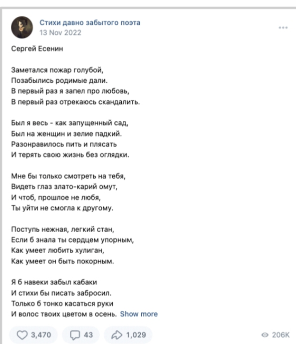


(a) Remarque’s

(b) Bulgakov’s

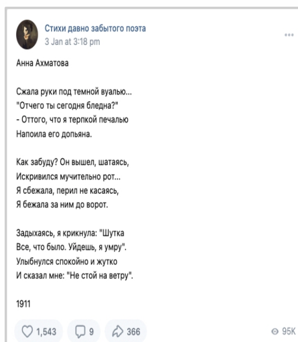


(c) Brodsky’s



(d) Esenin’s

(e) Tsvetaeva’s



(f) Akhmatova’s

Figure 3. The most shared quotations of the most mentioned literary authors

5.2. Text Analysis of Literary Quotations

To study the topics that primarily concern readers, we researched the texts that they cite.

We have analysed the text of quotations using Voyant Tools to term the topics of the most quoted texts. At this stage of the research, we analyse the texts of quotations from prose texts, but not poetic ones, because poetic texts contain a large number of metaphors that are not interpreted literally and require knowledge of the content to understand them.

We loaded the corpus with 5687 total words after cleaning data from prepositions, conjunctions, pronouns, interjections, proper names, and the same words that do not reflect the topic of the precedent text.

The word "love" (*любовь*) and its derivatives have a prominent appearance. In many of Remarque's texts, indeed, there is a love story. However, the most frequent words in the corpus emphasise the cross-cutting theme of all of Remarque's novels: the place of people broken by war in the world. The main figures for his characters are close social ties with friends and lovers. Thus, the word "friend" (*друг*) appears 22 times, and the word "human" (*человек*) – 17 times. These are some of the most common words (Figure 4).

Figure 4. Relation of quotations from Remarque's texts in sources

Figure 5. Relation of quotations from Bulgakov's "The Master and Margarita"

As we have already found out, among most novels of Remarque, there are novels included in the so-called emigrant pentalogy: "Flotsam", "Arch of Triumph", "The Night in Lisbon", "Shadows in Paradise", and "The Promised Land". As A. S. Porshneva, a researcher of Remarque's work, notes, "... in all these works, from the point of view of one or more emigrant heroes, a special world is being built – the world of emigration" [7]. The highest frequency of use of the words "friend" and "human" illustrates this thesis.

They are second only to the word "nothing" (*ничего*), which occurs 26 times, and pessimistically, but realistically predicts the most likely fate of the heroes, in which there will be nothing bright. By the way, we notice the same situation with the text of quotations from the novel "The Master and Margarita": the word "nothing" (*ничего*) appears ten times, and the most appeared word is "never" (*никогда*) – it is occurring 20 times and in combination with the word "nothing" increases the hopelessness of the main Bulgakov's characters (Figure 5).

Limitations of the study

Our study should be considered as an initial experiment which confirms the viability of research questions and introduces a framework which can be reused in subsequent studies. Here we would like to summarise some limitations:

- As with all studies using datasets, we extracted data from specific social media and selected groups matching our subject domain. Literary quotations are appearing less frequently in other social media content.
- We used one social media – VK. Any such study can be expanded to include content from different social media.
- We used several tools without making any attempts to improve their quality – this meant adding some time for checking the outcomes at each stage and correcting manually any potential mistakes.
- We also have not analysed quotations which appear on images – this is a very popular behaviour of social media users (and also some share videos with multiple quotations from the same author). These media can be analysed in a subsequent study using additional tools to extract texts from images/videos.

As a summary, our RQ1 and RQ2 are not confirmed by the evidence. However, we showed that the work with data from social media provides valuable insights thereby strengthening our understanding of the importance of literature in today's society. We have also provided a framework (see Figure 1) that can be adopted in other similar studies. Our last research question, RQ4, requires further exploration. Research has established channels to analyse the visibility of publications and there are specialised services around this (e.g. Altmetrics). Sharing literary quotations on social media requires further study as a form of information behaviour. It can be compared to the sharing of images of art but there are a limited number of in-depth studies on this topic, and even less on sharing poetic quotations.

On the other hand, the general tendency to share quotations is widespread and often results in spreading disinformation on the author's attribution of particular quotations. This aspect also requires further investigation.

6. Conclusions and next steps

In our study, we explored the nature of sharing literary quotations in the most popular Russian social media and arrived at the following conclusions.

Firstly, contrary to our hypothesis that Russian-native readers will prefer Russian-native authors, the most cited author turned out to be Erich Maria Remarque, the German-born writer of the Lost Generation.

Secondly, we also did not confirm the hypothesis that posts containing poems are the most popular. Nevertheless, they turn out to be the next few lines in the conditional rating according to the frequency of citation after Remarque.

The quantitative analysis showed that Russian readers prefer the poetry of the 20th century and representatives of a variety of literary trends: imagism (Sergei Esenin), acmeism (Anna Akhmatova), as well as poets who do not belong to any poetic school (Marina Tsvetaeva, Joseph Brodsky). In addition, we can see the importance of the work of Russian women poets for contemporary readers. Besides, we found a quotation of disputed authorship, which some people attribute to Joseph Brodsky.

We can draw some conclusions based on these results. The study suggests that an important place on the shelf of contemporary readers is occupied by texts, on the one hand, telling a tragic love story, and on the other hand, telling the difficult fate of a man of war. This is evidenced by several coincidences in the most frequent words used in such dissimilar novels as "The Master and Margarita" by Mikhail Bulgakov, the Russian writer, and Remarque's novels.

The compelling findings from this modest study not only affirm our commitment to further exploration but also inspire us to forge ahead, employing not just the same but even more cutting-edge quantitative methodologies in the future. The discoveries detected in the course of this study can serve as a basis for fundamental humanitarian research, in particular in the field of comparative literature. Finally, this study can be the basis for continuing the study of reading preferences using other quantitative and qualitative interdisciplinary methods and cover broader criteria that potentially affect the formation of reading preferences, such as the level of education, pop-cultural trends, social environment, and others.

In the general setting of information behaviour studies, we noted that community building within social media networks has not been extensively studied. We hope that this domain will attract some more attention and plan to expand our studies to other social media platforms and countries.

References

[1] T. Aichner, M. Grünfelder, O. Maurer and D. Jegeni, Twenty-five years of social media: A review of social media applications and definitions from 1994 to 2019, Cyberpsychology Behav. Soc. Netw. 24(4) (2021) 215–222, https://doi.org/10.1089/cyber.2020.0134.

[2] T. J. Bocala-Wiedemann, Social media as a tool for evangelism among youth and young adults, Great Commission Research Journal 14(1) (2022) 19–34, https://place.asburyseminary.edu/gcrj/vol14/iss1/2 (last accessed: 21.09.2023).

[3] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, From data mining to knowledge discovery in databases, AI Magazine, 17(3) (1996) 37–54.

[4] K. Fitzpatrick, Reporting from the Digital Humanities 2010 Conference, The Chronicle of Higher Education, July 13, 2010, https://www.chronicle.com/blogs/profhacker/reporting-from-the-digital-humanities-2010-conference.

[5] M. K. Gold (Ed.), Debates in the Digital Humanities, NED-New edition, University of Minnesota Press, 2012, http://www.jstor.org/stable/10.5749/j.ctttv8hq (last accessed: 21.09.2023).
[6] G. Griffin and M. Hayler, Research methods for reading digital data in the digital humanities, Edinburgh University Press, 2016, 214 pp.
[7] L. Haapanen and D. Perrin, Media and quoting: understanding the purposes, roles, and processes of quoting in mass and social media, in: The Routledge Handbook of Language and Media, C. Cotter and D. Perrin (Eds.), Routledge Handbooks in Linguistics, Routledge, London, 2017, 424–441, https://doi.org/10.4324/9781315673134-31.
[8] N. John, Sharing and social media: The decline of a keyword?, New Media & Society 26(4) (2024) 1891–1908, https://doi.org/10.1177/14614448221078603.
[9] A. Mas-Bleda, M. Makita, A. Mrva-Montoy and M. Thelwall, What makes a book tweet popular? Analysis of the most retweeted content posted by Spanish and non-Spanish book publishers, Revista Española de Documentación Científica 45(3) (2022) e322 (in Spanish), https://doi.org/10.3989/redc.2022.3.1904.
[10] A. S. Porshneva, The world of emigration in the German emigrant novel of the 1930s-1970s, Kabinetny Uchony Publishing, 2019.
[11] Russia: Most used social media platforms 2022, Statista, https://www.statista.com/statistics/867549/top-active-social-media-platforms-in-russia/ (last accessed: 21.09.2023).
[12] S. Schreibman, R. R. Siemens and J. Unsworth, A new companion to digital humanities, John Wiley & Sons, 2016, 567 pp, https://doi.org/10.1002/9781118680605.
[13] R. Vanherle, H. Hendriks and K. Beullens, Only for friends, definitely not for parents: Adolescents' sharing of alcohol references on social media features, Mass Communication and Society 26(1) (2023) 47–73, https://doi.org/10.1080/15205436.2022.2035767.
[14] VK for Developers, https://dev.vk.com/ru (last accessed: 21.09.2023).
[15] J. Vlieghe, J. Muls and K. Rutten, Everybody reads: Reader engagement with literature in social media environments, Poetics 54 (2016) 25–37, https://doi.org/10.1016/j.poetic.2015.09.001.

Received on March 25, 2024
Revised on April 28, 2024
Accepted on May 10, 2024

Inessa Kevler and Karina Agafonova

ITMO University
Saint Petersburg
RUSSIA

E-mails: inessakevler@gmail.com
agafonovakarin@gmail.com

Krassimira Ivanova

Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
Acad. G. Bonchev Str., Bl. 8
1113 Sofia
BULGARIA

E-mail: kivanova@math.bas.bg

Milena Dobreva

University of Strathclyde
Glasgow
SCOTLAND

E-mail:
milena.dobreva@strath.ac.uk

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“
ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Том 111

ANNUAL OF SOFIA UNIVERSITY “ST. KLIMENT OHRIDSKI”
FACULTY OF MATHEMATICS AND INFORMATICS
Volume 111

NO ZEROS OF THE PARTIAL THETA FUNCTION IN THE UNIT DISK

VLADIMIR PETROV KOSTOV

We prove that for $q \in (-1,0)\cup(0,1)$, the partial theta function $\theta(q,x) := \sum_{j=0}^{\infty} q^{j(j+1)/2}x^j$ has no zeros in the closed unit disk.

Keywords: partial theta function, Jacobi theta function, Jacobi triple product

2020 Mathematics Subject Classification: 26A06

1. Introduction

We consider the partial theta function $\theta(q,x) := \sum_{j=0}^{\infty} q^{j(j+1)/2}x^j$, where $q \in [-1,1]$ is a parameter and $x \in \mathbb{R}$ is a variable. In particular, $\theta(0,x) \equiv 1$, $\theta(1,x) = \sum_{j=0}^{\infty} x^j = 1/(1-x)$ and

$$\theta(-1,x) = \sum_{j=0}^{\infty}(-1)^j x^{2j} + \sum_{j=0}^{\infty}(-1)^{j+1}x^{2j+1} = (1-x)/(1+x^2).$$

For each $q \in (-1,0) \cup (0,1)$ fixed, θ is an entire function of x of order 0.

The name “partial theta function” is connected with the fact that the *Jacobi theta function* equals $\Theta(q,x) := \sum_{j=-\infty}^{\infty} q^{j^2}x^j$ while $\theta(q^2,x/q) = \sum_{j=0}^{\infty} q^{j^2}x^j$. “Partial” refers to the fact that summation in θ is performed only from 0 to ∞. One can observe that

$$\Theta^*(q,x) := \Theta(\sqrt{q},\sqrt{q}x) = \sum_{j=-\infty}^{\infty} q^{j(j+1)/2}x^j = \theta(q,x) + \theta(q,1/x)/x.$$

DOI: 10.60063/GSU.FMI.111.129-137

The function θ satisfies the relation

$$\theta(q,x) = 1 + qx\theta(q,qx). \tag{1.1}$$

Applications of θ to questions concerning asymptotics and modularity of partial and false theta functions and their relationship to representation theory and conformal field theory (see [6] and [4]) explain part of the most recent interest in it. Previously, this function has been studied with regard to Ramanujan-type q-series (see [25]), statistical physics and combinatorics (see [24]), the theory of (mock) modular forms (see [5]) and asymptotic analysis (see [3]); see also [1].

Another domain in which θ plays an important role is the study of section-hyperbolic polynomials. These are real polynomials with all roots real negative and all whose finite sections (i.e., truncations) have also this property, see [9, 21, 22]; the cited papers are motivated by results of Hardy, Petrovitch and Hutchinson (see [7, 8, 23]). Various analytic properties of the partial theta function are proved in [11–20] and other papers of the author.

The basic result of the present text is the following theorem (proved in Section 3):

Theorem 1. *For each $q \in (-1,0) \cup (0,1)$ fixed, the function θ has no zeros in the closed unit disk $\overline{\mathbb{D}_1}$.*

In the next section we discuss the question to what extent Theorem 1 proposes an optimal result. In Section 4 we make comments and formulate some open questions.

2. Optimality of the result

2.1. The theorem of Eneström-Kakeya

For $q \in (0,1)$, the theorem of Eneström-Kakeya about polynomials with positive coefficients (see [2]) implies that the modulus of each root of a polynomial $a_0 + a_1x + \cdots + a_nx^n$, $a_j > 0$, is not less than $\min_j |a_{j-1}/a_j|$. When this polynomial equals $1 + qx + \cdots + q^{n(n-1)/2}x^{n-1}$, the minimum equals $1/q$. Thus all zeros of all finite truncations of $\theta(q,.)$ (and hence all zeros of θ itself) lie outside the open disk $\mathbb{D}_{1/q}$.

Hence for $q \in (0,1)$ (but not for $q \in (-1,0)$), Theorem 1 follows from the theorem of Eneström-Kakeya. A hint how to obtain for $q \in (-1,0)$ a disk of a radius tending to ∞ as $q \to 0^-$ and free from zeros of θ is given in Remark 4.

Remark 2. For $q \in (-1,0)$, it is not true that $\theta(q,.)$ has no zeros inside the disk $\mathbb{D}_{1/|q|}$. Indeed, the function $\theta(-0.4,.)$ has a zero $1.96\ldots < 1/0.4 = 2.5$. More generally, the zero of $\theta(q,.)$ closest to the origin can be expanded in a Laurent series (convergent for $0 < |q|$ sufficiently small) of the form $-1/q - 1 + O(q)$, see [11]. For $q \in (-1,0)$ and $|q|$ sufficiently small, this number belongs to the interval $(0, 1/|q|)$. See also [20], where the zero set of θ is illustrated by pictures.

2.2. Optimality with respect to the parameter q

(1) This result cannot be generalized in the case when q and x are complex. Indeed, suppose that $q \in \mathbb{D}_1$ and $x \in \mathbb{C}$. Then the function θ has no zeros x with $|x| < 1/2$. In fact, it has no zero for $|x| \leq 1/2|q|$, see [10, Proposition 7]. On the other hand, the radius of the disk in the x-space centered at 0 in which θ has no zeros for any $q \in \mathbb{D}_1$ is not larger than $0.56\ldots$. Indeed, consider the series θ with $q = \omega := e^{3\pi i/4}$. It equals

$$\left(\sum_{j=0}^{7} \omega^{j(j+1)/2} x^j\right) \Big/ (1 - x^8).$$

Its numerator has a simple zero $x_* := 0.33\ldots + 0.44\ldots i$ whose modulus equals $0.56\ldots$. Hence for $\rho \in (0,1)$ sufficiently close to 1, the function $\theta(\rho e^{3i\pi/4}, .)$ has a zero close to x_*. To see this one can fix a closed disk $\mathcal{D}$ about x_* of radius < 0.1 in which x_* is the only zero of $\theta(e^{3i\pi/4}, .)$. As ρ tends to 1^-, the modulus of the difference $\theta(e^{3i\pi/4}, x) - \theta(\rho e^{3i\pi/4}, x)$ tends uniformly to 0 for $x \in \partial\mathcal{D}$ (the border of $\mathcal{D}$), because the series θ converges uniformly for $|x| < |x_*| + 0.1$, $|q| \leq 1$. The Rouché theorem implies that the function $\theta(\rho e^{3i\pi/4}, .)$ has the same number of zeros in $\mathcal{D}$ (counted with multiplicity) as the function $\theta(e^{3i\pi/4}, .)$.

(2) Set $q := |q|e^{i\phi}$. We show that there exists no interval (i.e., arc) J on the unit circle centered at 1 or -1 and such that for $\phi \in J$ and $|q| < 1$, the zeros of $\theta(q, .)$ are all of modulus ≥ 1. Suppose that $n \in \mathbb{N}$, $n > 2$, and that ω is a primitive root of unity of order n. If n is odd, then the sequence of numbers $\omega^{k(k+1)/2}$ is n-periodic, because $(n+1)/2 \in \mathbb{N}$, and one obtains

$$\theta(\omega, x) = P(x)/(1 - x^n), \quad P := \sum_{j=0}^{n-1} a_j x^j, \quad a_j = \omega^{j(j+1)/2}.$$

If n is even, then this sequence is clearly $(2n)$-periodic, but it is not n-periodic, because $\omega^{n(n+1)/2} = -1$. One has

$$\theta(\omega, x) = Q(x)/(1 - x^{2n}), \quad Q := \sum_{j=0}^{2n-1} b_j x^j, \quad b_j = \omega^{j(j+1)/2}.$$

The polynomials P and Q are self-reciprocal, i.e., $a_{(n-1)/2-s} = a_{(n-1)/2+s}$ and $b_{(2n-1)/2-s} = b_{(2n-1)/2+s}$. Indeed, for the polynomial P this follows from

$$\begin{aligned}((n-1)/2 - s)((n-1)/2 - s + 1)/2 &\equiv (n - (n-1)/2 + s)(n - (n-1)/2 + s - 1)/2 \\ &= ((n-1)/2 + s)((n-1)/2 + s + 1)/2 \mod [n].\end{aligned}$$

For the polynomial Q one gets

$$\begin{aligned}((2n-1)/2 - s)((2n-1)/2 - s + 1)/2 &\equiv (2n - (2n-1)/2 + s)(2n - (2n-1)/2 + s - 1)/2 \\ &= ((2n-1)/2 + s)((2n-1)/2 + s + 1)/2 \mod [2n].\end{aligned}$$

We show that at least one root of the polynomial P and at least one root of Q belong to the interior of the unit disk. Indeed, these polynomials are monic and $P(0) = Q(0) = 1$. The product of their roots being equal to ± 1, the only possibility for P and Q not to have roots in $\mathbb{D}_1$ is all their roots to be of modulus 1. These polynomials are self-reciprocal, so $P(z) = 0$ (resp. $Q(z) = 0$) implies $P(1/z) = 0$ (resp. $Q(1/z) = 0$). But if $|z| = 1$, then $1/z = \bar{z}$. This means that P and Q can have as roots either ± 1 or complex conjugate pairs, i.e., P and Q must be real which is false as their coefficients of x equal $\omega \neq \pm 1$.

So P and Q have each at least one root in $\mathbb{D}_1$. As in part (1) of this subsection one deduces that for $|q|$ sufficiently close to 1 and for $e^{i\phi} = \omega$, the function $\theta(q, .)$ has a zero in $\mathbb{D}_1$. Primitive roots are everywhere dense on the unit circle. This implies the absence of an interval J as above.

2.3. Optimality with respect to the variable x

Suppose first that $q \in (-1, 0)$. Then in the formulation of Theorem 1 one cannot replace the unit disk by a disk of larger radius. Indeed, the zero of the numerator of $\theta(-1, x)$ (which equals 1) is the limit as q tends to -1^+ of the smallest positive zero of $\theta(q, x)$, see [20, Part (2) of Theorem 3], so in any disk $\mathbb{D}_{1+\varepsilon}$, $\varepsilon > 0$, there is a zero of θ for some $q \in (-1, 0)$.

Suppose now that $q \in (0, 1)$.

Conjecture 3. *Theorem 1 does not hold true if one replaces in its formulation the unit disk by a disk of larger radius.*

The following numerical example shows why this conjecture should be considered plausible. Set $\theta_{100} := \sum_{j=0}^{100} q^{j(j+1)/2} x^j$ (the 100th truncation of θ). For $q = 0.98$, the function $\theta_{100}(0.98, .)$ has a zero $\lambda_0 := 1.209\ldots + 0.511\ldots i$, of modulus $1.312\ldots$. For $q = 0.98$ and $x = 1.32$, the first two terms of θ which are not in θ_{100} equal $y_{101} := 7.407\ldots \times 10^{-33}$ and $y_{102} := 1.270\ldots \times 10^{-33}$, respectively. Their ratio is $y_{101}/y_{102} > 5.5$ and the moduli of the terms of θ decrease faster than a geometric progression. Hence for $|x| < 1.32$, one has

$$T_0 := |\theta(0.98, x) - \theta_{100}(0.98, x)| < y_{101}/(1 - 5.5^{-1}) = 9.053\ldots \times 10^{-33}.$$

On the other hand, $\Lambda_0 := (\partial\theta/\partial x)(0.98, \lambda_0) = 27.180\ldots + 18.959\ldots i$ with $|\Lambda_0| > 33$. Thus one should expect to find a zero of $\theta(0.98, .)$ close to λ_0 (the truncated terms are expected to change the position of λ_0 by $\approx T_0/|\Lambda_0|$ which quantity is of order 10^{-34}. So in the formulation of Theorem 1 one should not be able to replace the unit disk by a disk of radius larger than 1.32.

3. Proof of Theorem 1

We remind first that the *Jacobi triple product* is the identity

$$\Theta(q, x^2) = \prod_{m=1}^{\infty} (1 - q^{2m})(1 + x^2 q^{2m-1})(1 + x^{-2} q^{2m-1})$$

which implies $\Theta^*(q,x) = \prod_{m=1}^{\infty}(1-q^m)(1+xq^m)(1+q^{m-1}/x)$. Thus

$$\prod_{m=1}^{\infty}(1-q^m)(1+xq^m)(1+q^{m-1}/x) = \theta(q,x) + \theta(q,1/x)/x. \tag{3.1}$$

Suppose that $q \in (-1,0) \cup (0,1)$, $x_0 \in \mathbb{C}$, $|x_0| = 1$ (hence $\overline{x_0} = 1/x_0$), and that $\theta(q,x_0) = 0$. The coefficients of θ being real, one has $\theta(q,\overline{x_0}) = \overline{\theta(q,x_0)} = 0$, so the right-hand side of equation (3.1) equals 0 for $x = x_0$. However for $x = x_0$, the left-hand side vanishes only for $x_0 = -1$.

For $q \in (0,1)$, one has $\theta(q,-1) = \sum_{j=0}^{\infty}(-1)^j q^{j(j+1)/2}$, and the latter function takes only values from the interval $(1/2,1)$, with $\lim_{q\to 1^-} = 1/2$, see [10, Propositions 14 and 16]. For $q \in (-1,0)$, one sets $u := -q$, so

$$\begin{aligned}\theta(q,-1) = \theta(-u,-1) &= 1 + u - u^3 - u^6 + u^{10} + u^{15} - u^{21} - u^{28} + \cdots \\ &= 1 - u^3 + u^{10} - u^{21} + \cdots + u - u^6 + u^{15} - u^{28} + \cdots > 0,\end{aligned}$$

because this is the sum of two Leibniz series with positive initial terms. Thus for $q \in (-1,0) \cup (0,1)$, the partial theta function has no zeros of modulus 1.

For $-1/2 < q < 1/2$, one has $\theta(q,x) \neq 0$ for any $x \in \overline{\mathbb{D}_1}$, because

$$|\theta(q,x)| \geq 1 - |q| - |q|^3 - |q|^6 - \cdots \geq 1 - |q| - |q|^2 - |q|^3 - \cdots = (1-2|q|)/(1-|q|) > 0.$$

As the parameter q varies in $(0,1)$ or in $(-1,0)$, the zeros of θ depend continuously on q. For $|q| < 1/2$, there are no zeros of θ in $\overline{\mathbb{D}_1}$ and for $q \in (-1,0) \cup (0,1)$, no zero of θ crosses $\partial\mathbb{D}_1$ (the border of the unit disk). Hence for $q \in (-1,0) \cup (0,1)$, there are no zeros of θ in $\overline{\mathbb{D}_1}$.

Remark 4. One can prove that for $|q| \leq 0.4$, the function $\theta(q,.)$ has no zeros in the closed disk $\overline{\mathbb{D}_{1/\sqrt{|q|}}}$. Indeed,

$$|\theta(q,1/\sqrt{|q|})| \geq 1 - \sum_{j=1}^{\infty}|q|^{j(j+1)/2-j/2} = 1 - \sum_{j=1}^{\infty}|q|^{j^2/2} \geq 1 - \sum_{j=1}^{\infty}0.4^{j^2/2} = 0.19\ldots > 0.$$

4. Comments and open questions

4.1. The case $q \in (0,1)$

In Figure 1 we show the images for $q = 0.2$ (the smaller oval) and for $q = 0.7$ (the larger oval) of the unit circle in the x-plane under the mapping $x \mapsto \theta(q,x)$, together with the vertical line $\operatorname{Re} x = 1/2$.

It would be interesting to know whether:

1. for any $q \in (0,1)$, the image of the unit circle is a convex oval about the point $(1,0)$ and belonging to the half-plane $\operatorname{Re} x > 1/2$;

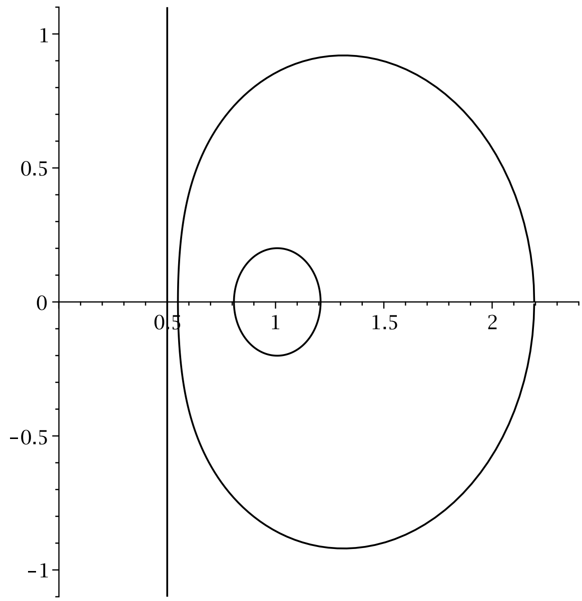


Figure 1. The vertical line $\mathrm{Re}\, x = 1/2$ and the images of the unit circle under the map $x \mapsto \theta(q, x)$ for $q = 0.2$ and $q = 0.7$

2. for $0 < q_1 < q_2 < 1$, the image of the unit circle for $q = q_1$ lies inside its image for $q = q_2$.

These questions are motivated by the fact that for $q = 1$, one has $\theta(1, x) = 1/(1 - x)$, and for $|x| = 1$, it is true that $\mathrm{Re}(1/(1 - x)) = 1/2$, i.e., the vertical line $\mathrm{Re}(1/(1 - x)) = 1/2$ is the image of the unit circle for $q = 1$; on the other hand, the point $(1, 0)$ is the image of the unit circle for $q = 0$.

4.2. The case $q \in (-1, 0)$

In Figure 2 we show the images for $q = -0.2$ (small oval in dashed line), $q = -0.53$ (closed contour in dotted line), $q = -0.7$ (curve with self-intersection in dashed line) and $q = -0.85$ (curve with self-intersection in solid line) of the unit circle in the x-plane under the mapping $x \mapsto \theta(q, x)$.

The following questions are natural to ask:

3. Is it true, and for which value of $v \in (0, 1)$, that for $q \in (-v, 0)$, the corresponding image is a convex oval about the point $(1, 0)$?

4. Is it true that for $q \in (-w, -v)$, $-1 < -w < -v < 0$, the corresponding image changes convexity twice "at its right" (as this seems to be the case of the curve given in dotted line)?

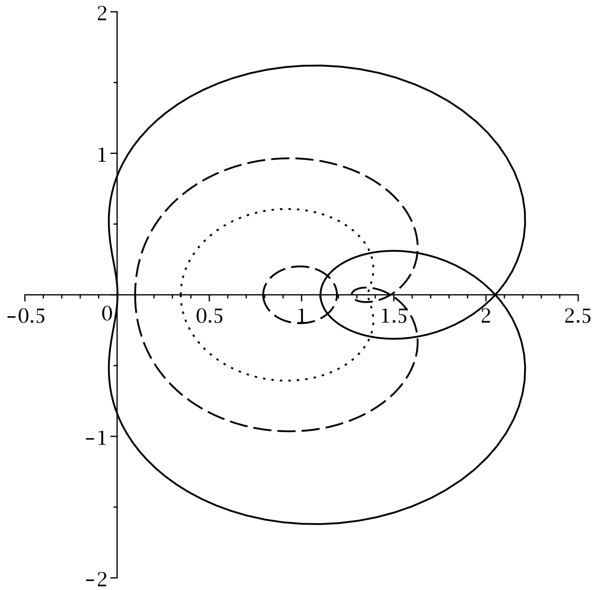


Figure 2. The images of the unit circle under the map $x \mapsto \theta(q, x)$ for $q = -0.2$, $q = -0.53$, $q = -0.7$, and $q = -0.85$

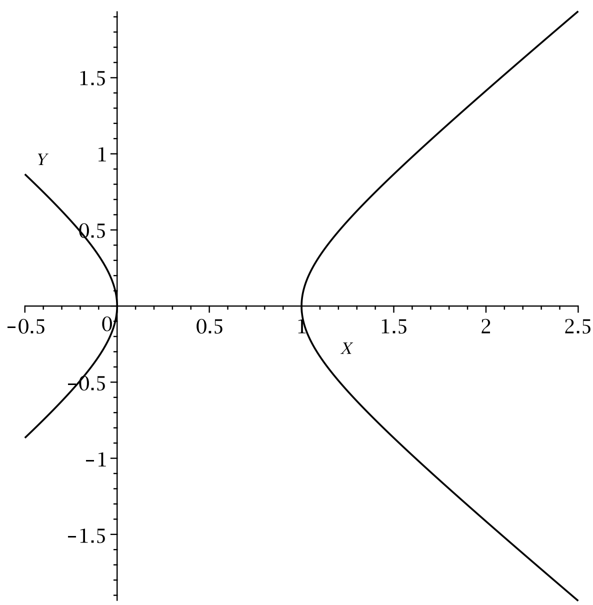


Figure 3. The image of the unit circle under the map $x \mapsto \theta(q, x)$ for $q = -1$

5. Is it true that for $q \in (-1, -w)$, the image has a self-intersection point? One can expect that for $q = -w$, the image has a cusp point.

6. Is it true that for $q \in (-1, -w')$, $-1 < -w' < -w$, the image has still self-intersection and changes convexity twice "at its left"?

7. Is it true that for $q \in (-1, -w'')$, $-1 < -w'' < -w'$, the image has still self-intersection, changes convexity twice "at its left" and intersects the vertical axis at four points? (For $q = w''$, the image is supposed to have two tangencies with the vertical axis.)

8. Is it true that these are all transformations which the image undergoes for $q \in (-1, 0)$?

9. Is it true that for $-1 < q_2 < q_1 < 0$, the image of the unit circle for $q = q_1$ lies inside its image for $q = q_2$? "Inside" means "inside the contour excluding (for $q_2 > w$) the loop".

It should be observed that for values of q close to -1, the image seems to pass through the origin. In reality, it passes very close to it, but nevertheless to its right, according to Theorem 1. The image of the unit circle for $q = -1$ is the hyperbola $Y^2 - X^2 - X = 0$, where $X := \operatorname{Re} x$ and $Y := \operatorname{Im} x$, see Figure 3. (The centre of the hyperbola is at $(1/2, 0)$, its asymptotes are the lines $Y = \pm(X - 1/2)$.) Following a similar logic one can assume that the point $(1, 0)$ remains in the exterior of the loop of the image (the loop existing for $q > w$). The proximity of the image to the origin makes it seem unlikely that one could prove the absence of zeros of θ in a disk of a radius larger than 1 (for all $q \in (-1, 0)$).

Acknowledgements. The author is grateful to A. Vishnyakova and B. Shapiro for the useful comments of this text.

References

[1] G. E. Andrews and B. C. Berndt, Ramanujan's lost notebook, Part II, Springer, NY, 2009.
[2] N. Anderson, E. B. Saff, and R. S. Varga, On the Eneström-Kakeya theorem and its sharpness, Lin. Algebra Appl. 28 (1979) 5–16.
[3] B. C. Berndt and B. Kim, Asymptotic expansions of certain partial theta functions, Proc. Amer. Math. Soc. 139(11) (2011) 3779–3788.
[4] K. Bringmann, A. Folsom, and A. Milas, Asymptotic behavior of partial and false theta functions arising from Jacobi forms and regularized characters, J. Math. Phys. 58(1) (2017) 011702, 19 pp.
[5] K. Bringmann, A. Folsom, and R. C. Rhoades, Partial theta functions and mock modular forms as q-hypergeometric series, Ramanujan J. 29(1–3) (2012) 295–310, http://arxiv.org/abs/1109.6560.
[6] T. Creutzig, A. Milas, and S. Wood, On regularised quantum dimensions of the singlet vertex operator algebra and false theta functions, Int. Math. Res. Not. IMRN 2017(5) (2017) 1390–1432.
[7] G. H. Hardy, On the zeros of a class of integral functions, Messenger of Mathematics 34 (1904) 97–101.

[8] J. I. Hutchinson, On a remarkable class of entire functions, Trans. Amer. Math. Soc. 25 (1923) 325–332.
[9] O. M. Katkova, T. Lobova, and A. M. Vishnyakova, On power series having sections with only real zeros, Comput. Methods Funct. Theory 3(2) (2003) 425–441.
[10] V. P. Kostov, On the zeros of a partial theta function, Bull. Sci. Math. 137(8) (2013) 1018–1030.
[11] V. P. Kostov, Stabilization of the asymptotic expansions of the zeros of a partial theta function, C. R. Acad. Bulg. Sci. 68(10) (2015) 1217–1222.
[12] V. P. Kostov, On the double zeros of a partial theta function, Bull. Sci. Math. 140(4) (2016) 98–111.
[13] V. P. Kostov, On the multiple zeros of a partial theta function, Funct. Anal. Appl. 50(2) (2016) 153–156.
[14] V. P. Kostov, On the double zeros of a partial theta function, Bull. Sci. Math. 140(4) (2016) 98–111.
[15] V. P. Kostov, The closest to 0 spectral number of the partial theta function, C. R. Acad. Bulg. Sci. 69(9) (2016) 1105–1112.
[16] V. P. Kostov, A domain containing all zeros of the partial theta function, Publicationes Mathematicae Debrecen 93(1–2) (2018) 189–203.
[17] V. P. Kostov, A separation in modulus property of the zeros of a partial theta function, Anal. Math. 44 (2018) 501–519.
[18] V. P. Kostov, On the complex conjugate zeros of the partial theta function, Funct. Anal. Appl. 53(2) (2019) 149–152.
[19] V. P. Kostov, Partial theta function and separation in modulus property of its zeros, Vietnam J. Math. 48(1) (2020) 145–157.
[20] V. P. Kostov, On the zero set of the partial theta function, Serdica Math. J. 45 (2019), 225–258.
[21] V. P. Kostov and B. Z. Shapiro, Hardy-Petrovitch-Hutchinson's problem and partial theta function, Duke Math. J. 162(5) (2013) 825–861.
[22] I. V. Ostrovskii, On zero distribution of sections and tails of power series, Israel Math. Conf. Proceedings 15 (2001) 297–310.
[23] M. Petrovitch, Une classe remarquable de séries entières, Atti del IV Congresso Internationale dei Matematici, Rome (Ser. 1) 2 (1908) 36–43.
[24] A. Sokal, The leading root of the partial theta function, Adv. Math. 229(5) (2012) 2603–2621, arXiv:1106.1003.
[25] S. O. Warnaar, Partial theta functions. I. Beyond the lost notebook, Proc. London Math. Soc. 87(2) (2003) 363–395.

Received on October 26, 2023
Accepted on December 6, 2023

Vladimir Petrov Kostov
Université de Nice
Laboratoire de Mathématiques
Parc Valrose
06108 Nice Cedex 2
FRANCE
E-mail: `vladimir.kostov@unice.fr`

ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“
ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Том 111

ANNUAL OF SOFIA UNIVERSITY “ST. KLIMENT OHRIDSKI”
FACULTY OF MATHEMATICS AND INFORMATICS
Volume 111

EXCHANGE OF OCCUPATIONAL HEALTH ASSESSMENT SUMMARIES BASED ON THE EN 13606 STANDARD

EVGENIY KRASTEV, PETKO KOVACHEV, SIMEON ABANOS and DIMITAR TCHARAKTCHIEV

The globalization of world economy stimulates large number of people to search abroad for better employment opportunities. This is the reason to consider the exchange of occupational health assessment summary (OHAS) content from cross-border view point. It is a problem of larger social-significance than the cross-border exchange for International Patient Summary (IPS). While the IPS dataset is well studied in the framework of an international standard, the OHAS remains rather insufficiently explored in the existing literature. This paper proposes a novelty systematic approach to provide a practicable solution to this problem by ensuring semantic interoperability in the exchange of OHAS. It starts by presenting a description of the use case for cross-border exchange. Next, the OHAS dataset is defined on the basis of exploring common data requirements the national legislation of EU countries. The final step of this approach is the design of an EN 13606 archetype of OHAS satisfying all the requirements for semantic interoperability in the exchange of clinical data. Further on, the static, non-volatile and reusable information model of OHAS is used to create EN 13606 instances that are valid with respect to the Reference model and the datatypes of this standard. Finally, this paper demonstrates the software implementation of basic activities in the OHAS use case in a fully functional web application, where the EN 13606 instances are managed by means of a native XML database.

Keywords: eHealth, occupational medicine, occupational health summary, use cases, semantic interoperability, EN 13606, workplace hazards, occupational disease, proactive prevention

2020 Mathematics Subject Classification: 68P05, 68-04

CCS Concepts:

• CCS∼Software and its engineering∼Software creation and management∼Software development techniques∼Software prototyping;

• CCS∼Information systems∼Information systems applications∼Enterprise information systems∼Enterprise applications

DOI: 10.60063/GSU.FMI.111.139-152

1. Introduction

One of the primary tasks of occupational health is to provide public health services for maintenance of the workers' health state by continuous health assessment of their capacity to execute the professional duties of the assigned them job position. These services aim to protect the health and well-being of employees and workers as well as to improve the productivity and quality of work [29]. It is widely recognized that the conditions in the workplace environment significantly impact human's health state and contribute to contracting illnesses caused by exposure to risk factors accompanying the execution of work activity [12]. Prolonged exposure to hazards in the workplace environment, such as interaction with dangerous chemical or biological agents, radiation or even continuous stress at work stimulates the development or progression of occupational diseases [7].

Occupational diseases, unlike other illnesses, have causal relationship with specific hazards established in the workplace environment or specific occupations. Therefore, such diseases are most often referred to as occupational diseases [23]. The International Labor Organization (ILO) acknowledges occupational diseases as a serious world-wide problem having significant burden on the economy [11,25,30]. According to public statistic data the expenses on occupational diseases are increasing every year and besides, exceed the costs spent on treating other socially significant diseases like cancer and diabetes [5,21]. The medical costs associated with occupational disease treatment appear to be quite significant. Additionally, productivity losses resulting from work-related ill health or disabilities generate another substantially larger fraction of these costs. These costs are partially covered by private and social insurance funds due for occupational disease. However, a major part of these expenses is paid by the workers and their families suffering from occupational disease. Thus, occupational diseases have many associated costs that cannot be expressed just in financial terms. These diseases impact the employee, the employer and the whole society, while recent statistical data suggest the need for improving the existing approaches and public services for prevention of occupational diseases.

A starting point for successful management and prevention of occupational diseases is the identification of these diseases. According to the ILO, there are four groups of internationally recognized occupational diseases [15]. The diseases in the ILO list are rather generally described in that list without a reference to dangerous levels of workplace hazards, duration or type of exposure allowing to establish a link between a risk factor in work activities and the observed health disorders. Besides, this list had been revised for the last time 10 years ago and the representation of occupational diseases needs to be updated in order to be used efficiently with modern information technologies. More detailed classification of occupational diseases accompanied by related descriptions of risk factors may be discovered at national level [2, 14].

The increased globalization of the workforce market requires deeper harmonization of international standards, national legislation and practices aiming to achieve effective preventive and protective measures [16]. For example, the European Union

(EU) has introduced lists of indicative limit values for occupational exposure to hazardous chemical agents during work activities [8], where agents are accurately described in terms of their Chemical Abstract Service (CAS) registry number [1]. Similar EU directives establish limit values for exposure to physical hazards, biological agents as well as provisions on avoiding other risks in the workplace environment like psychological, workload, or ergonomical risks. While this approach for quantitative exposure assessment aims to ensure a safer workplace, lately employers prefer to apply a proactive approach to occupational diseases prevention. Diagnosis, early discovery and avoiding the progression of occupational diseases are in the focus of this approach. At the same time, employees often change job positions or workplaces, each one of which, impacts the employee's health in one or another way. Therefore, a proper diagnosis of occupational disease needs to take into account information from the occupational history of the employee relating quantitative exposure assessments with observed clinical symptoms and hazards in the workplace as well as conclusions from occupational health assessments for fitness to execute work activities.

Such information is usually recorded in the occupational health assessment summary (OHAS) of the employee. It is not just a single medical record, rather a summary of medical assessments accumulated during the occupational history of the worker. The OHAS is applicable in use cases similar to those of the International Patient Summary (IPS) [9]. Although OHAS and the IPS serve different purposes, both of them have similar use cases that involve exchange of clinical data both at national and cross-border level. The EN ISO 27269 [6] standard introduces a uniform structure for the IPS dataset and EN 13606 [17, 18] enables semantic interoperability of IPS exchange [20, 27]. Unlike the similarity in use cases and the importance of proactive strategies for prevention of occupational diseases, currently, there is no international standard for representing the OHAS of workers and employees. This complicates a lot the development of services for sharing of information from the OHAS between the various professionals. Typically, the issues of medical confidentiality and privacy of data are underestimated in practice. Studies of existing literature sources show that only few countries have provisions in their legislation for management of occupational health assessments in relation to the occupational history of the employee or worker thus, limiting significantly the information required for decision making regarding surveillance and prevention of occupational disease disorders [10, 16, 26]. For example, the French Law No. 2021-1018 of August 2, 2021 makes medical records a major preventive tool in occupational health part of the French "Dossier Medical Personel" (DMP) that is equivalent to the Patient's Summary in other countries, while Decree No. 2022-1434 of November 15, 2022 specifies in particular the content and the conditions of transmission.

The focus of this paper is two-fold. The first objective is to outline the typical use cases of OHAS involving exchange of clinical data for the purpose of proactive prevention of occupational diseases. Secondly, it considers the French and Bulgarian OHAS [22, 26] in a case study aiming to demonstrate semantic interoperability in exchange of OHAS instances compatible with an EN 13606 archetype. Both the use cases and the representation of an OHAS in the reference model of EN 13606 are novelties in the existing literature, while the results from computer experiments prove that the exchange of OHAS has a practicable solution with significant number of applications in occupational health.

The structure of this paper is organized as follows. The following section describes the use cases for exchange of OHAS and the EN 13606 archetype representing OHAS according to the provisions of the Bulgarian national legislation. In section 3 we consider a typical case study for the implementation of the use cases. For this purpose we demonstrate exchange of valid instances of the EN 13606 archetype by means of a native XML database. In conclusion, we discuss the applicability of the obtained results and our plans for future research.

2. Methods

Proactive prevention of occupational diseases relies on the ability to make decisions employing clinical data collected systematically from occupational health assessments conducted during the occupational history of the employee or worker. We refer to this dataset as occupational health assessment summary (OHAS). Unlike the IPS, data entries in the OHAS are provided by occupational physicians or by health professionals from the multidisciplinary occupational health team of experts during health surveillance visits or during health assessment for job position fitness. Its structure and scope are usually established by provisions in the national legislation of each country with the primary objective of providing continuity of care and coordination of efforts in surveillance and prevention of occupational diseases. Therefore, without loss of generality we assume that management of OHAS is strictly defined by the national legislation. On the other side, the globalization of the workforce market requires taking into consideration the case of cross-border exchange of OHAS. From this viewpoint, it follows that OHAS must be also cross-border accessible both for viewing and updating content, subject to the existing norms for medical confidentiality and data privacy. The use case to communicate OHAS and its content is introduced in Subsection 2.1 as an initial step that motivates the need for the standardization of OHAS content. Apparently, this use case requires a more detailed presentation that exceeds the scope of this paper. Its description brings up the outstanding problem of semantic interoperability of clinical data in OHAS exchanges. Therefore, in Subsection 2.2 we focus the attention on building a real-life case study for semantic interoperability exchange of OHAS content [22] represented in terms of the dual reference model of EN 13606 [18].

2.1. Use case

The here considered use case describes the different ways the OHAS and its content are communicated among the occupational health agencies and participants in the international workforce market. People are free to get employed in the country of their origin (Country A) or abroad (Country B). The appointment of employees without knowledge about their occupational health background is risky, especially, when it concerns job positions involving exposure to known hazards in the workplace environment. Therefore, in most cases of employment benefits for proactive prevention of occupational disease and also for the well-being of the worker can be gained, provided the exchanged data in the OHAS is up to date and besides, interpreted semantically correct, by the respective occupational health professional (OHP). The following use case proposes a solution to this problem. Its scope could be cross-border, national or regional.

Participants in this use case are the employee/worker, the OHP in the employee/worker's country of origin (Country A) and the OHP in the country of employment (Country B). In a particular case, Country B could designate Country A or a region of Country A when the use case scope is national or regional.

Important prerequisites for executing the functional requirements include the following observations. Different countries and even, different occupational agencies in the same country operate heterogeneous software systems and may use different informational technologies for representing the OHAS content. Moreover, OHAS content may be recorded in different languages, clinical vocabularies or measurement systems. All of it presents challenges including semantic interoperability in the practical implementation of sharing OHAS content between various professionals.

Typically, the OHAS dataset contains at least the following sections:

- A section with medico-administrative data.
- An ordered group of sections recorded during the occupational history of the employee/worker where each record contains data about:
- Workplace activities relevant to job position.
- Hazards specific to the workplace environment.
- Risks based on formal quantitative exposure assessment procedures.
- Established occupational diseases (diagnoses), documented disabilities, existence or absence of a pathology possibly linked to occupational exposure.
- Medical conclusions and medical contraindications for job position fitness.

There are two distinct versions of the use case depending on the actors involved in the exchange activities. The first version is the case when the employee/worker works occasionally or regularly in Country B, whereby he may already have an OHAS stored in Country A. This describes the Use case of the employee/worker working abroad. The second use case considers exchange of OHAS when employment occurs in the country of origin, i.e. when Country B coincides with Country A. This describes the Use case of the native employee/worker. In terms of cross-border exchange of OHAS, the first use case is the Primary use case (Figure 1).

Because of the absence of a standard for a comprehensive definition of the OHAS dataset at the EU level, the implementation of the Primary use case is accompanied by significant semantic interoperability problems. Without loss of generality we

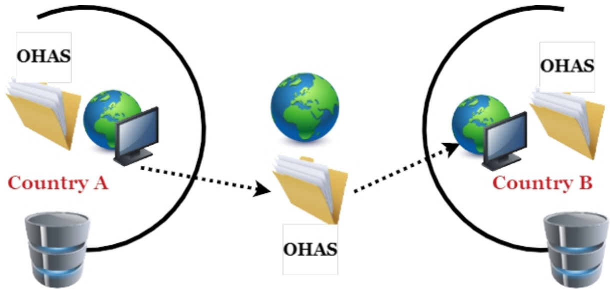


Figure 1. Primary use case for cross-border exchange of OHAS (employee/worker working abroad)

assume at this point that compliant implementations are feasible at national or regional level, although there remain serious obstacles to overcome at this level as well. Therefore, a consensus agreement is getting growing support on the need to develop an international standard allowing to overcome the semantic interrogability problems in use cases of OHAS.

Provided the use case prerequisites are satisfied, then the main activities during the exchange of OHAS content execute in the following order:

1. The employee/worker contacts an OHP in Country B.
2. The OHP is identified, authenticated and authorized.
3. The employee/worker is identified by the OHP to have OHAS data in Country A.
4. The OHP informs the employee/worker what part of the OHAS will be retrieved.
5. The employee/worker authorizes the OHP to extract and use data from his OHAS.
6. The OHP provides the employee/worker with the requested occupational health service.

An essential quality requirement in the use case execution is the correct semantic interpretation of the clinical aspect contained in the OHAS regarding language, medical diagnoses and conclusions, terminologies and vocabularies, measurement units. The following subsection proposes a methodology that transforms this quality requirement into a software solution.

2.2. Common OHAS dataset model

The standardization of both the OHAS dataset structure and contents is an essential prerequisite for achieving interoperability among the heterogeneous information systems participating in the exchange of data. In order to develop the common information model of the OHAS dataset, we reviewed the existing practices in the area of occupational health of EU countries like France, Bulgaria and

OCCUPATIONAL HEALTH ASSESSMENT SUMMARY		
Health summary content status		
Personal data		
	Employment history	
		Employment details
		Workplace hazards
		Preliminary medical exams
		Preliminary obligatory medical exam
		Periodic medical exams
		Periodic obligatory medical exams
		Illness/Diability/Accident data
		Individual exams
		Conclusions

Figure 2. The major data sections of the OHAS dataset

Belgium [22, 26]. Thus, the obtained OHAS dataset is built on the data discovered typically in the national legislation of these countries, although such data are found in these documents in terms of different notations. A high level hierarchical organization of the major data sections is represented in Figure 2.

The root of this hierarchy contains data fields describing the identity of the employee/worker in terms of Personal data (identity code, names, local address and general practitioner contact details) and a conditional field (health summary content status) controlling access to this dataset (Health summary content status). An ordered list below this root is used to store the Employment history where each record of the Employment history has nine sections of data:

Employment details. It is a required section with data fields describing the employer's contact details, job position and the period for employment of the employee/worker.

Workplace factors for health disorders. This section has a brief description of the job activities and a list of the factors in the workplace environment that might cause health disorders. There are several internationally recognized factors such as noise, vibrations, lightning at the workplace, dust, chemical agents, biological agents, ionizing and nonionizing radiation etc. The levels of these factors at the workplace are qualified as Recommended exposure limits, Threshold limit values (close to the upper recommended exposure) or Immediately dangerous to life limit values (significantly exceeding the recommended exposure) [8]. With respect to exposure duration each such factor is identified as long-term (repeated exposure) or short-term (accidental exposure). The development of a common job classification system relating job encodings to known hazards in typical workplace settings would provide a standardized framework for cross-border interoperability in management of this kind of data.

Preliminary medical exams. These exams are represented in **two sections**, a section for obligatory or a section for nonobligatory preliminary exams. The obligatory exams are done by a commission of experts in occupational health and concludes with a statement for fitness to execute job activities or to be exempted from executing certain activities, while the nonobligatory exams are done by a single occupational health professional who issues a diagnose confirming or rejecting an occupational disease. These sections of OHAS make references to diagnoses and symptoms for occupational diseases.

Periodic medical exams. Data about the periodic exams is recorded at regular intervals during the employment of the employee/worker. Similarly by content to the preliminary medical exam sections, there are **two separate sections** dedicated to recording periodic nonobligatory and periodic obligatory exams.

Illness/Disability/Accident data. This section records documented events of this nature registered during the employment period.

Individual exams. This section contains records about medical exams undertaken by the employee/worker himself.

Conclusions. This section contains records of positive/negative conclusions about the ability to execute job activities done by commissions on occasions different than obligatory preliminary and periodic exams.

It is noteworthy, that the composition of the OHAS dataset shown in Figure 2 can be extended to a folder of OHAS generated in a single country. Thus, in the implementation of the primary use case in Subsection 2.1, one folder can be dedicated to store OHAS data composed in Country A and eventually, multiple folders for representing compositions of OHAS data in case of employment in a set of countries referred to as Country B in that use case.

2.3. OHAS archetype in EN 13606

Unlike other application domains, the healthcare domain requires clinical data to be exchanged together with the semantic context in which the data values are created and remain valid. The correct interpretation of the semantic context in the OHAS dataset is vital for the evaluation of the worker's health state, determining a diagnosis or identifying symptoms for related complications and, in particular, for proactive prevention of occupational diseases. Therefore, the implementation of the use case for exchange of OHAS must rely on standards and technologies supporting the application of common vocabularies and terminologies for representing the semantic context in the process of communication between different computer systems. The most notable example for such a standard is EN 13606 [17, 18].

EN 13606 is a complete international standard for semantic interoperability in the exchange of extracts of electronic health data such as OHAS. This standard uses a dual architecture model and object oriented technologies, combining a static Reference information model (RM) with an Archetype object model (AOM) quite similar to those discovered in the openEHR specification [24]. The classes in its reference model, shown in Figure 3, are employed to represent the non-volatile structure of clinical concepts, while the archetype model serves to express specific knowledge constraints on instances of the underlying reference model [4].

In contrast to the alternative implementation of this approach in FHIR [13], all the EN 13606 archetype instances are validated against one and the same RM. Additional advantages of EN 13606 archetypes are their support for visualization, composition, specialization and redefinition, enabling reusability and participation of clinical domain experts in their development. Updating clinical content in EN 13606 requires no change in the RM rather adding new or extending existing archetypes.

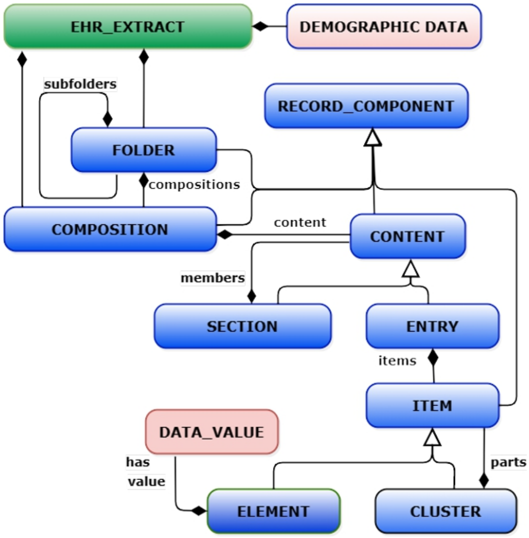


Figure 3. UML class diagram of the EN 13606 Reference model

In FHIR it means the development of new software for each new resource. Therefore, the EN 13606 standard is selected in this paper as the primary means for building a long-term open health platform for cross-border exchange of OHAS.

The EN 13606 archetype of the OHAS shown in Figure 4 allows to accurately represent the constraints in the clinical domain that are discovered in the use case described in Subsection 2.1. It is a COMPOSITION archetype developed employing LinkEHR Studio [19, 28]. The archetype is composed of an ENTRY for Personal data and a SECTION holding the Employment history. The Employment history is represented in terms of an ordered sequence SECTIONs with ENTRY datatypes describing the employment from the occupational health perspective outlined in Subsection 2.2.

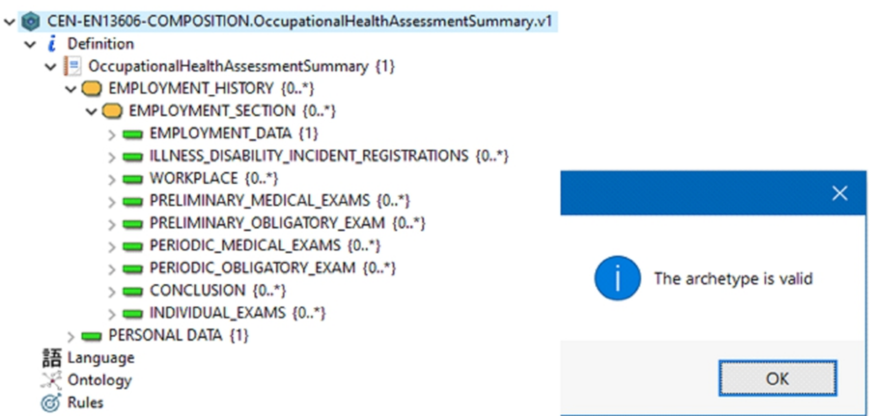


Figure 4. The EN 13606 archetype of OHAS

An important feature of LinkEHR Studio is that it allows to validate the archetype with respect to the RM datatypes of EN 13606 as well as in relation to the datatypes of the AOM. These are two kinds of datatypes that serve different purposes.

The purpose of the AOM datatypes is to wrap attributes and values of the RM datatypes with properties allowing to impose on them constraints discovered in the application domain. Besides, both kinds of datatypes are expressed in terms of primitive XML datatypes like Integer, Real, String, Boolean or Date/Time/DateTime employing the respective XML schemes accompanying the EN 13606 standard.

Important constraints that are directly related to achieving semantic interoperability of OHAS in the clinical domain (occupational health) are bindings to systemized nomenclatures as SNOMED-CT (Figure 5), international classifications of diseases as ICD-10, encodings of hazardous factors in the workplace environment or measurement system standards for identification of units used to describe quantitative exposure assessment values. Other examples are the cardinality constraints in AOM on instances of classes from the RM, shown in curled brackets in Figures 4 and 5, or using regular expressions, assumed and default values for controlling the intended clinical meaning of values stored in the instances of that archetype.

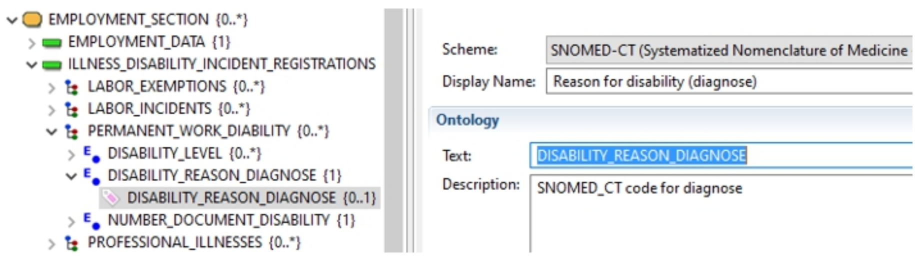


Figure 5. Binding diagnoses to SNOMED-CT in the OHAS archetype

3. Results

The creation of valid instances of EN 13606 archetypes is an essential requirement for using the archetypes in practice. It is common to obtain these instances in XML format. Besides, the archetype instances must be valid with respect to the XML scheme of classes and datatypes in EN 13606 like EN 13606-RM.xsd depicted in Figure 6. Unlike openEHR, building instances of EN 13606 archetypes is not a straightforward process because it involves the development of custom software utilities. Note that validating the archetype instance by taking in consideration just the primitive datatypes in the XML standard does not make it a valid EN 13606 instance. This difficulty appears to be one of the major obstacles for a broader recognition of EN 13606 in exchanging extracts of electronic health records.

Another important decision is the choice of the platform for management of the EN 13606 instances of the OHAS archetype like the one displayed in Figure 6. At

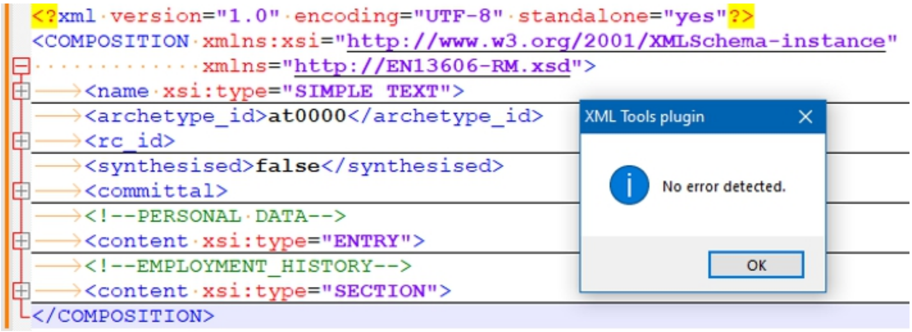


Figure 6. Archetype instance of OHAS validated against EN 13606 classes and datatypes

this point ,we can take advantage of the proven validation of the archetype instance against the XML scheme of the EN 13606 standard RM and employ as a platform a native XML database like BaseX [3].

This database provides access to data stored in XML format through RESTful web services. Besides, management of XML data is done in terms of XQuery as a substitute for the absence of support for a query language in EN 13606. The static non-volatile nature of the OHAS archetype translates to static non-volatile structure of all the instances of that archetype. Therefore, likewise the EN 13606 archetype, the XPath-s in XQueries are also reusable. It allows to create a stable API for management of XML data in native XML databases.

In order to prove this concept, we have developed a three-tier application comprising BaseX, an Apache 2 web server and a desktop client application. The application demonstrates the execution of the main activities in the OHAS use case such as exporting and storing valid EN 13606 instances of OHAS. Figure 7 demonstrates a case study for management of real-life data by means of RESTful web services for accessing OHAS instances stored in BaseX.

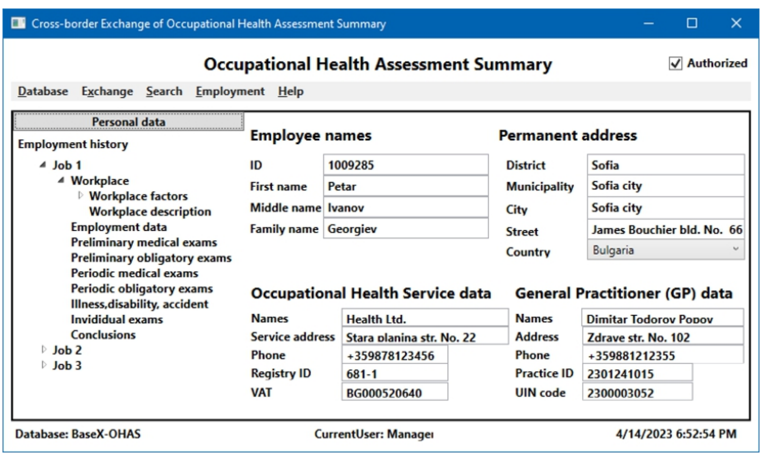


Figure 7. Case study: Data management of exchanged EN 13606 instances of OHAS

4. Conclusion

The exchange of OHAS is essential requirement for proactive prevention of occupational diseases as well as for ensuring the health of employees/workers in executing job activities. Exercising professional activities in most cases is accompanied with the development of occupational diseases that in certain cases lead to disabilities. Treatment of occupational disease costs a lot to the industry, the national budget and above all, causes sufferings and loss of jobs. Therefore, regular medical exams and early discovery of symptoms is a key prevention strategy. Apparently, changing workplaces might cause deterioration of worker's health, provided records from previous medical exams are not taken into consideration, especially, when the employee has been exposed to harmful health factors in former workplace environments. The purpose of OHAS is to reduce the uncertainty in making decisions about fitness to execute job activities or in evaluating the health state of the employee/worker.

The globalization of world economy stimulates a large number of people to search abroad for better employment opportunities. This is the reason to consider the exchange of OHAS content from cross-border view point. It is a problem of larger social-significance than the cross-border exchange for International Patient Summary (IPS). The IPS standard concerns most of all people travelling abroad, while the OHAS affects all the employed people. While the IPS dataset is studied better within the framework of an international standard, the OHAS remains rather insufficiently explored in the existing literature.

This paper proposes a novelty systematic approach to provide a practicable solution to this problem by ensuring semantic interoperability in the exchange of OHAS. A description of the use case for cross-border exchange is pre3sented. Next, the OHAS dataset is defined on the basis of exploring common data requirements the national legislation of EU countries. The final step of this approach is the design of an EN 13606 archetype of OHAS satisfying all the requirements for semantic interoperability in the exchange of clinical data. Further on, the static, non-volatile and reusable information model of OHAS is used to create EN 13606 instances that are valid with respect to the RM and the datatypes of this standard. Finally, this paper demonstrates the software implementation of basic activities in the OHAS use case in a fully functional web application, where the EN 13606 instances are managed by a native XML database.

Acknowledgements

This work is supported by the Fund for Scientific Research at Sofia University "St. Kliment Ohridski" under grant 4.80-10-59/09.04.2024.

References

[1] American Chemical Society, CAS registry, 2023. [Online]. Available: https://www.cas.org/. [Accessed 18 February 2024].

[2] Australian Office of Parliamentary Counsel, Work Health and Safety Act 2011 and the Work Health and Safety (Transitional and Consequential Provisions) Act 2011, Canberra, Federal Register of Legislation, 2021.

[3] BaseX, The XML Framework. Lightweight and High-Performance Data Processing, https://basex.org/, 2023. [Accessed 11 February 2024].

[4] T. Beale, Archetypes: Constraint-based domain models for future-proof information systems, in: 11th OOPSLA Workshop on Behavioral Semantics: Serving the Customer, Boston, 2002, 16–32.

[5] Bureau of Labor Statistics, Economic news release: Employer-reported workplace injuries and illnesses, 2021, U.S. Bureau of Labor Statistics, 2022.

[6] CEN/TC 251, EN ISO 27269-2021 – The International Patient Summary Standard, Brussels, European Committee for Standardization, 2022.

[7] J. Euchner, Occupational Hazards, Research-Technology Management 57(2) (2015) 9–10.

[8] European Commission, Consolidated text: Commission Directive 2009/161/EU of 17 December 2009 establishing a third list of indicative occupational exposure limit values in implementation of Council Directive 98/24/EC and amending Commission Directive 2000/39/EC, Official Journal of the European Communities, vol. L 27, p. 115, 31 January 2017.

[9] European Commission, Patient Summary. Guideline on the electronic exchange of health data under Cross-Border Directive 2011/24/EU, eHealth Network, 2022.

[10] Haute Autorité de Santé, Le dossier médical en santé au travail, HAS/Service des bonnes pratiques professionnelles (https://www.has-sante.fr), 2009 (in French).

[11] Health and Safety Executive, Health and safety at work. Summary statistics for Great Britain 2022, 2022. [Online]. Available: https://www.hse.gov.uk/statistics/assets/docs/hssh2122.pdf. [Accessed 30 October 2024].

[12] A. B. Hill, The environment and disease: association or causation?, J. R. Soc. Med. 108(1) (2015) 32–37.

[13] HL 7 FHIR, FHIR R4. International Patient Summary Implementation Guide, Health Level Seven International – Patient Care Work Group, 2022.

[14] International Labour Organization, List with occupational diseases, State Gazette (Darzhaven Vestnik), Bulgaria, vol. 66, 25 July 2008 (in Bulgarian).

[15] International Labour Organization, List of occupational diseases (revised 2010), Occupational Health and Safety Series No. 74, Geneva, International Labour Office, 2010.

[16] International Labour Organization, NATLEX. Database of national labour, social security and related human rights legislation, International Labour Organization (ILO), 2023. [Online]. Available: https://www.ilo.org/dyn/natlex/. [Accessed 11 February 2024].

[17] ISO, ISO/EN 13606-1:2019(en): Health informatics-Electronic health record communication-Part 1: Reference model, International Organization for Standardization, 2019.

[18] ISO, ISO/EN 13606-2:2019(en): Health informatics-Electronic health record communication-Part 2: Archetype interchange specification, International Organization for Standardization, 2019. [Online]. Available: https://sabs.isolutions.iso.org/obp/ui#iso:std:iso:13606:-2:ed-2:v1:en. [Accessed 19 February 2024].

[19] E. Krastev, EN 13606 Archetype of Occupational Health Assessment Summary, 2023. [Online]. Available: https://github.com/euggit/EN13606-archetypes. [Accessed 11 February 2024].

[20] E. Krastev, S. Abanos and D. Tcharaktchiev, Health data exchange based on archetypes of clinical concepts, in: Information Systems & Grid Technologies: 15th International Conference ISGT'2022, Sofia, May 27–28, 2022, 98–112.
[21] J. Leigh, Economic burden of occupational injury and illness in the United States, Milbank Quarterly 89(4) (2011) 728–772.
[22] Ministry of Health, Ordinance No 3 of 25.01.2008 on the terms and conditions for carrying out the activities of the occupational health services, State Gazette (Darzhaven Vestnik), Bulgaria, vol. 14, 2 December 2008 (in Bulgarian).
[23] V. Nestorova, Z. Stoyneva and G. Prakova, Occupational diseases, Varna, Medical University Press, 2020.
[24] openEHR, Archetype Object Model 2 (AOM2), 19 February 2024. [Online]. Available: https://specifications.openehr.org/releases/AM/latest/AOM2. [Accessed 11 February 2024].
[25] F. Pega, B. Náfrádi, N. C. Momen et al., Global, regional and national burdens of ischemic heart disease and stroke attributable to exposure to long working hours for 194 countries, 2000–2016: A systematic analysis from the WHO/ILO Joint Estimates of the Work-related Burden of Disease and Injury, Environment International 154 (2021) 1–15.
[26] A. Pourchel, Dossier médical en santé au travail: un outil essentiel pour le suivi individuel, Labour and Safety 835 (2023) 44–46 (in French).
[27] D. Tcharaktchiev, E. Krastev, P. Petrossians, S. Abanos, H. Kyurkchiev and P. Kovatchev, Cross-border exchange of clinical data using archetype concepts compatible with the international patient summary, Stud. Health Technol. Inform. 270 (2020) 552–556.
[28] VeraTech for Health, LinkEHR Studio, 2022. [Accessed 11 February 2024].
[29] WHO, Basic documents, 49 ed., World Health Organization, 2020.
[30] World Health Organization, WHO/ILO joint estimates of the work-related burden of disease and injury, 2000–2016: global monitoring report, World Health Organization and International Labour Organization, 2021.

Received on January 9, 2024
Accepted on March 5, 2024

Evgeniy Krastev, Petko Kovachev and Simeon Abanos

Faculty of Mathematics and Informatics
Sofia University "St. Kliment Ohridski"
5, James Bourchier Blvd.
1164 Sofia
BULGARIA

E-mails: eck@fmi.uni-sofia.bg
az@petko.info
simeonabanos@gmail.com

Dimitar Tcharaktchiev

Medical Faculty, Medical University of Sofia
2, Zdrave Str.
1431 Sofia
BULGARIA

E-mail: dimitardt@gmail.com