Preface

Óbuda University has an intensive research cooperation in engineering high tech fields including mechanical, electric and electronic engineering, materials science, robotics, optimal control and informatics. All these activities are related to or involve applied and industrial mathematics research as well. This volume is a selection of 12 papers that contain either new results in applied mathematics or use mathematics to solve an important application problem. Most of the papers are written by staff members of Óbuda University and/or their research partners from all over the world. The topics of presented results vary from graph theory to fuzzy decision making. The authors indeed use a wide spectrum of mathematical methods for their investigation. The first paper by T. K. Pogány is related to sampling of stochastic L^2-processes. The paper of J. Abaffy and A. Galántai gives a new method for global Lipschitz optimization and related numerical experiments. The third paper of this volume is writen by R. Briggs and P. T. Nagy who derive a classification of sub-Riemannian manifolds and also give an application to an invariant optimal control problem. The paper of A. Baricz and T. K. Pogány presents monotonicity and convexity properties for the one dimensional regularization of the Coulomb potential and gives Turán type inequalities used in some applications. A. Kristály and S. Nagy investigate the existence of Stackelberg equilibrium in games defined on manifolds. The paper of J. Abaffy and S. Fodor presents a new method for solving mixed-integer problems by applying the ABS approach to Gomory's cutting plane algorithm.

Authors D. H. Hoang, M. Benes and T. Oberhuber develop a numerical simulation of anisotropic mean curvature of graphs in the context of relative geometry. T. Réti and D. Dimitrov's paper compares various irregularity measures for bidegreed graphs. The paper of D. L. Debeljovic, S. B. Stojanovic and A. M. Jovanovic gives a condition of algebraic character for the finite-time stability of linear time-delay systems, while the work of K. R. Hedrig and L. Veljovic gives a new description of kinetic pressures on shaft bearings of a rigid body nonlinear dynamics. The paper of C. O. Morariu and S-M. Zaharia suggest a new method for reliability testing. The work of P. Rezaei, K. Rezaie, S. Nazari-Shirkouhi and M- R- J. Tajabadi gives an interesting application of fuzzy decision making to allocate an underground dam to improve water management.

Aurél Galántai, Péter T. Nagy

Special Issue Guest Editors

Bessel–sampling restoration of stochastic signals

Tibor K. Pogány

Óbuda University, John von Neumann Faculty of Informatics, Institute of Applied Mathematics, Bécsi út 96/b, Budapest Hungary
e-mail: tkpogany@gmail.com *and*

University of Rijeka, Faculty of Maritime Studies, Studentska 2, Rijeka, Croatia
e-mail: poganj@pfri.hr

Abstract: *The main aim of this article is to establish sampling series restoration formulae in for a class of stochastic L^2-processes which correlation function possesses integral representation close to a Hankel-type transform which kernel is either Bessel function of the first and second kind J_ν, Y_ν respectively. The results obtained belong to the class of irregular sampling formulae and present a stochastic setting counterpart of certain older results by Zayed* [25] *and of recent results by Knockaert* [13] *for J–Bessel sampling and of currently established Y–Bessel sampling results by Jankov Maširević* et al. [7]. *The approach is twofold, we consider sampling series expansion approximation in the mean–square (or L^2) sense and also in the almost sure (or with the probability* 1*) sense. The main derivation tools are the Piranashvili's extension of the famous Karhunen–Cramér theorem on the integral representation of the correlation functions and the same fashion integral expression for the initial stochastic process itself, a set of integral representation formulae for the Bessel functions of the first and second kind J_ν, Y_ν and various properties of Bessel and modified Bessel functions which lead to the so–called Bessel–sampling when the sampling nodes of the initial signal function coincide with a set of zeros of different cylinder functions.*

Keywords: *Almost sure convergence, Bessel functions of the first and second kind J_ν, Y_ν, correlation function, harmonizable stochastic processes, Karhunen–Cramér–Piranashvili theorem, Karhunen processes, Kramer's sampling theorem, mean–square convergence, sampling series expansions, sampling series truncation error upper bound, spectral representation of correlation function, spectral representation of stochastic process.*

MSC(2010): 42C15, 60G12, 94A20.

1 Introduction

The development and application of sampling theory in technics, engineering but in parallel in pure mathematical investigations was rapid and continuous since the middle of the 20th century [4, 6, 9, 15, 16, 17]. It is one of the most important mathematical techniques used in communication engineering and information theory, and it is also widely represented in many branches of physics and engineering, such as signal analysis, image processing, optics, physical chemistry, medicine etc. [9, 25]. In general sampling theory can be used where functions need to be restored from their discretized–measured–digitalized sampled values, usually from the values of the functions and/or their derivatives at certain points. Here we are focused to a kind of Bessel–sampling restoration of finite second order moment stochastic processes (signals), which correlation function possesses Hankel–transform type integral representation. In the Bessel sampling procedure the sampling nodes we take to be the positive zeros $j_{\nu,k}, y_{\nu,k}$ of the Bessel functions J_ν, Y_ν respectively, depending on the appearing Bessel function in the kernel of the integral expression representing the correlation function of the considered initial stochastic signal.

The results obtained form a stochastic setting counterpart to recent results by Zayed [25, 26, 24, 27], Knockaert [13] and Jankov *et al.* [7].

This paper is organized as follows: in the sequel we give a short account in correlation and spectral theory of stochastic signals, which consists from a necessary introductionary knowledge about different kind stochastic processes appearing in the engineering literature together with associated mathematical models. Secondly, J–Bessel and Y–Bessel deterministic sampling theorems are recalled together with their ancestor result, that is the Kramer's sampling theorem. In Section 2 we prove our main results on the Bessel sampling restoration of stochastic signals in both mean–square and almost sure manner. Finally, we proceed with restoration error analysis, presenting associated results in finding the uniform upper bounds for newly derived truncated sampling series, which is a counterpart of deterministic results which has been considered in a number of publications in the mathematical literature, consult for instance [7, 8, 9] and the appropriate references therein. In Conclusion section we give an overview of the exposed matter together with new research directions and improvement possibilities. The exhaustive references list finishes the exposition.

1.1 Brief invitation to correlation theory of stochastic processes

Let $(\Omega, \mathfrak{A}, \mathsf{P})$ a standard fixed probability space and consider the random variables $\xi\colon T \times \Omega \mapsto \mathbb{C}, T \subseteq \mathbb{R}$; the double–indexed infinite family of random variables $\{\xi(t) \equiv \xi(t,\omega)\colon t \in \mathbb{T}, \omega \in \Omega\}$ is a *stochastic process*. Here T is the *index set* of the process ξ. Denote $L^2(\Omega, \mathfrak{A}, \mathsf{P})$ [abbreviated to $L^2(\Omega)$ in the sequel] be the space of all finite second order complex–valued random variables defined on $(\Omega, \mathfrak{A}, \mathsf{P})$, equipped with the norm $\sqrt{\mathsf{E}|\cdot|^2} := \|\cdot\|_2$, where E means the expectation operator. Notice that $L^2(\Omega)$ is a Hilbert–space with the inner (or scalar) product $\mathsf{E}\xi\,\overline{\eta}$ endowed. However, it is enough to restrict ourselves to the linear mean–square–closure $\mathscr{H}_t(\xi) := \overline{\{L^2\{\xi(s)\colon s \leq t\}}$ spanned by all finite linear combinations and/or their *in medio* limits generated by the family $\{\xi(s)\colon s \leq t\}, t \in \mathbb{R}$, which is the linear subspace of the Hilbert space $L^2(\Omega)$. It is well-known that $\mathscr{H}_\infty(\xi) \equiv \mathscr{H}(\xi)$ possesses also a Hilbert–space structure, keeping the norm and inner product of

$L^2(\Omega)$. We recall that when $\bigcap_{t\in\mathbb{R}} H_t(\xi) = \emptyset$, then ξ is *purely indeterministic*[1], say; moreover in the case $\bigcap_{t\in\mathbb{R}} H_t(\xi) = \mathscr{H}(\xi)$, process ξ is *purely deterministic*[2].

The function $m_\xi(t) = \mathsf{E}\xi(t)$ is the *expectation function*. Let us assume throughout that the considered stochastic processes are centered, that is $m_\xi(t) \equiv 0, t \in \mathbb{R}$[3]. The function $B_\xi(t,s) = \mathsf{E}\xi(t)\overline{\xi(s)}$ is the *correlation function* (or autocorrelation function) of the centered process ξ at two "times values" $t,s \in T$. By the Cauchy–Buniakovskiy-Schwarz inequality it is straightforward that

$$|B_\xi(t,s)|^2 \le B_\xi(t,t)B_\xi(s,s), \qquad t,s,\in T\,, \tag{1}$$

being ξ with the finite second order moment rv, with any fixed $t \in T$. The function $\mathsf{D}\xi(t) := B_\xi(t,t)$ is the *variance* of the process ξ[4].

Very wide class of stochastic processes has been introduced by Piranashvili [18]. He has studied the sampling reconstruction of a class of nonstationary processes, which correlation function (and *a fortiori* the initial process itself) possess spectral representations in a form of a double integral. In fact Piranashvili extended the Karhunen-Cramér theorem for a wider class stochastic processes; see the works of Karhunen [11] and Cramér [3], also see [29, p. 156].

Theorem A. [Karhunen–Cramér–Piranashvili Theorem] *Let a centered stochastic $L^2(\Omega)$–process ξ has correlation function (associated to some domain $\Lambda \subseteq \mathbb{R}$ with some sigma–algebra $\sigma(\Lambda)$) in the form:*

$$B(t,s) = \int_\Lambda \int_\Lambda f(t,\lambda)\overline{f(s,\mu)}\,F_\xi(\mathrm{d}\lambda,\mathrm{d}\mu), \tag{2}$$

with analytical exponentially bounded kernel function $f(t,\lambda)$, while F_ξ is a positive definite measure on $\mathbb{R}^2$ provided the total variation $\|F_\xi\|(\Lambda,\Lambda)$ of the spectral distribution function *F_ξ satisfies*

$$\|F_\xi\|(\Lambda,\Lambda) = \int_\Lambda \int_\Lambda \big|F_\xi(\mathrm{d}\lambda,\mathrm{d}\mu)\big| < \infty.$$

Then, the process $\xi(t)$ has the spectral representation as a Lebesgue integral

$$\xi(t) = \int_\Lambda f(t,\lambda) Z_\xi(\mathrm{d}\lambda); \tag{3}$$

in (2) *and* (3)

$$F_\xi(S_1,S_2) = \mathsf{E}Z_\xi(S_1)\overline{Z_\xi(S_2)}, \qquad S_1,S_2 \subseteq \sigma(\Lambda),$$

and vice versa.

[1] In the Western terminology; however, according to the Eastern, Soviet/Russian probabilistic terminology this kind process is *regular*.

[2] Singular. It is worth to mention that we deal here with a class of weakly stationary singular processes.

[3] Otherwise we pick up the so–called centered process $\xi_0(t) = \xi(t) - m_\xi(t)$, which expectation function is obviously zero.

[4] By (1) we see, that $\mathsf{D}\xi(t) \le \sup_{u\in\mathbb{R}} B_\xi^2(u,u) := \mathfrak{B}_\xi^2 < \infty$.

Note that in the case of finite Λ we will talk on processes *bandlimited to* Λ.

If F_ξ of (2) concentrates of diagonal $\lambda = \mu$, that is $F_\xi(\lambda,\mu) = \delta_{\lambda,\mu} F_\xi(\lambda)$, then the resulting correlation is called of *Karhunen class*, and B_ξ becomes

$$B_\xi(t,s) = \int_\Lambda f(t,\lambda)\overline{f(s,\lambda)}F_\xi(d\lambda).$$

The spectral representation of the resulting *Karhunen process* $\xi(t)$ remains of the form given by (3).

Also, putting $f(t,\lambda) = e^{it\lambda}$ in (2) one gets the *Loève-representation*:

$$B(t,s) = \int_\Lambda\int_\Lambda e^{i(t\lambda - s\mu)}F_\xi(\mathrm{d}\lambda,\mathrm{d}\mu).$$

Then, the Karhunen process with the Fourier kernel $f(t,\lambda) = e^{it\lambda}$ we recognize as the *weakly stationary stochastic process* having covariance

$$B(\tau) = \int_\Lambda e^{i\tau\lambda}F_\xi(\mathrm{d}\lambda), \qquad \tau = t - s.$$

The stochastic processes having correlation function expressible in the form (2) we call *harmonizable*. Further reading about different kind harmonizabilities present the works [10, 20, 21] and the appropriate references therein. Finally, when $\Lambda = (-w,w)$ for some finite $w > 0$ in this considerations, we get the *band–limited* variants of the above introduced processes. So, for $\xi(t)$, being weak sense stationary band–limited to $w > 0$, there holds the celebrated Whittaker–Kotel'nikov–Shannon sampling theorem:

$$\xi(t) = \sum_{k\in\mathbb{Z}} \xi\Big(\frac{\pi}{w}k\Big)\,\frac{\sin(wt - k\pi)}{wt - k\pi}\,, \tag{4}$$

uniformly convergent on all compact t–subsets of $\mathbb{R}$, in both mean–square and almost sure sense; the latter has been proved by Belyaev [2].

1.2 Kramer's theorem and Bessel sampling

Here we recall three theorems which will help us to derive our first set of Bessel sampling restoration results for a class of harmonizable stochastic processes having Karhunen representable correlation functions.

Theorem B. [Kramer's Theorem], [12, 13] *Let $K(x,t)$ be in $L^2[a,b], -\infty < a < b < \infty$ a function of x for each real number t and let $E = \{t_k\}_{k\in\mathbb{Z}}$ be a countable set of real numbers such that $\{K(x,t_k)\}_{k\in\mathbb{Z}}$ is a complete orthogonal family of functions in $L^2[a,b]$. If*

$$f(t) = \int_a^b g(x)K(x,t)\,\mathrm{d}x,$$

for some $g \in L^2[a,b]$, then f admits the sampling expansion

$$f(t) = \sum_{k\in\mathbb{Z}} f(t_k)\, S^\star(t,t_k)\,,$$

where

$$S^\star(t,t_k) = \frac{\displaystyle\int_a^b K(x,t)\overline{K(x,t_k)}\,\mathrm{d}x}{\displaystyle\int_a^b |K(x,t_k)|^2\,\mathrm{d}x}\,.$$

Remark 1. *Annaby reported, that points $\{t_k\}_{k\in\mathbb{Z}}$, which are for practical reasons preferred to be real, can also be complex,* [1, p. 25].

Obviously, the function f, having above integral representation property bandlimited to the region $\Lambda = [a,b]$.

Now we give the two Bessel–sampling theorems, the J–Bessel derived e.g. by Zayed [25, p. 132], but the J–Bessel sampling method was known already by Whittakers [22, 23], Helms and Thomas [5] and Yao [30].

Theorem C. *It there is some $G \in L^2(0,b)$ with a finite Hankel–transform*

$$f(\lambda) = \frac{2^\nu\,\Gamma(\nu+1)}{b^{\nu+\frac{1}{2}}\lambda^\nu} \int_0^b G(x)\sqrt{x}J_\nu(x\lambda)\,\mathrm{d}x\,, \tag{5}$$

then there holds

$$f(t) = \frac{2J_\nu(bt)}{b^\nu z_{,}t^\nu} \sum_{k\geq 1} \frac{j_{\nu,k}^{\nu+1}\, f(a^{-1}\, j_{\nu,k})}{(b^2t^2 - j_{\nu,k}^2)\,J_\nu'(j_{\nu,k})}\,,$$

where the series converges uniformly on any compact subset of the complex t–plane. Here λ_k denote the kth zero of $J_\nu(b\sqrt{\lambda})$.

In turn the Y–Bessel sampling theorem has been recently derived by Jankov Maširević *et al.* in [7, p. 81, Theorem 4].

Theorem D. *Let for some $G \in L^2(0,a), a > 0$, function f possesses a finite Hankel–transform*

$$f(t) = \int_0^a G(x)\sqrt{x}Y_\nu(tx)\,\mathrm{d}x\,, \tag{6}$$

then, for all $t \in \mathbb{R}$, $\nu \in [0,1)$, the function f admits the sampling expansion

$$f(t) = 2Y_\nu(at)\sum_{k\geq 1} f(b^{-1}y_{\nu,k})\,\frac{y_{\nu,k}}{(y_{\nu,k}^2 - a^2t^2)\,Y_{\nu+1}(y_{\nu,k})}\,,$$

where $y_{\nu,k}$, $k \in \mathbb{N}$ are the positive real zeros of the Bessel function $Y_\nu(t)$. Here the convergence os uniform in all compact t–subsets of $\mathbb{C}$.

2 Main results

Although formula (4), Theorem C and Theorem D yield an explicit restoration of bandlimited weakly stationary stochastic process $\xi(t)$ by the WKS sampling theorem, and Hankel-transformable $f(t)$ by either J–Bessel or Y–Bessel sampling procedures respectively, these results are usually considered to be of theoretical interest only, because the restoration procedures require computations of infinite sums. In practice, we truncate the sampling expansion series. The sampling size N is determined by the relative error accepted in the reconstruction. Thus the error analysis plays a crucial role in setting up the interpolation formula, and it is of considerable interest to find sampling series truncation error upper bounds (the exact value of the truncation error is in general a "mission impossible") which vanishes with the growing sampling size.

Here and in what follows we will concentrate to a class of harmonizable stochastic processes having spectral representation of the form (3) with the kernel function

$$f(t,\lambda)\in L^2(0,b), \qquad b>0\,,$$

with respect to the time–parameter t.

According to these requirements, we introduce the notations for both kind Bessel sampling procedures:

$$\begin{aligned} \mathscr{S}_N^J(\mathfrak{G};t) &:= \frac{2J_\nu(bt)}{b^\nu t^\nu}\sum_{k=1}^N \frac{j_{\nu,k}^{\nu+1}\,\mathfrak{G}(b^{-1}j_{\nu,k})}{(b^2t^2-j_{\nu,k}^2)\,J_\nu'(j_{\nu,k})}\\ \mathscr{S}_N^Y(\mathfrak{G};t) &:= 2Y_\nu(bt)\sum_{k=1}^N \frac{y_{\nu,k}\,\mathfrak{G}(b^{-1}y_{\nu,k})}{(y_{\nu,k}^2-b^2t^2)\,Y_{\nu+1}(y_{\nu,k})}\,, \end{aligned}$$

for the truncated (partial) Bessel sampling series expansions either of $L^2(0,b)$–bandlimited signal f, or for the stochastic process ξ, that is $\mathfrak{G}\in\{f,\xi\}$. Next, we introduce the sampling series restoration truncation error, read as follows

$$\begin{aligned} \mathscr{T}_N^J(\mathfrak{G};t) &:= \mathfrak{G}(t)-\mathscr{S}_N^J(\mathfrak{G};t) = \frac{2J_\nu(bt)}{b^\nu t^\nu}\sum_{k\ge N+1} \frac{j_{\nu,k}^{\nu+1}\,\xi(b^{-1}j_{\nu,k})}{(b^2t^2-j_{\nu,k}^2)\,J_\nu'(j_{\nu,k})} \qquad (7)\\ \mathscr{T}_N^Y(\mathfrak{G};t) &:= \mathfrak{G}(t)-\mathscr{S}_N^Y(\mathfrak{G};t) = 2Y_\nu(bt)\sum_{k\ge N+1} \frac{y_{\nu,k}\,\xi(b^{-1}y_{\nu,k})}{(y_{\nu,k}^2-b^2t^2)\,Y_{\nu+1}(y_{\nu,k})}\,, \end{aligned}$$

Our main goal in that stage of investigation is to establish as sharp as possible mean square truncation error upper bounds in both Bessel–sampling procedures, that is for

$$\Delta_N^{\mathscr{B}}(\xi;t) = \mathsf{E}\big|\xi(t)-\mathscr{S}_N^{\mathscr{B}}(\xi;t)\big|^2 = \mathsf{E}\big|\mathscr{T}_N^{\mathscr{B}}(\xi;t)\big|^2, \qquad \mathscr{B}\in\{J,Y\}\,.$$

Firstly, we establish the spectral representation formula for $\mathscr{S}_N^J(\xi;t)$.

Theorem 1. *Let $\xi(t), t \in T \subseteq \mathbb{R}$ a harmonizable stochastic process of Piranashvili class, that is*

$$\xi(t) = \int_\Lambda f(t,\lambda)\, Z_\xi(\mathrm{d}\lambda)$$

with the kernel function $f(t,\lambda) \in L^2(0,b)$ with respect to t and any fixed $\lambda \in \Lambda$. Then we have

$$\mathscr{S}_N^{\mathscr{B}}(\xi;t) = \int_\Lambda \mathscr{S}_N^{\mathscr{B}}(f;t)\, Z_\xi(\mathrm{d}\lambda), \qquad \mathscr{B} \in \{J, Y\}.$$

Moreover, there holds true

$$\mathscr{T}_N^{\mathscr{B}}(\xi;t) = \int_\Lambda \mathscr{T}_N^{\mathscr{B}}(f;t)\, Z_\xi(\mathrm{d}\lambda), \qquad \mathscr{B} \in \{J, Y\};$$

both formulae are valid in the mean square sense.

Proof. The sampling series expansion of the kernel function $f(t,\lambda)$ which appears in the representation (3), when truncated to the terms indexed by N becomes $\mathscr{S}_N^J(f;t)$. Now, by (7) we get

$$\begin{aligned}
\mathscr{S}_N^J(\xi;t) &= \frac{2J_\nu(bt)}{b^\nu t^\nu} \sum_{k=1}^N \frac{j_{\nu,k}^{\nu+1}\, \xi(b^{-1} j_{\nu,k})}{(b^2t^2 - j_{\nu,k}^2) J_\nu'(j_{\nu,k})} \\
&= \frac{2J_\nu(bt)}{b^\nu t^\nu} \sum_{k=1}^N \frac{j_{\nu,k}^{\nu+1}}{(b^2t^2 - j_{\nu,k}^2) J_\nu'(j_{\nu,k})} \int_\Lambda f(a^{-1} j_{\nu,k}, \lambda) Z_\xi(\lambda) \\
&= \int_\Lambda \left\{ \frac{2J_\nu(bt)}{b^\nu t^\nu} \sum_{k=1}^N \frac{j_{\nu,k}^{\nu+1}}{(b^2t^2 - j_{\nu,k}^2) J_\nu'(j_{\nu,k})} f(b^{-1} j_{\nu,k}, \lambda) \right\} Z_\xi(\lambda);
\end{aligned}$$

here all equalities are in the mean square sense used. This is exactly the statement for $\mathscr{B} = J$. The case of Y–Bessel sampling we handle in the same way.

The second assertion we prove directly:

$$\begin{aligned}
\mathscr{T}_N^J(\xi;t) = \xi(t) - \mathscr{S}_N^J(\xi;t) &= \int_\Lambda f(t,\lambda) Z_\xi(\mathrm{d}\lambda) - \int_\Lambda \mathscr{S}_N^J(f;t) Z_\xi(\mathrm{d}\lambda) \\
&= \int_\Lambda \left\{ f(t,\lambda) - \mathscr{S}_N^J(f;t) \right\} Z_\xi(\mathrm{d}\lambda) \\
&= \int_\Lambda \mathscr{T}_N^J(f;t) Z_\xi(\mathrm{d}\lambda).
\end{aligned}$$

The equalities are also in the mean square sense used. The rest is clear. □

Theorem 2. *Let the situation be the same as in* Theorem 1. *Then we have*

$$\Delta_N^{\mathscr{B}}(\xi;t) = \int_\Lambda \int_\Lambda \mathscr{T}_N^{\mathscr{B}}(f;t) \overline{\mathscr{T}_N^{\mathscr{B}}(f;t)} F_\xi(\mathrm{d}\lambda, \mathrm{d}\mu), \qquad \mathscr{B} \in \{J, Y\}, \tag{8}$$

in the mean square sense.

The proof is a straightforward consequence of the Karhunen–Cramér–Piranshvili Theorem A and the spectral representation formulae of stochastic process ξ, therefore we omit it.

Remark 2. Obviously Theorem 2 is devoted to the case of Piranashvili processes. For the Karhunen processes this result reduces to

$$\Delta_N^{\mathscr{B}}(\xi;t) = \int_\Lambda \big|\mathscr{T}_N^{\mathscr{B}}(f;t)\big|^2 F_\xi(\mathrm{d}\lambda), \qquad \mathscr{B} \in \{J,Y\}. \tag{9}$$

Denote $L^2(\Lambda;F_\xi)$ the class of square–integrable on the support domain Λ, complex functions with respect to the measure $F_\xi(\mathrm{d}\lambda)$, i.e.

$$L^2(\Lambda;F_\xi) := \left\{\varphi\colon \int_\Lambda |\varphi|^2 F_\xi(\mathrm{d}\lambda) < \infty\right\}.$$

This class form also a Hilbert–space and the correspondence $\xi(t) \longleftrightarrow f(t,\lambda)$ defines an isomorphism between $\mathscr{H}(\xi)$ and $L^2(\Lambda;F_\xi)$. Therefore by the existing isometry, we conclude (9).

Next, a special case of the Karhunen process is the weakly stationary stochastic process[5]. Choosing $\Lambda = (-w,w)$, we arrive at

$$\Delta_N^{\mathscr{B}}(\xi;t) = \int_{-w}^{w} |\mathscr{T}_N^{\mathscr{B}}(\mathrm{e}^{\mathrm{i}t\lambda})|^2 F_\xi(\mathrm{d}\lambda), \qquad \mathscr{B} \in \{J,Y\}.$$

Now, we are ready to state our Bessel–sampling series finding for stochastic processes.

Theorem 3. *Let* $\{\xi(t)\colon t \in \mathbb{T} \subseteq \mathbb{R}\}$ *a Piranashvili process* (3) *with a kernel function* $f(t,\lambda) \in L^2(0,b)$ *which possesses a Hankel–transform representation either of the form* (5) (*J–Bessel sampling*) *or* (6) (*Y–Bessel sampling*). *Then we have*

$$\xi(t) = \mathscr{S}^J(\xi;t) = \frac{2J_\nu(bt)}{b^\nu t^\nu} \sum_{k\geq 1} \frac{j_{\nu,k}^{\nu+1}\,\xi(b^{-1}j_{\nu,k})}{(b^2t^2 - j_{\nu,k}^2)J'_\nu(j_{\nu,k})}$$

$$\xi(t) = \mathscr{S}^Y(\xi;t) = 2Y_\nu(bt) \sum_{k\geq 1} \frac{y_{\nu,k}\,\xi(b^{-1}y_{\nu,k})}{(y_{\nu,k}^2 - b^2t^2)Y_{\nu+1}(y_{\nu,k})},$$

respectively. Both equalities hold in the mean square sense.

Proof. Having in mind that (8)

$$\Delta_N^{\mathscr{B}}(\xi;t) = \mathsf{E}|\mathscr{T}_N^{\mathscr{B}}(\xi:t)|^2 = \int_\Lambda\int_\Lambda \mathscr{T}_N^{\mathscr{B}}(f;t)\,\overline{\mathscr{T}_N^{\mathscr{B}}(f;t)}\,F_\xi(\mathrm{d}\lambda,\mathrm{d}\mu),$$

and $\mathscr{T}_N^{\mathscr{B}}(f;t)$ vanishes pointwise and uniformly [25, p. 132] (J–Bessel sampling), that is [7, p. 83, Theorem 4] (Y–Bessel sampling) with the growing N, we deduce

$$\lim_{N\to\infty} \Delta_N^{\mathscr{B}}(\xi;t) = 0, \qquad \mathscr{B} \in \{J,Y\},$$

which completes the proof. □

[5] Also known as *stationary in the Khintchin sense*.

3 Truncation error bounds for Y–Bessel sampling of Karhunen processes

In this section we would derive uniform upper bound for the truncation error for the Y–Bessel sampling expansion of the Karhunen process $\xi(t), t \in T \subseteq \mathbb{R}$:

$$\mathscr{S}^Y(\xi;t) = 2Y_\nu(t)\sum_{k=1}^{N} \frac{y_{\nu,k}\,\xi(y_{\nu,k})}{(y_{\nu,k}^2 - t^2)\,Y_{\nu+1}(y_{\nu,k})}\,,$$

setting for the sake of simplicity $b = 1$, $\nu \in [0,1)$ and the function f has a band–region contained in $(0,1)$. Having in mind (9) exposed in Remark 2, we specify:

$$\Delta_N^Y(\xi;t) = \int_\Lambda \left|\mathscr{T}_N^Y(f;t)\right|^2 F_\xi(\mathrm{d}\lambda)\,. \tag{10}$$

The truncation error upper bound has been already calculated in under the polynomial decay condition (see e.g. [14])

$$|f(t)| \le \frac{A}{|t|^{r+1}}, \qquad A>0,\, r>0\,, t \neq 0. \tag{11}$$

The corresponding truncation error upper bound [7, p. 83, Theorem 5] for all

$$\nu \in [0,1), \quad t \in (\nu, y_{\nu,2}), \quad \min\{A,r\} > 0, \quad N \ge 2$$

reads as follows

$$\mathscr{T}_N^Y(f;t) < \frac{2AH(t)M_N(\nu)}{\pi^2 L_{N+1}(\nu)} := U_N^Y(t)\,,$$

where

$$\begin{aligned} H(t) &= 1 + \frac{2t}{\pi(t^2-\nu^2)} \\ M_N(\nu) &= \exp\left\{\left(N + \frac{1-\pi+2(\nu - y_{\nu,2})}{2\pi}\right)^{-1}\right\} - 1 \\ L_{N+1}(\nu) &= \frac{2}{\sqrt{\pi}}\, y_{\nu,N+1}^r \left\{\frac{y_{\nu,N+1}^2 - (2\nu+3)(2\nu+7)}{(4y_{\nu,N+1} - \nu - 1)^{\frac{3}{2}} + \mu^*}\right\}^{\frac{1}{2}} \end{aligned}$$

and $\mu^* = (2\nu+3)(2\nu+5)$.

Moreover, for any fixed $t \in (\nu, y_{\nu,2})$ and growing N the following the asymptotic behavior results holds [7, p. 83, Eq. (15)]

$$\mathscr{T}_N^Y(f;t) = \mathscr{O}\left(N^{-r-\frac{5}{4}}\right).$$

Now, we are ready to formulate our next main result.

Theorem 4. *Let $\xi(t), t \in \mathbb{R}$ a Karhunen process with the kernel function f satisfying polynomial decay condition* (11). *Then for all $\nu \in [0,1)$, for all $t \in (\nu, y_{\nu,2})$,* $\min\{A,r\} > 0$ *and all $N \geq 2$, we have*

$$\Delta_N^Y(\xi;t) \leq \frac{A^2\,\|F_\xi\|(\Lambda)\,(\pi\nu t+2)^2\,[(4y_{\nu,N+1}-\nu-1)^{\frac{3}{2}}+(2\nu+3)(2\nu+5)]}{\pi^5\,\nu^2\,t^2\,y_{\nu,N+1}^{2r}[y_{\nu,N+1}^2-(2n+3)(2n+7)]} \times\left(\exp\left\{\left(N+\frac{1-\pi+2(\nu-y_{\nu,2})}{2\pi}\right)^{-1}\right\}-1\right)^2,$$

where $\|F_\xi\|(\Lambda)$ stands for the total variation of the spectral distribution function F_ξ.

Moreover, the decay magnitude of the truncation error is

$$\Delta_N^Y(\xi;t) = \mathscr{O}\left(N^{-2r-\frac{5}{2}}\right). \tag{12}$$

Proof. Because of the spectral representation formula (10) and the functional truncation error upper bound (11) by Jankov Maširević *et al.* we have

$$\Delta_N^Y(\xi;t) = \int_\Lambda \left|\mathscr{T}_N^Y(f;t)\right|^2 F_\xi(\mathrm{d}\lambda) \leq \int_\Lambda \left|U_N^Y(f;t)\right|^2 F_\xi(\mathrm{d}\lambda).$$

Now routine calculations lead to the statement. Relation (12) is the immediate consequence of this upper bound result. □

Next, we consider the almost sure convergence in the Y–Bessel sampling series restoration of the Karhunen process.

Theorem 5. *Let $\xi(t)$ a Karhunen process with the kernel function f satisfying polynomial decay condition* (11). *Then for all $\nu \in [0,1)$, for all $t \in (\nu, y_{\nu,2})$,* $\min\{A,r\} > 0$ *and all $N \geq 2$, we have*

$$\mathsf{P}\left\{\lim_{N\to\infty} \mathscr{S}_N^Y(\xi;t) = \xi(t)\right\} = 1.$$

Proof. Firstly, for some positive ε we evaluate the probability

$$\mathsf{P}_N = \mathsf{P}\left\{\left|\xi(t) - \mathscr{S}_N^Y(\xi;t)\right| \geq \varepsilon\right\}.$$

Applying the Čebyšev inequality, then Theorem 3 we conclude th estimate

$$\mathsf{P}_N \leq \varepsilon^{-2}\,\mathsf{E}\left|\mathscr{T}_N^Y(\xi;t)\right|^2 = \mathscr{O}\left(N^{-2r-\frac{5}{2}}\right).$$

For certain enough large absolute constant C the following bound follows in terms of the Riemann Zeta function:

$$\sum_{N\geq 2} \mathsf{P}_N \leq C \sum_{N\geq 2} N^{-2r-\frac{5}{2}} = C\left[\zeta\left(2r+\tfrac{5}{2}\right)-1\right],$$

and the series converges, being $r > 0$. Now, by the Borel–Cantelli lemma it follows the a.s. convergence, which completes the proof. □

4 Final remarks

In the footnote 2 it was mentioned that we work throughout with singular, or purely deterministic processes. Indeed, having in mind that the initial input process of Piranashvili type $\xi(t)$ possesses spectral representation (3) in which the kernel function is a Hankel transform of some convenient $G \in L^2(0,b)$, we deduce

$$\begin{aligned}\xi(t) &= \int_\Lambda f(t,\lambda)\, Z_\xi(\mathrm{d}\lambda)\\ &= \frac{2^\nu \Gamma(\nu+1)}{b^{\nu+\frac{1}{2}}} \int_\Lambda \left\{ \frac{1}{\lambda^\nu} \int_0^b G(x)\sqrt{x} J_\nu(x\lambda)\, \mathrm{d}x \right\} F_\xi(\mathrm{d}\lambda)\\ &= \frac{2^\nu \Gamma(\nu+1)}{b^{\nu+\frac{1}{2}}} \int_0^b G(x)\sqrt{x} \left\{ \int_\Lambda \frac{J_\nu(x\lambda)}{\lambda^\nu} F_\xi(\mathrm{d}\lambda) \right\} \mathrm{d}x\\ &= \frac{2^\nu \Gamma(\nu+1)}{b^{\nu+\frac{1}{2}}} \int_0^b G(x)\sqrt{x}\, \Psi_\nu(x)\, \mathrm{d}x.\end{aligned}$$

Obviously $\xi(t)$ is bandlimited to $(0,b)$. (We mention that the sample function $\xi(t) \equiv \xi(t,\omega_0)$ and $f(t,\lambda)$ possess the same exponential types [2, Theorem 4], [18, Theorem 3], and also by the Paley–Wiener theorem we conclude that $\xi(t)$ is bandlimited to the support set $(0,b)$).

The Kolmogorov–Krein analytical singularity criterion states that the singular processes possesses divergent integral:

$$\int_{\mathbb{R}} \frac{\log \frac{\mathrm{d}}{\mathrm{d}\lambda} F_\xi(\mathrm{d}\lambda)}{1+\lambda^2}\, \mathrm{d}\lambda = -\infty,$$

which is obviously true, being the Radon–Nikodým derivative (or in other words spectral density) in the integrand equal to zero on a set of positive Lebesgue measure.

References

[1] M. H. Annaby, Sampling Integrodifferential Transforms Arising from Second Order Differential Operators, *Math. Nachr.* 216 (2000), 25–43.

[2] Yu. K. Belyaev, Analytical random processes, *Teor. Verojat. Primenen.* **IV** (1959), No. 4, 437–444. (in Russian)

[3] H. Cramér, A contribution to the theory of stochastic processes, *Proc. Second Berkely Symp. Math. Stat. Prob.*, 329–339, 1951.

[4] B. Draščić, Sampling Reconstruction of Stochastic Signals - The Roots in the Fifties, *Austrian J. Statist.* **36** (2007), No.1, 65–72.

[5] H. D. Helms, J. B. Thomas, On truncation error of sampling theorem expansion, *Proc. I.R.E.* **50** (1962), 179–184.

[6] J. R. HIGGINS, *Sampling Theory in Fourier and Signal Analysis*, Clarendon Press, Oxford, 1996.

[7] D. JANKOV MAŠIREVIĆ, T. K. POGÁNY, Á. BARICZ, A. GALÁNTAI, Sampling Besssel functions and Bessel sampling, Proceedings of IEEE 8th International Symposium on Applied Computational Intelligence and Informatics, May 23-25, Timişoara (2013) 79-84.

[8] A. J. JERRI, I. A. JOSLIN, Truncation error for the generalized Bessel type sampling series, *J. Franklin Inst.*, **314** (1982), No. 5, 323–328.

[9] A. J. JERRI, The Shannon sampling theoremits various extensions and applications: A tutorial review, *Proc. IEEE* **11** (1977), 1565-1596.

[10] Y. KAKIHARA, *Multidimensional Second Order Stochastic Processes*, World Scientific, Singapore, 1997.

[11] K. KARHUNEN, Über lineare Methoden in der Wahrscheinlichkeitsrechnung, *Ann. Acad. Sci. Fennicae, Ser A, I* **37** (1947), 3–79.

[12] H. P. KRAMER, A generalized sampling theorem, *J. Math. Phys.* **38** (1959), 68–72.

[13] L. KNOCKAERT, A class of scaled Bessel sampling theorems, *IEEE Trans. Signal Process.* **59** (2011), No. 11, 5082–5086.

[14] X. M. LI, Uniform Bounds for Sampling Expansions, *Journal of approximation theory* **93** (1998), 100–113.

[15] A. Y. OLENKO, T. K. POGÁNY, Time shifted aliasing error upper bounds for truncated sampling cardinal series, *J. Math. Anal. Appl.* **324** (2006), 262–280.

[16] A. Y. OLENKO, T. K. POGÁNY, Universal truncation error upper bounds in irregular sampling restoration, *Appl. Anal.* **90(3–4)** (2011), 595–608.

[17] A. Y. OLENKO, T. K. POGÁNY, Average sampling reconstruction od harmonizable processes, *Commun. Stat. Theor. Methods* **40** (2011) , No. 19-20, 3587-3598.

[18] Z.A. PIRANASHVILI, On the problem of interpolation of random processes, *Teor. Ver. Primenen.* **XII** (1967), No. 4, 708–717. (in Russian)

[19] T. POGÁNY, Almost sure sampling reconstruction of band–limited stochastic signals, Chapter 9. in *Sampling Theory in Fourier and Signal Analysis. Advanced Topics*, Higgins, J.R. & Stens, R.L. (eds.), Oxford University Press, Oxford, 209-232, 284–286, 1999.

[20] M. B. PRIESTLEY, *Non–Linear and Non–Stationary Time Series*, Academic Press, London, New York, 1988.

[21] M. M. RAO, Harmonizable processes: structure theory. *Einseign. Math. (2)* **28**(1982), No. 3–4, 295–351.

[22] E. WHITTAKER, On the functions which are represented by the expansion of the interpolation theory, *Proc. Roy. Soc. Edinburgh Sect. A* **35** (1915), 181-194.

[23] J. WHITTAKER, *Interpolatory Function Theory*, Cambridge University Press, Cambridge, 1935.

[24] A. I. ZAYED, On Kramer's sampling theorem associated with general Sturm-Liouville problems and Lagrange interpolation, SIAM *J. Appl. Math.* **51** (1991), 575-604.

[25] A. I. ZAYED, *Advances in Shannon's Sampling Theory*, CRC Press, New York, 1993.

[26] A. I. ZAYED, A proof of new summation formulae by using sampling theorems, *Proc. Amer. Math. Soc.* **117(3)** (1993), 699–710.

[27] A. ZAYED, G. HINSEN, P. BUTZER, On Lagrange interpolation and Kramer-type sampling theorems associated with Sturm-Liouville problems, SIAM *J. Appl. Math.* **50** (1990), 893-909.

[28] A. M. YAGLOM, *Correlations Theory of Stationary and Related Random Function I. Basic Results*, Spriner–Verlag, 1987.

[29] A. M. YAGLOM, *Correlations Theory of Stationary and Related Random Function II. Supplementary Notes and References*, Spriner–Verlag, 1987.

[30] K. YAO, Application of reproducing kernel Hilbert spaces–band-limited signal models, *Inf. Contr.* **11** (1967), 427–444.

An always convergent algorithm for global minimization of univariate Lipschitz functions

József Abaffy

Institute of Applied Mathematics, Óbuda University, H-1034 Budapest, Bécsi út 96/b, Hungary
abaffy.jozsef@nik.uni-obuda.hu

Aurél Galántai

Institute of Applied Mathematics, Óbuda University, H-1034 Budapest, Bécsi út 96/b, Hungary
galantai.aurel@nik.uni-obuda.hu

Abstract: We develop and analyze a bisection type global optimization algorithm for real Lipschitz functions. The suggested method combines the branch and bound method with an always convergent solver of nonlinear equations. The computer implementation and performance are investigated in detail.

Keywords: global optimum; nonlinear equation; always convergent method; Newton method; branch and bound algorithms; Lipschitz functions

1 Introduction

In paper [2] we defined the following branch and bound method to find the global minimum of the problem

$$\begin{aligned}&f(z) \to \min\\&l \le z \le u,\end{aligned}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is sufficiently smooth and $l, u \in \mathbb{R}^n$. Assume that

$$z_{output} = \texttt{alg\_min}(f, z_{input})$$

is a local minimization algorithm that satisfies $f(z_{output}) \le f(z_{input})$, for any z_{input}. Similarly, assume that

$$[z_{sol}, iflag] = \texttt{equation\_solve}(f, c)$$

denotes a solution algorithm of the single multivariate equation $f(z) = c$ such that $iflag = 1$, if a true solution z_{sol} exists (that is $f(z_{sol}) = c$), and $iflag = -1$, otherwise.

Let $f_{\min}$ denote the global minimum of f, and let $B_{lower} \in \mathbb{R}$ is a lower bound of f such that $f_{\min} \geq B_{lower}$. Let $z_0 \in D_f$ be any initial approximation to the global minimum point ($f(z_0) \geq B_{lower}$). The suggested algorithm of [2] then takes the form:

Algorithm 1

```
$z_1 = \texttt{alg\_min}(f, z_0)$
$a_1 = f(z_1)$, $b_1 = B_{lower}$, $i = 1$
**while** $a_i - b_i > tol$
    $c_i = (a_i + b_i)/2$
    $[\xi, iflag] = \texttt{equation\_solve}(f, c_i)$
    **if** $iflag = 1$
        $z_{i+1} = \texttt{alg\_min}(f, \xi)$, $a_{i+1} = f(z_{i+1})$, $b_{i+1} = b_i$
    **else**
        $z_{i+1} = z_i$, $a_{i+1} = a_i$, $b_{i+1} = c_i$
    **end**
    $i = i + 1$
**end**
```

Using the idea of Algorithm 1 we can also determine a lower bound of f, if such a bound is not known a priori (for details, see [2]). Algorithm 1 shows conceptual similarities with other multidimensional bisection type algorithms such as those of Shary [34] and Wood [50], [52].

Theorem 1. Assume that $f : \mathbb{R}^n \to \mathbb{R}$ is continuous and bounded from below by B_{low}. Then Algorithm 1 is globally convergent in the sense that $f(z_i) \to f_{\min}$.

Proof. At the start we have z_1 and the lower bound b_1 such that $f(z_1) \geq f_{\min} \geq b_1$. Then we take the midpoint of this interval, i.e. $c_1 = (f(z_1) + b_1)/2$. If a solution ξ exists such that $f(\xi) = c_1$ ($iflag = 1$), then $c_1 \geq f_{\min}$ holds. For the output z_2 of the local minimizer, the inequality $c_1 \geq f(z_2) \geq f_{\min} \geq b_1$ holds by the initial assumptions. If there is no solution of $f(\xi) = c_1$ (i.e. $iflag = -1$), then $c_1 < f_{\min}$. By continuing this way we always halve the inclusion interval $(b_i, f(z_i))$ at the worst case. So the method is convergent in the sense that $f(z_i) \to f_{\min}$. Note that sequence $\{z_i\}$ is not necessarily convergent. □

The practical implementation of Algorithm 1 clearly depends on the local minimizer, the equation solver and also on f. Since we have several local minimizers satisfying the above requirements we must concentrate on the equation solvers. There are essentially two questions to be dealt with. Namely, the existence of the solution and the very existence of methods that are always convergent in the sense that either they give a solution when exists or give a warning sign if no solution exists.

The existence of solution follows from the Weierstrass theorem, if $f_{\min} \leq c \leq f(z_0)$. As for the solvers we may observe that for $n > 1$, our equation is an underdetermined nonlinear equation of the form

$$g(z) = f(z) - c = 0 \quad (g : \mathbb{R}^n \to \mathbb{R}). \tag{1}$$

There are several locally convergent methods for such equations (see, e.g. [25], [3], [45], [26], [27], [28], [47], [48], [12], [13], [14]). In paper [2] we tested Algorithm 1 with a nonlinear Kaczmarz projection algorithm [45], [26], [27], [25], which showed fast convergence in most of the test cases, but also showed numerical instability in some cases, when $\nabla f(z_k)$ was close to zero.

There also exist always convergent methods for equation (1) (see, e.g. [37], [9], [20], [22], [21], [43], [44], [1], [31], [46]). For the multivariate case, most methods are related to subdivision and seem to be quite slow. For univariate equations, however, the always convergent methods of Szabó [43], [44], Abaffy and Forgó [1], Pietrus [31] and Várterész [46] are using other principles than subdivision and they are quite fast.

Here we study Algorithm 1 for one-dimensional real Lipschitz functions. The global minimization of real Lipschitz functions has a rich literature with many interesting and useful algorithms. For these, we refer to Hansen, Jaumard, Lu [15], [17], [18] and Pintér [32].

The outline of paper is the following. We develop and analyze the equation solver in Section 2. In Section 3 we develop a modified implementation of Algorithm 1 called Algorithm 2 that use this equation solver and double bisection. The final section contains the principles and results of numerical testing. The comparative numerical testing indicates that Algorithm 2 can be a very efficient minimizer in practice.

2 An always convergent solver for real equations

Consider the real equation

$$g(t) = 0 \quad (g : \mathbb{R} \to \mathbb{R}, \, t \in [\alpha, \beta]) \tag{2}$$

An iterative solution method of the form $x_{n+1} = F(g; x_n)$ is said to be always convergent, if for any $x_0 \in [\alpha, \beta]$ $(g(x_0) \neq 0)$

(i) the sequence $\{x_n\}$ is monotone,

(ii) $\{x_n\}$ converges to the zero in $[\alpha,\beta]$ that is nearest to x_0, if such zero exists,

(iii) if no such zero exists, then $\{x_n\}$ exits the interval $[\alpha,\beta]$.

Assuming high order differentiability, Szabó [43], [44] and Várterész [46] developed some high order always convergent iterative methods. Assuming only continuous differentiability Abaffy and Forgó [1] developed a linearly convergent method, which was generalized to Lipschitz functions by Pietrus [31] using generalized gradient in the sense of Clarke.

Since we assume only the Lipschitz continuity of g, we select and analyze an always convergent modification of the Newton method. This method was first investigated by Szabó [43], [44]) under the condition that g is differentiable and bounded in the interval $[\alpha,\beta]$. We only assume that g satisfies the Lipschitz condition.

Theorem 2. (a) Assume that $|g(t)-g(s)| \leq M|t-s|$ holds for all $t,s\in[\alpha,\beta]$. If $x_0\in(\alpha,\beta]$ and $g(x_0)\neq 0$, then the iteration

$$x_{n+1}=x_n-\frac{|g(x_n)|}{M}\quad (n=0,1,\dots) \tag{3}$$

either converges to the zero of g that is nearest left to x_0 or the sequence $\{x_n\}$ exits the interval $[\alpha,\beta]$. (b) If $y_0\in[\alpha,\beta)$ and $g(y_0)\neq 0$, then the iteration

$$y_{n+1}=y_n+\frac{|g(y_n)|}{M}\quad (n=0,1,\dots) \tag{4}$$

either converges to the zero of g that is nearest right to y_0 or the sequence $\{y_n\}$ exits the interval $[\alpha,\beta]$.

Proof. We prove only part (a). The proof of part (b) is similar. It is clear that $x_{n+1}\leq x_n$. If a number γ exists such that $\alpha\leq\gamma\leq x_0$ and $x_n\to\gamma$, then $g(\gamma)=0$. Otherwise there exists an index j such that $x_j<\alpha$. Assume now that $\alpha\leq\gamma<x_0$ is the nearest zero of g to x_0. Also assume that $\gamma\leq x_n$ $(n\geq 1)$. We can write

$$x_{n+1}-\gamma=x_n-\gamma-\frac{|g(x_n)-g(\gamma)|}{M}=\left(1-\frac{\xi_n}{M}\right)(x_n-\gamma)\quad (\xi_n\in[0,M]). \tag{5}$$

Since $0\leq 1-\frac{\xi_n}{M}\leq 1$, we obtain that $\gamma\leq x_{n+1}$ and $x_{n+1}-\gamma\leq x_n-\gamma$. Hence the method, if converges, then converges to the nearest zero to x_0. Assume that no zero exists in the interval $[\alpha,x_0]$ and let $|g|_{\min}=\min_{\alpha\leq t\leq x_0}|g(t)|$. Then

$$x_{n+1}=x_n-\frac{|g(x_n)|}{M}\leq x_n-\frac{|g|_{\min}}{M}\leq x_0-(n+1)\frac{|g|_{\min}}{M},$$

and algorithm (3) leaves the interval in at most $\frac{M(x_0-\alpha)}{|g|_{\min}}$ steps. A similar claim holds for algorithm (4). □

The convergence speed is linear in a sense. Assume that $\alpha \leq \gamma < x_0$ is the nearest zero to x_0 and $\varepsilon > 0$ is the requested precision of the approximate zero. Also assume that a number $m_\varepsilon > 0$ exists such that $m_\varepsilon |t-\gamma| \leq |g(t)| \leq M|t-\gamma|$ holds for all $\gamma+\varepsilon \leq t \leq x_0$. If g is continuously differentiable in $[\alpha,\beta]$, then $m_\varepsilon = \min_{t\in[\gamma+\varepsilon,x_0]} |g'(t)|$. Having such a number m_ε we can write (5) in the form

$$x_n - \gamma \leq \left(1-\frac{m_\varepsilon}{M}\right)^n (x_0-\gamma) \leq \left(1-\frac{m_\varepsilon}{M}\right)^n (\beta-\alpha).$$

This indicates a linear speed achieved in at most $\left\lceil \frac{\log\frac{\varepsilon}{\beta-\alpha}}{\log\left(1-\frac{m_\varepsilon}{M}\right)} \right\rceil$ steps. We can assume that $m_\varepsilon > \varepsilon$, which gives the bound $\left\lceil \frac{\log\frac{\varepsilon}{\beta-\alpha}}{\log\left(1-\frac{\varepsilon}{M}\right)} \right\rceil$. Relation $\log(1+\varepsilon) \approx \varepsilon$ yields the approximate expression $M\left|\log\frac{\varepsilon}{\beta-\alpha}\right|\varepsilon^{-1}$ for the number of required iterations.

For the optimum step number of algorithms in the class of Lipschitz functions, see Sukharev [42] and Sikorski [35].

Assume now that $L > 0$ is the smallest Lipschitz constant of g on $[\alpha,\beta]$ and $M = L+c$ with a positive c. It then follows from (5) that

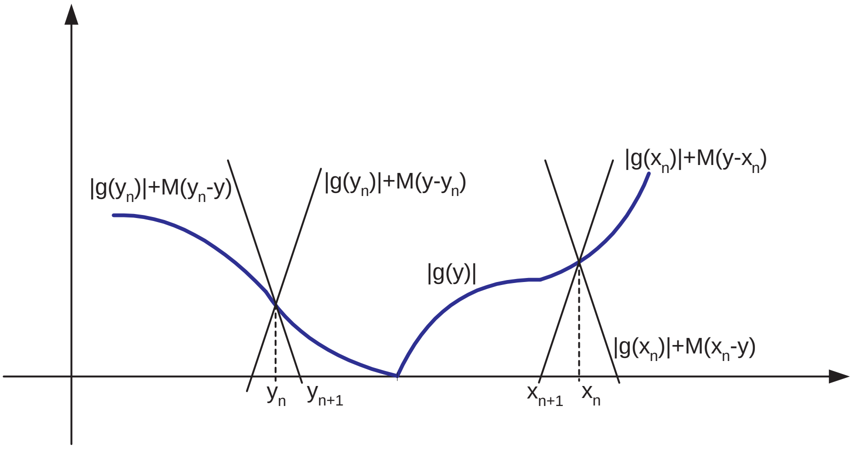
$$x_{n+1} - \gamma \geq \left(1-\frac{L}{L+c}\right)(x_n-\gamma) = \left(\frac{c}{L+c}\right)^{n+1}(x_0-\gamma).$$

This indicates a linear decrease of the approximation error. Note that the method can be very slow, if $c/(L+c)$ is close to 1 (if M significantly overestimates L) and it can be fast, if $c/(L+c)$ is close to 0 (if M is close to L). Equation (5) also shows that M can be replaced in the algorithms (3)-(4) by an appropriate M_n that satisfies the condition $0 \leq \frac{\xi_n}{M_n} \leq 1$. For differentiable g, M_n might be close to $|g'(x_n)|$ in order to increase the speed (case of small c).

A simple geometric interpretation shows that the two algorithms are essentially the same. The Lipschitz condition implies that $||g(t)|-|g(s)|| \leq M|t-s|$ $(t,s\in[\alpha,\beta])$ also holds. The resulting inequality

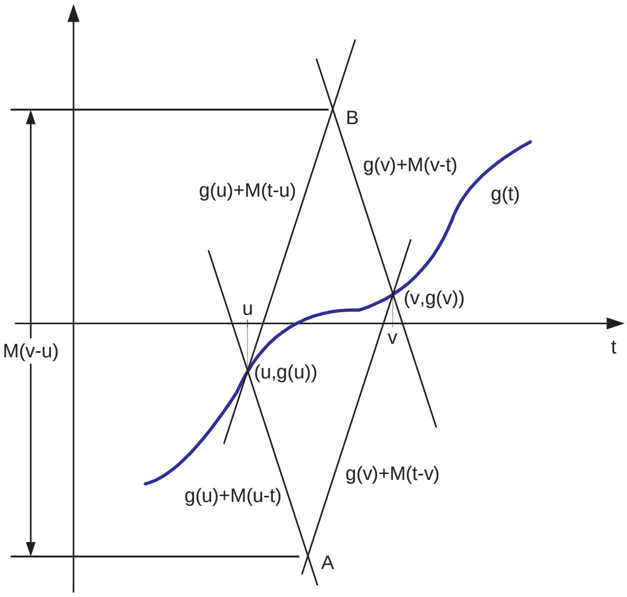
$$|g(x)| - M|x-t| \leq |g(t)| \leq |g(x)| + M|x-t|$$

gives two linear bounding functions for $|g(t)|$, namely $|g(x)|+M(x-t)$ and $|g(x)|+M(t-x)$ for a fixed x. If the zero γ is less than x_n, then for $t \leq x_n$, the linear function $|g(x_n)|+M(t-x_n)$ will be under $|g(t)|$. Its zero $x_{n+1} = x_n - \frac{|g(x_n)|}{M} \leq x_n$ is the next approximation to γ and $x_{n+1} \geq \gamma$ clearly holds. Similarly, if $y_n < \gamma$, then $|g(y_n)|+M(y_n-t)$ will be under $|g(t)|$ and for its zero, $y_n \leq y_{n+1} = y_n + \frac{|g(y_n)|}{M} \leq \gamma$ clearly holds. The next figure shows both situations with respect to an enclosed unique zero γ.



It also follows that if $g(x_0) > 0$ ($g(x_0) < 0$) then $g(t) > 0$ ($g(t) < 0$) for $\gamma < t \leq x_0$, if such a zero γ exists. If not, $g(t)$ keeps the sign of $g(x_0)$ in the whole interval $[\alpha, x_0]$. An analogue result holds for algorithm (4).

Consider the following general situation with arbitrary points $u, v \in [\alpha, \beta]$ ($u < v$).



The points $(u, g(u))$ and $(v, g(v))$ and the related linear bounding functions define a parallelogram that contains function g over the interval $[u, v]$ with the bounds

$$\frac{g(u)+g(v)}{2} + M\frac{u-v}{2} \leq g(t) \leq \frac{g(u)+g(v)}{2} + M\frac{v-u}{2} \quad (u \leq t \leq v).$$

This property is the basis of Piyavskii's minimization algorithm and related methods (see, e.g. [17], [32]). It is also exploited in Sukharev's modified bisection method [41], [42].

Function $g(t)$ may have a zero in $[u,v]$ only if

$$g(u)+g(v)+M(u-v)\leq 0\leq g(u)+g(v)+M(v-u),$$

that is if

$$|g(u)+g(v)|\leq M(v-u). \tag{6}$$

If $g(t)$ has a zero $\gamma\in(u,v)$, then by the proof of Theorem 2.

$$u+\frac{|g(u)|}{M}\leq\gamma\leq v-\frac{|g(v)|}{M} \tag{7}$$

holds and (6) is clearly satisfied. If u and v are close enough and (u,v) does not contain a zero of $g(t)$, then (6) does not hold. This happens, if $u\geq v-\frac{|g(v)|}{M}$ and $g(u)\neq 0$ or $v\leq u+\frac{|g(u)|}{M}$ and $g(v)\neq 0$.

Note that iterations (3)-(4) satisfy the bounds

$$\frac{g(x_{n+1})+g(x_n)-|g(x_n)|}{2}\leq g(t)\leq\frac{g(x_{n+1})+g(x_n)+|g(x_n)|}{2} \tag{8}$$

for $x_{n+1}\leq t\leq x_n$, and the bounds

$$\frac{g(y_{n+1})+g(y_n)-|g(y_n)|}{2}\leq g(t)\leq\frac{g(y_{n+1})+g(y_n)+|g(y_n)|}{2} \tag{9}$$

for $y_n\leq t\leq y_{n+1}$.

Note also that if u and v are distant enough (in a relative sense), then condition (6) may hold without having a zero in (u,v).

Using the above geometric characterization we can develop practical exit conditions for the nonlinear solver (3)-(4). The most widely used exit conditions are $|x_{n+1}-x_n|<\varepsilon$ and $|g(x_n)|<\varepsilon$, which are not fail safe neither individually nor in the combined form $\max\{|x_{n+1}-x_n|,|g(x_n)|\}<\varepsilon$. For a thorough analysis of the matter, see Delahaye [8], Sikorski and Wozniakowski [36] and Sikorski [35]. Another problem arises in the floating precision arithmetic that requires stopping, if either $|x_{n+1}-x_n|<\varepsilon_{machine}$ or $|g(x_n)|<\varepsilon_{machine}$ holds. Since $|x_{n+1}-x_n|=\frac{|g(x_n)|}{M}$, the tolerance precision ε is viable, if $\max\{1,M\}\varepsilon_{machine}<\varepsilon$. By the same argument the tol parameter of Algorithm 1 must satisfy the lower bound $tol\geq 2\varepsilon_{machine}$.

If $g(t)$ has a zero $\gamma\in[\alpha,x_0)$, the monotone convergence of $\{x_n\}$ implies the relation $|x_{n+1}-x_n|\leq|x_n-\gamma|$. Hence $|x_{n+1}-x_n|$ is a lower estimate of the approximation error.

There are some possibilities to increase the reliability of the combined exit condition. The first one uses algorithm (4) in the following form. If interval $(x_n-\varepsilon,x_n)$

is suspect to have a zero of $g(t)$ (and $g(x_n-\varepsilon), g(x_n)\neq 0$), then we can apply condition (6) with $u=x_n-\varepsilon$ and $v=x_n$ in the form

$$M\varepsilon \geq |g(x_n-\varepsilon)+g(x_n)|. \tag{10}$$

If $M\varepsilon < |g(x_n-\varepsilon)+g(x_n)|$, then there is no zero in $[x_n-\varepsilon, x_n]$ and we have to continue the iterations. Even if $M\varepsilon \geq |g(x_n-\varepsilon)+g(x_n)|$ holds, it is not a guarantee for the existence of a zero in the interval $[x_n-\varepsilon, x_n]$.

In the latter case we can apply algorithm (4) with $y_0=x_n-\varepsilon$. If there really exists a zero $\gamma\in(x_n-\varepsilon, x_n)$, then the sequence $\{y_n\}$ converges to γ and remains less than x_n. If no zero exists in the interval, then $m=\min_{t\in[x_n-\varepsilon,x_n]}|g(t)|>0$ and the iterations $\{y_n\}$ satisfy $y_n\geq y_0+n\frac{m}{M}$. Hence the sequence $\{y_n\}$ exceeds x_n in a finite number of steps. The same happens at the point $x_n-\varepsilon$, if we just continue the iterations $\{x_n\}$.

The two sequences $\{y_n\}$ and $\{x_n\}$ exhibit a two-sided approximation to the zero (if exists) and x_j-y_k is an upper estimate for the error. This error control procedure is fail safe, but it may be expensive. We can make it cheaper by fixing the maximum number of extra iterations at the price of losing absolute certainty. For example, if we use the first extra iteration x_{n+1} ($x_n-\varepsilon<x_{n+1}$) and set $v=x_{n+1}$, then condition (6) changes to

$$M\varepsilon \geq |g(x_n-\varepsilon)+g(x_{n+1})|+|g(x_n)|. \tag{11}$$

Similar expressions can be easily developed for higher number of iterations as well.

A second possibility for improving the exit conditions arises if a number $m>0$ exists such that $m|t-\gamma|\leq|g(t)|\leq M|t-\gamma|$ holds for all $t\in[\alpha,\beta]$. Then $|x_n-\gamma|\leq\frac{1}{m}|g(x_n)|$ is an upper bound for the error. Similarly, we have

$$|x_n-\gamma|\leq\delta+\frac{1}{m}|g(x_n-\delta)|$$

and by selecting $\delta=x_n-x_{n+1}$ we arrive at the bound

$$|x_n-\gamma|\leq x_n-x_{n+1}+\frac{1}{m}|g(x_{n+1})|.$$

This type of a posteriori estimate depends however on the existence and value of m.

3 The one-dimensional optimization algorithm

We now use algorithms (3)-(4) to implement an Algorithm 1 type method for the one-dimensional global extremum problem

$$f(t)\rightarrow\min \quad (l\leq t\leq u,\ f:\mathbb{R}\rightarrow\mathbb{R},\ l,u\in\mathbb{R}) \tag{12}$$

under the assumption that $|f(t)-f(s)| \leq L|t-s|$ holds for all $t,s \in [l,u]$. Here the solution of equation $f(t)=c$ is sought on the interval $[l,u]$.

It first seems handy to apply Algorithm 1 directly with solver (3) or (4). It may happen that equation $f(t)=c_i$ has no solution for some i, and this situation is repeated ad infinitum. Since for $\min f > c_i$, the number of iterations is $O\left(\frac{1}{\min f - c_i}\right)$, this may cause severe problems for $c_i \nearrow \min f$. Assume that $a_k = a_{k+\ell} > \min f > c_{k+\ell} > b_{k+\ell}$ for $\ell \geq 0$. Then $a_{k+\ell} - b_{k+\ell} = a_k - b_{k+\ell} = \frac{a_k - b_k}{2^\ell} \to 0$, which is contradiction to $a_k > \min f > b_{k+\ell}$ $(\ell \geq 0)$. Hence the situation can occur infinitely many times, if by chance $a_k = f(z_k) = \min f$. However preliminary numerical testing indicated a very significant increase of computational time in cases, when c_i just approached $\min f$ from below with a small enough error. This unexpected phenomenon is due to the always convergent property of solver, that we want to keep. Since the iteration numbers also depend on the length of computational interval (see the proof of Theorem 2) we modify Algorithm 1 so that in case $c_i < \min f$ and $c_i \approx \min f$ the computational interval should decrease.

The basic element of the modified algorithm is the solution of equation $g(x) = f(x) - c = 0$ on any subinterval $[\alpha,\beta] \subset [l,u]$. Assume that the upper and lower bounds

$$a = f(x_a) \geq \min_{x\in[\alpha,\beta]} f(x) > b \quad (x_a \in [\alpha,\beta])$$

are given and $c \in (a,b)$. If equation $f(x)=c$ has a solution in $[\alpha,\beta]$, then

$$\min_{x\in[\alpha,\beta]} f(x) \leq c < a,$$

otherwise

$$\min_{x\in[\alpha,\beta]} f(x) > c > b.$$

If $f(\beta) \neq c$, then we compute iterations $\xi_0 = \beta$ and

$$\xi_{i+1} = \xi_i - \frac{|f(\xi_i) - c|}{M} \quad (i \geq 0). \tag{13}$$

There are two cases:

(i) There exists $x^* \in [\alpha,\beta)$ such that $f(x^*) = c$.

(ii) There exists a number k such that $\xi_k = \alpha$ or $\xi_k < \alpha < \xi_{k-1}$.

In case (i) the sequence $\{\xi_k\}$ is monotone decreasing and converges to $x_c \in [\alpha,\beta)$, which is the nearest zero of $f(t)=c$ to β. It is an essential property that

$$\operatorname{sign}(f(t)-c) = \operatorname{sign}(f(\beta)-c) \quad (t \in (x_c,\beta)). \tag{14}$$

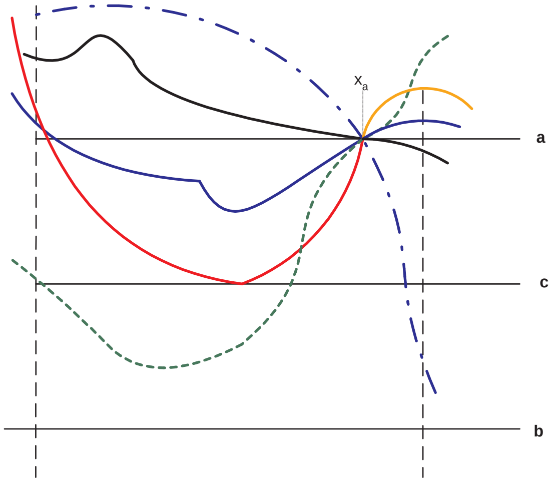
The new upper estimate of the global minimum on $[\alpha,\beta]$ is $a' := c$, $x_{a'} := x_c$ (b unchanged). If $f(\beta) > c$, the inclusion interval $[\alpha,\beta]$ of the global minimum can be restricted to the interval $[\alpha,x_c]$, because $f(t) > c$ $(x_c < t \leq \beta)$. If $f(\beta) < c$, the inclusion interval remains $[\alpha,\beta]$ but the new upper bound $a' = f(\beta)$, $x_{a'} = \beta$, (b unchanged) is better than c. In such a case we do not solve the equation (and save computational time).

In case (ii) we have the iterations $\xi_k < \xi_{k-1} < \cdots < \xi_1 < \xi_0$ such that either $\xi_k = \alpha$ or $\xi_k < \alpha < \xi_{k-1}$ holds. If $\xi_k < \alpha$, or $\xi_k = \alpha$ and $f(\xi_k) \neq c$, we have no solution and $\operatorname{sign}(f(t) - c) = \operatorname{sign}(f(\beta) - c) \quad (t \in [\alpha,\beta))$. If $f(\beta) > c$, the new upper estimate of the global minimum is $a' := a_{est} = \min\left\{f(\alpha), \min_{\xi_i > \alpha} f(\xi_i)\right\}$, $x_{a_{est}}$ ($f(x_{a_{est}}) = a_{est}$). In case $f(\beta) < c$ the best new upper bound is

$$a := \min\left\{f(\alpha), \min_{\xi_i > \alpha} f(\xi_i)\right\}, \quad x_a = \arg\min\left\{f(\alpha), \min_{\xi_i > \alpha} f(\xi_i)\right\},$$

if the iterations are computed. If $f(\beta) < c$, we set the new upper bound as $a' = f(\beta)$, $x_{a'} = \beta$ and do not solve the equation.

A few of the possible situations are shown on the next figure.



Assume that alg1_d is an implementation of algorithm (3) such that

$$\left[\alpha', \beta', a', x_{a'}, b', iflag\right] = \text{alg1}_\text{d}(\alpha, \beta, a, x_a, b; c)$$

denotes its application to equation $f(t) = c$ with the initial value $x_0 = \beta$. If $f(\beta) = c$, then it returns the solution $x_c = \beta$, immediately. If $f(\beta) > c$ it computes iteration (13) and sets the output values according to cases (i) or (ii). If $f(\beta) < c$, then it returns $a' = f(\beta)$ and $x_{a'} = \beta$. We may also require that

$$a \geq a' = f(x_{a'}) \geq \min_{x \in [\alpha,\beta]} f(x) > b' \geq b \quad \wedge \; x_{a'} \in [\alpha,\beta].$$

The $iflag$ variable be defined by

$$iflag = \begin{cases} 1, & \text{if } f(\beta) \geq c \wedge \exists x_c \in [\alpha,\beta] : f(x_c) = c \\ 0, & \text{if } f(\beta) > c \wedge \nexists x_c \in [\alpha,\beta] : f(x_c) = c \\ -1, & \text{if } f(\beta) < c \end{cases}$$

Hence the output parameters are the following:

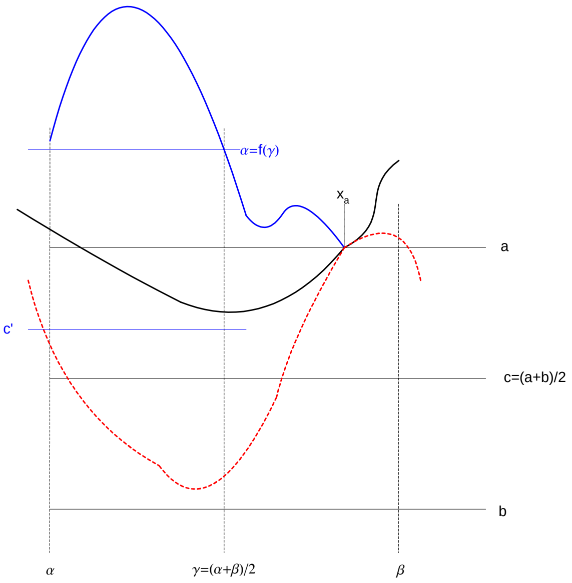
$$\left(\alpha',\beta',a',x_{a'},b'\right)=\begin{cases}(\alpha,x_c,c,x_c,b) & ,iflag=1\\(\alpha,\beta,a_{est},x_{a_{est}},c) & ,iflag=0\\(\alpha,\beta,f(\beta),\beta,b) & ,iflag=-1\end{cases}$$

Instead of $a_{est}=\min\left\{f(\alpha),\min_{\xi_i>\alpha}f(\xi_i)\right\}$ we can take $a_{est}=f(\beta)$, $f(\alpha)$ or any function value at a randomly taken point of $[\alpha,\beta]$. Note that α never changes, a and x_a have no roles in the computations (except for the selection of c), the output a' and $x_{a'}$ are extracted from the computed function values $f(\xi_i)$.

Next we investigate the case, when we halve the interval $[\alpha,\beta]$ and apply $\mathrm{alg1_d}$ to both subintervals $[\alpha,\gamma]$ and $[\gamma,\beta]$ (we assume that $\gamma=(\alpha+\beta)/2$). Consider the possible situations (for simplicity, we assume that $x_a\in[\gamma,\beta]$):

$x\in[\alpha,\gamma]$	$x\in[\gamma,\beta]$
$\min_{x\in[\alpha,\gamma]}f(x)>a$	$\min_{x\in[\gamma,\beta]}f(x)\geq a$
$c<\min_{x\in[\alpha,\gamma]}f(x)\leq a$	$c<\min_{x\in[\gamma,\beta]}f(x)\leq a$
$\min_{x\in[\alpha,\gamma]}f(x)=c$	$\min_{x\in[\gamma,\beta]}f(x)=c$
$\min_{x\in[\alpha,\gamma]}f(x)<c$	$\min_{x\in[\gamma,\beta]}f(x)<c$

There are altogether 16 possible cases. Some possible situations are shown in the next figure for $c=(a+b)/2$.



Assume now that (α,β,a,x_a,b) is given (or popped from a stack) and we have an upper estimate a_{est} (and $x_{a_{est}}$) of $\min_{x\in[l,u]}f(x)$. Estimate a_{est} is assumed to be the smallest among the upper estimates contained in the stack.

If $a_{est}\leq b$, then we can delete (α,β,a,x_a,b) from the stack. Otherwise $b<a_{est}\leq a$

holds. Then we halve the interval $[\alpha,\beta]$ and apply $\mathrm{alg1_d}$ to both subintervals as follows.

Algorithm 2

```
1. Set the estimates $a_{est} = f(u)$ ($x_{a_{est}} = u$), $b$, and push $(l,u,f(u),u,b)$ onto the (empty) stack.

2. **While** *stack is nonempty*
        **pop** $(a,\beta,a,x_a,b)$ from the stack
        **if** $a_{est} \leq b$ **delete** $(a,\beta,a,x_a,b)$ from the stack
        $\left[\alpha,\gamma',a_l',x_{a_l'},b_l',iflag\right] = \mathrm{alg1_d}\left(\alpha,\frac{\alpha+\beta}{2},a,x_a,b;c_l\right)$
        **if** $a_l' < a_{est}$ **then** $a_{est} = a_l^{'}$, $x_{a_{est}} = x_{a_l'}$
        **push** $\left(\alpha,\gamma',a_l',x_{a_l'},b_l'\right)$ onto the stack.
        $\left[\frac{\alpha+\beta}{2},\beta',a_r',x_{a_r'},b_r',iflag\right] = \mathrm{alg1_d}\left(\frac{\alpha+\beta}{2},\beta,a,x_a,b;c_r\right)$
        **if** $a_r' < a_{est}$ **then** $a_{est} = a_r^{'}$, $x_{a_{est}} = x_{a_r'}$
        **push** $\left(\frac{\alpha+\beta}{2},\beta',a_r',x_{a_r'},b_r'\right)$ onto the stack.
    **endwhile**
```

In the practical implementation of Algorithm 2 we used an additional condition ($\beta - \alpha < tol$ and $a - b < tol$) for dropping a stack element. There are many possibilities for choosing c_l and c_r. For simplicity, we selected $c_l = \left(f\left(\frac{\alpha+\beta}{2}\right)+b\right)/2$ and $c_r = (f(\beta)+b)/2$ in the numerical testing.

Molinaro, Sergeyev [30], Sergeyev [33] and Kvasov, Sergeyev [24] investigated the following problem. One must check if a point x^* exists such that

$$g(x^*) = 0, \quad g(x) > 0, \quad x \in [a,x^*) \cup (x^*,b]. \tag{15}$$

These authors suggested the use of Piyavskii type global minimization algorithms to solve the problem in case of Lipschitz functions. However a direct application of algorithms (3)-(4) may also give a satisfactory answer to the problem.

1. Apply algorithm (3) with $x_0 = b$.

2. If a zero ξ of g is found in (a,b), then apply algorithm (4) with $y_0 = a$.

3. If the first zero ζ found by (4) is equal to ξ then the problem is solved. If $\zeta < \xi$, the answer is negative.

4 Numerical experiments

The performance of global Lipschitz optimization clearly depends on the estimation of the unknown Lipschitz constant. Estimates of the Lipschitz constant were suggested and/or analyzed by Strongin [39], [40] Hansen, Jaumard, Lu [16], Wood, Zhang [51] and many others (see, e.g. [29], [24]). Preliminary testing indicated that none of the suggested algorithms performed well, probably due to the local character of the applied equation solver. Instead we used the following although more expensive estimate

$$L \approx L_n^{est} = k \max_{i<n} \left\{ \frac{|f(x_i+h) - f(x_i-h)|}{2h} \right\} + d \quad (h \approx \sqrt{\varepsilon_{machine}})$$

with the values $K = 8$ and $d = 1$. Here $\frac{|f(x_i+h)-f(x_i-h)|}{2h}$ is a second order estimate of the first derivative at the point x_i, if f is differentiable three times and it is optimal in the presence of round-off error.

We used the test problem set of Hansen, Jaumard, Lu [18] numbered as 1–20, four additional functions numbered as 21–24, namely,

$$f(x) = e^{-x} \sin(1/x) \quad \left(x \in \left[10^{-5}, 1\right]\right),$$

$$f(x) = \sin x \quad (x \in [0, 1000]),$$

the Shekel function ([53])

$$f(x) = -\sum_{i=1}^{10} \frac{1}{(k_i(x-a_i))^2 + c_i} \quad (x \in [0, 10])$$

with parameters

i	1	2	3	4	5	6	7	8	9	10
a_i	4	1	8	6	7	9	3	1.5	2	3.6
c_i	0.1	0.2	0.1	0.4	0.4	0.6	0.3	0.7	0.5	0.5

and the Griewank function

$$f(x) = 1 + \frac{1}{4000} x^2 - \cos x \quad (x \in [-600, 600]).$$

In addition, we took 22 test problems of Famularo, Sergeyev, Pugliese [10] without the constraints. This test problems were numbered as 25–46.

All programs were written and tested in Matlab version R2010a (64 bit) on an Intel Core I5 PC with 64 bit Windows. We measured the achieved precision and the computational time for three different exit tolerances (10^{-3}, 10^{-5}, 10^{-7}). Algorithm 2 was compared with a Matlab implementation of the GLOBAL method of Csendes [6], Csendes, Pál, Sendín, Banga [7]. The GLOBAL method is a well-established and maintained stochastic algorithm for multivariable functions that is based on the ideas of Boender etal [5]. The GLOBAL program can be downloaded from the web site

http : //www.inf.u − szeged.hu/˜csendes/index_en.html

The following table contains the averages of output errors for different exit or input tolerances.

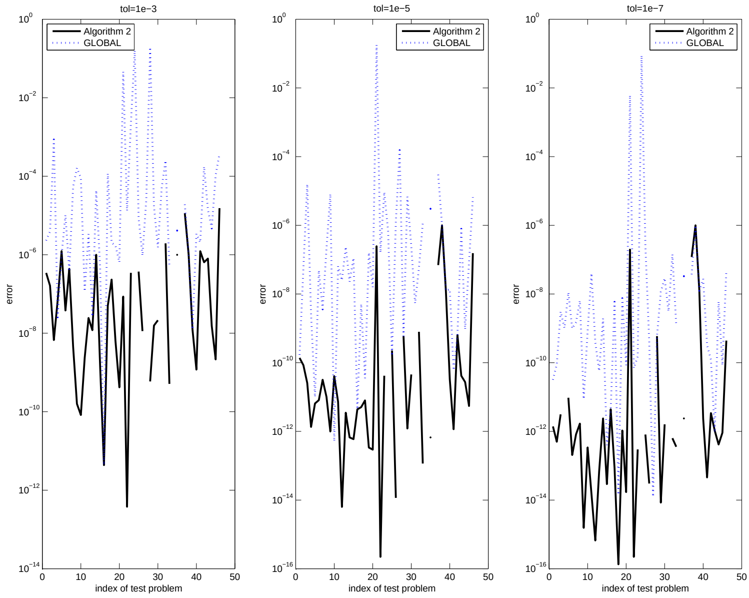
	Algorithm 2	GLOBAL
$1e-3$	$8.2343e-007$	0.0088247
$1e-5$	$3.2244e-008$	0.0039257
$1e-7$	$2.8846e-008$	0.0020635

The average execution times in [sec] are given in the next table:

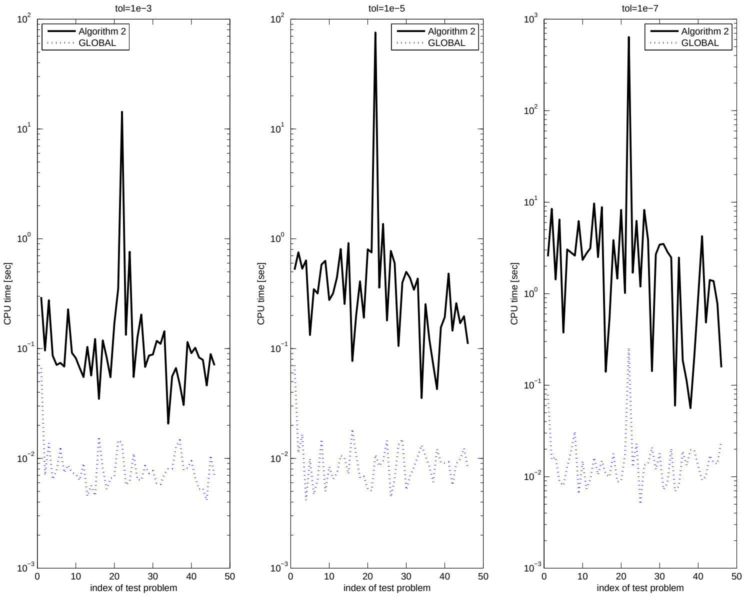
	Algorithm 2	GLOBAL
$1e-3$	0.42863	0.0093795
$1e-5$	2.027	0.010489
$1e-7$	16.6617	0.020512

It is clear that Algorithm 2 has better precision, while GLOBAL is definitely faster. The exit tolerance $1e-7$ does not give essentially better precision, while the computational time significantly increased in the case of both algorithms.

The following two figures show particular details of the achieved precision and computational time.

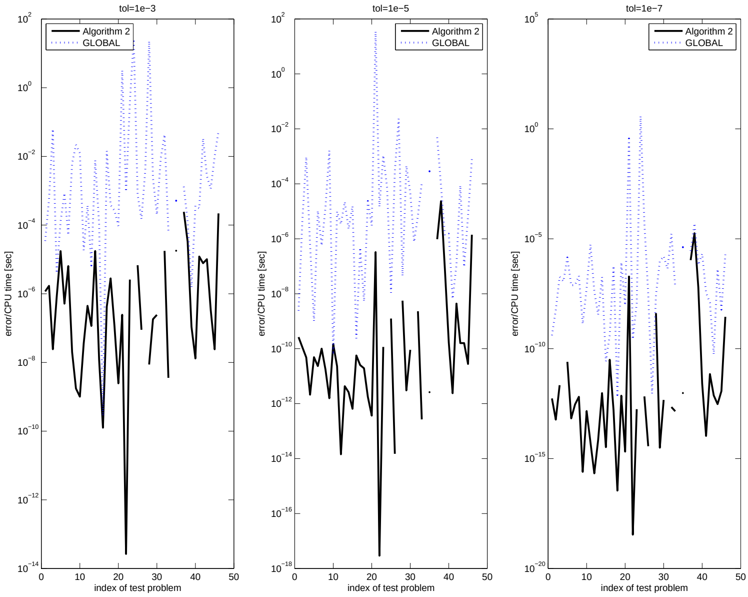


Absolute errors



CPU time

The plots are semi-logarithmic. Hence the missing values of the first figure indicate zero output errors for both algorithms. Considering the obtained precision per CPU time we obtain the following plot.



Precision vs CPU time

The latter plot indicates that Algorithm 2 has a better precision rate per time unit in spite of the fact, that GLOBAL is definitely faster. Upon the basis of the presented numerical testing we conclude that Algorithm 2 might be competitive in univariate global optimization.

References

[1] Abaffy J., Forgó F.: Globally convergent algorithm for solving nonlinear equations, JOTA, 77, 2, 1993, 291–304

[2] Abaffy J., Galántai A.: A globally convergent branch and bound algorithm for global minimization, in LINDI 2011 3rd IEEE International Symposium on Logistics and Industrial Informatics, August 25–27, 2011, Budapest, Hungary, IEEE, 2011, pp. 205-207, ISBN: 978-1-4577-1842-7, DOI: 10.1109/LINDI.2011.6031148

[3] An, H.-B., Bai, Z.-Z.: Directional secant method for nonlinear equations, Journal of Computational and Applied Mathematics 175, 2005, 291–304

[4] Berman, G.: Lattice approximations to the minima of functions of several variables, JACM, 16, 1969, 286–294

[5] Boender, C.G.E., Rinnooy Kan, A.H.G., Timmer, G.T., Stougie, L.: A stochastic method for global optimization. Mathematical Programming, 22, 1982, 125–140

[6] Csendes T.: Nonlinear parameter estimation by global optimization - efficiency and reliability, Acta Cybernetica 8, 1988, 361–370

[7] Csendes T., Pál L., Sendín, J.-Ó. H., Banga, J.R.: The GLOBAL optimization method revisited, Optimization Letters, 2, 2008, 445–454

[8] Delahaye, J.-P.: Sequence Transformations, Springer, 1988

[9] Dellnitz, M., Schütze, O., Sertl, S.: Finding zeros by multilevel subdivision techniques, IMA Journal of Numerical Analysis, 22, 2002, 167–185

[10] Famularo, D., Sergeyev, Ya.D., Pugliese, P.:Test problems for Lipschitz univariate global optimization with multiextremal constraints, in G. Dzemyda, V. Saltenis, A. Zilinskas (eds.): Stochastic and Global Optimization, Kluwer Academic Publishers, Dordrecht, 2002, 93–110

[11] Floudas, C.A., Pardalos, P.M. (eds.): Encyclopedia of Optimization, 2nd ed., Springer, 2009

[12] Ge, R.-D., Xia, Z.-Q.: An ABS algorithm for solving singular nonlinear systems with rank one defect, Korean J. Comput. & Appl. Math. 9, 2002, 167–183

[13] Ge, R.-D., Xia, Z.-Q.: An ABS algorithm for solving singular nonlinear systems with rank defects, Korean J. Comput. & Appl. Math. 12, 2003, 1–20

[14] Ge, R.-D., Xia, Z.-Q., Wang, J.: A ABS algorithm for solving singular nonlinear system with space transformation, JAMC, 30, 2009, 335–348

[15] Hansen, P., Jaumard, B., Lu, S.H.: On the number of iterations of Piyavskii's global optimization algorithm, Mathematics of Operations Research, 16, 1991, 334–350

[16] Hansen, P., Jaumard, B., Lu, S.H.: On using estimates of Lipschitz constants in global optimization, JOTA, 75, 1, 1992, 195–200

[17] Hansen, P., Jaumard, B., Lu, S.H.: Global optimization of univariate Lipschitz functions: I. Survey and properties, Mathematical Programming, 55, 1992, 251–272

[18] Hansen, P., Jaumard, B., Lu, S.H.: Global optimization of univariate Lipschitz functions: II. New algorithms and computational comparison, Mathematical Programming, 55, 1992, 273–292

[19] Huang, Z.: A new method for solving nonlinear underdetermined systems, Computational and Applied Mathematics 1, 1994, 33–48

[20] Kálovics F.: Determination of the global minimum by the method of exclusion, Alkalmazott Matematikai Lapok, 5, 1979, 269–276, in Hugarian

[21] Kálovics F., Mészáros G.: Box valued functions in solving systems of equations and inequalities, Numerical Algorithms, 36, 2004, 1–12

[22] Kearfott, R.B.: Rigorous Global Search: Continuous Problems, Kluwer, 1996

[23] Kvasov, D.E., Sergeyev, Ya.D.: A multidimensional global optimization algorithm based on adaptive diagonal curves, Zh. Vychisl. Mat. Mat. Fiz., 43, 1, 2003, 42–59

[24] Kvasov, D.E., Sergeyev, Ya.D.: Univariate geometric Lipschitz global optimization algorithms, Numerical Algebra, Control and Optimization, 2, 2012, 69–90

[25] Levin, Y., Ben-Israel, A.: Directional Newton method in *n* variables, Mathematics of Computation, 71, 2001, 251–262

[26] McCormick, S.: An iterative procedure for the solution of constrained nonlinear equations with application to optimization problems, Numerische Mathematik, 23, 1975, 371–385

[27] McCormick, S.: The methods of Kaczmarz and row orthogonalization for solving linear equations and least squares problems in Hilbert space, Indiana University Mathematics Journal, 26, 6, 1977, 1137–1150

[28] Meyn, K.-H.: Solution of underdetermined nonlinear equations by stationary iteration methods, Numerische Mathematik, 42, 1983, 161–172

[29] Molinaro, A., Pizzuti, C., Sergeyev, Y.D.: Acceleration tools for diagonal information global optimization algorithms, Computational Optimization and Applications, 18, 2001, 5–26

[30] Molinaro, A., Sergeyev, Y.D.: Finding the minimal root of an equation with the multiextremal and nondifferentiable left-part, Numerical Algorithms, 28, 2001, 255–272

[31] Pietrus, A.: A globally convergent method for solving nonlinear equations without the differentiability condition, Numerical Algorithms, 13, 1996, 60–76

[32] Pintér, J.D.: Global Optimization in Action, Kluwer, 1996

[33] Sergeyev, Y.D.: Finding the minimal root of an equation, in: J.D. Pintér (ed.), Global Optimization, Springer, 2006, 441–460

[34] Shary, S.P.: A surprising approach in interval global optimization, Reliable Computing, 7, 2001, 497–505

[35] Sikorski, K.: Optimal Solution of Nonlinear Equations, Oxford University Press, 2001

[36] Sikorski, K., Wozniakowski, H.: For which error criteria can we solve nonlinear equations?, technical report, CUCS-41-83, Department of Computer Science, Columbia University, New York, 1983

[37] Smiley, M.W., Chun, C.: An algorithm for finding all solutions of a nonlinear system, Journal of Computational and Applied Mathematics, 137, 2001, 293–315

[38] Spedicato, E. and Z. Huang: 1995, 'Optimally stable ABS methods for nonlinear underdetermined systems'. Optimization Methods and Software 5, 17–26

[39] Strongin, R.G.: On the convergence of an algorithm for finding a global extremum, Engineering Cybernetics, 11, 1973, 549–555

[40] Strongin, R.G. Numerical Methods on Multiextremal Problems, Moscow: Nauka.1978, in Russian

[41] Sukharev, A.G.: Optimal search of a root of a function that satisfies a Lipschitz condition, Zh. Vychisl. Mat. Mat. Fiz., 16, 1, 1976, 20–29, in Russian

[42] Sukharev, A.G.: Minimax Algorithms in Problems of Numerical Analysis, Nauka, Moscow, 1989, in Russian

[43] Szabó Z.: Über gleichungslösende Iterationen ohne Divergenzpunkt I-III, Publ. Math. Debrecen, 20 (1973) 222-233, 21 (1974) 285–293, 27 (1980) 185-200

[44] Szabó Z.: Ein Erveiterungsversuch des divergenzpunkfreien Verfahrens der Berührungsprabeln zur Lösung nichtlinearer Gleichungen in normierten Vektorverbänden, Rostock. Math. Kolloq., 22, 1983, 89–107

[45] Tompkins, C.: Projection methods in calculation, in: H. Antosiewicz (ed.): Proc. Second Symposium on Linear Programming, Washington, D.C., 1955, 425–448

[46] Várterész M.: Always convergent iterations for the solution of nonlinear equations, PhD Thesis, Kossuth University, Debrecen, 1998, in Hungarian

[47] Walker, H.F., Watson, L.T.: Least-change secant update methods for underdetermined systems, Report TR 88-28, Comp. Sci. Dept., Viginia Polytechnic Institute and State University, 1988

[48] Walker, H.F.: Newton-like methods for underdetermined systems, in E.L. Allgower, K. Georg (eds.): Computational Solution of Nonlinear Systems, Lectures in Applied Mathematics, 26, AMS, 1990, pp. 679–699

[49] Wang, H.-J., Cao, D.-X.: Interval expansion method for nonlinear equation in several variables, Applied Mathematics and Computation 212, 2009, 153–161

[50] Wood, G.R.: The bisection method in higher dimensions, Mathematical Programming, 55, 1992, 319–337

[51] Wood, G.R., Zhang, B.P.: Estimation of the Lipschitz constant of a function, Journal of Global Optimization, 8, 1, 1996, 91–103

[52] Wood, G.: Bisection global optimization methods, in: C.A. Floudas, P.M. Pardalos (eds.): Encyclopedia of Optimization, 2nd ed., Springer, 2009, pp. 294–297

[53] Zilinskas, A: Optimization of one-dimensional multimodal functions, Journal of the Royal Statistical Society, Series C (Applied Statistics), 27, 3, 1978, 367–375

A Classification of Sub-Riemannian Structures on the Heisenberg Groups

Rory Biggs

Department of Mathematics (Pure and Applied), Rhodes University, PO Box 94, 6140 Grahamstown, South Africa
rorybiggs@gmail.com

Péter T. Nagy

Institute of Applied Mathematics, Óbuda University, H-1034 Budapest, Bécsi út 96/b, Hungary
nagy.peter@nik.uni-obuda.hu

Abstract: We apply Williamson's theorem for the diagonalization of quadratic forms by symplectic matrices to sub-Riemannian (and Riemannian) structures on the Heisenberg groups. A classification of these manifolds, under isometric Lie group automorphisms, is obtained. A (parametrized) list of equivalence class representatives is identified; a geometric characterization of this equivalence relation is provided. A corresponding classification of (drift-free) invariant optimal control problems is exhibited.

Keywords: Heisenberg group; sub-Riemannian geometry; isometry; symplectic group; invariant optimal control; cost-equivalence

1 Introduction

Riemannian geometry is concerned with the (higher dimensional theory of) metric geometry of Euclidean surfaces and in particular the length-minimizing curves on these surfaces. Sub-Riemannian geometry may be interpreted as a generalization of Riemannian geometry. The fundamental difference is that for a sub-Riemannian manifold motion is restricted to certain admissible (or horizontal) directions. Due to such constraints it may not be possible, in general, to connect any two points by a (horizontal) curve. Sub-Riemannian geometry has been a full research domain since the 1980's; it has motivations and ramifications in several areas of pure and applied mathematics. Moreover, there is a substantial overlap between sub-Riemannian geometry ([7, 16]), geometric optimal control ([2, 12, 18]) and nonholonomic mechanics ([5, 8]).

Among the sub-Riemannian manifolds, the Carnot groups are the most fundamental. In the words of Montgomery [16] "Carnot groups are to sub-Riemannian geometry as Euclidean spaces are to Riemannian geometry." The Heisenberg groups are the simplest, non-Euclidean Carnot groups. Structures on the Heisenberg groups (and their generalizations) have been extensively studied in the last few decades (see, e.g., [4, 9, 14, 15, 19]).

In this paper we shall classify, under isometric Lie group automorphisms, the left-invariant bracket-generating sub-Riemannian (and Riemannian) structures on the $(2n+1)$-dimensional (polarized) Heisenberg group

$$\mathsf{H}_n = \left\{ \begin{bmatrix} 1 & x_1 & x_2 & \cdots & x_n & z \\ 0 & 1 & 0 & & 0 & y_1 \\ 0 & 0 & 1 & & 0 & y_2 \\ \vdots & & & \ddots & & \vdots \\ 0 & & \cdots & & 1 & y_n \\ 0 & & \cdots & & 0 & 1 \end{bmatrix} : x_i, y_i, z \in \mathbb{R} \right\}.$$

H_n is a simply-connected two-step nilpotent Lie group with one-dimensional center; its Lie algebra

$$\mathfrak{h}_n = \left\{ \begin{bmatrix} 0 & x_1 & x_2 & \cdots & x_n & z \\ 0 & 0 & 0 & & 0 & y_1 \\ 0 & 0 & 0 & & 0 & y_2 \\ \vdots & & & \ddots & & \vdots \\ 0 & & \cdots & & 0 & y_n \\ 0 & & \cdots & & 0 & 0 \end{bmatrix} = zZ + \sum_{i=1}^{n} (x_i X_i + y_i Y_i) \, : \, x_i, y_i, z \in \mathbb{R} \right\}$$

has non-zero commutators $[X_i, Y_j] = \delta_{ij} Z$. Moreover, any simply-connected two-step nilpotent Lie group with one-dimensional center is isomorphic to H_n.

Let us fix a sub-Riemannian structure on H_n. A standard computation yields the automorphism group of H_n, a subgroup of which is a symplectic group. By use of the automorphisms, we normalize the distributions on H_n. Equivalence class representatives are then constructed by successively applying automorphisms, that preserve the normalized distribution, to the metric. (The Riemannian case is treated similarly.) Central to our argument is Williamson's theorem, which states that any positive definite symmetric matrix can be diagonalized, in a certain way, by symplectic matrices. Furthermore, we shall characterize (in coordinate-free form) when two sub-Riemannian (resp. Riemannian) structures on H_n are equivalent. (This characterization is based on decomposing $\mathfrak{h}_n$, as a vector space, into the product of a symplectic vector space and $\mathbb{R}$.)

To every invariant sub-Riemannian (resp. Riemannian) structure we can naturally associate an invariant optimal control problem (cf. [18]). Accordingly, a classification of sub-Riemannian and Riemannian structures may induce a classification of invariant optimal control problems (or rather, cost-extended systems). In the last section, we exhibit the corresponding classification of invariant optimal control problems on H_n.

1.1 Left-Invariant Sub-Riemannian Structures

By a *left-invariant sub-Riemannian manifold*, we mean a triplet $(\mathsf{G}, \mathscr{D}, \mathbf{g})$, where G is a (real, finite dimensional) connected Lie group with unit element $\mathbf{1}$, $\mathscr{D}$ is a smooth left-invariant distribution on G, and $\mathbf{g}$ is a left-invariant Riemannian metric on $\mathscr{D}$. More precisely, $\mathscr{D}(\mathbf{1})$ is a linear subspace of the Lie algebra $\mathfrak{g}$ of G which is left-translated to the tangent bundle $T\mathsf{G}$ via

$$\mathscr{D}(g) = g\mathscr{D}(\mathbf{1}) \quad \text{for} \quad g \in \mathsf{G}.$$

The metric $\mathbf{g_1}$ is a positive definite symmetric bilinear from on $\mathfrak{g}$ which is extended to $T\mathsf{G}$ by left translation:

$$\mathbf{g}_g(gA, gB) = \mathbf{g_1}(A,B) \quad \text{for} \quad A, B \in \mathfrak{g},\, g \in \mathsf{G}.$$

Here, by the product gA we mean $T_\mathbf{1}L_g \cdot A$, where $L_g : h \mapsto gh$ is a left-translation. We recover a *left-invariant Riemannian manifold* if $\mathscr{D} = T\mathsf{G}$, i.e., $\mathscr{D}(\mathbf{1}) = \mathfrak{g}$.

Remark. Right-invariant sub-Riemannian structures are defined similarly. Such structures are isometric to left-invariant ones (via Lie group anti-isomorphisms).

An absolutely continuous curve $g(\cdot) : [0,T] \to \mathsf{G}$ is called a *horizontal curve* if $\dot{g}(t) \in \mathscr{D}(g(t))$ for almost all $t \in [0,T]$. We shall assume that $\mathscr{D}$ satisfies the bracket generating condition, i.e., $\mathscr{D}(\mathbf{1})$ generates $\mathfrak{g}$; this condition is necessary and sufficient for any two points in G to be connected by a horizontal curve. The *length* of a horizontal curve $g(\cdot)$ is given by

$$\ell(g(\cdot)) = \int_0^T \sqrt{\mathbf{g}(\dot{g}(t), \dot{g}(t))}\, dt.$$

A sub-Riemannian manifold $(\mathsf{G}, \mathscr{D}, \mathbf{g})$ is endowed with a natural metric space structure, namely the *Carnot-Carathéodory distance*:

$$d(g,h) = \inf\{\ell(g(\cdot)) \, : \, g(\cdot) \text{ is a horizontal curve joining } g \text{ and } h\}.$$

A horizontal curve $g(\cdot)$ that realizes the Carnot-Carathéodory distance between two points is called a *minimising geodesic*; these geodesics are fundamental objects of interest in the investigation of sub-Riemannian manifolds. Minimising geodesics exist between any two points if and only if the metric space (G, d) associated with Carnot-Carathéodory distance is complete ([16]).

By an *isometry* between two left-invariant sub-Riemannian (or Riemannian) manifolds $(\mathsf{G}, \mathscr{D}, \mathbf{g})$ and $(\mathsf{G}', \mathscr{D}', \mathbf{g}')$ we mean a diffeomorphism $\phi : \mathsf{G} \to \mathsf{G}'$ such that

$$\phi_* \mathscr{D} = \mathscr{D}' \qquad \text{and} \qquad \mathbf{g} = \phi^* \mathbf{g}'.$$

Any such isometry preserves the Carnot-Carathéodory distance in the sense that $d(g,h) = d'(\phi(g), \phi(h))$. Isometries establish a one-to-one correspondence between the minimizing geodesics of $(\mathsf{G}, \mathscr{D}, \mathbf{g})$ and $(\mathsf{G}', \mathscr{D}', \mathbf{g}')$.

2 Automorphisms

The automorphisms of $\mathfrak{h}_n$ are exactly those linear isomorphisms that preserve the center $\mathfrak{z}$ of $\mathfrak{h}_n$ and for which the induced map on $\mathfrak{h}_n/\mathfrak{z}$ preserves an appropriate symplectic structure (cf. [11]). More precisely, let Ω be the skew-symmetric bilinear form on $\mathfrak{h}_n$ specified by

$$[A,B] = \Omega(A,B)Z, \quad A,B \in \mathfrak{h}_n.$$

Note that $\Omega(X_i,Y_j) = \delta_{ij}$ and that Ω is zero on the remaining pairs of basis vectors. Accordingly, we get the following characterization of automorphisms.

Lemma 1. *A linear isomorphism $\psi : \mathfrak{h}_n \to \mathfrak{h}_n$ is a Lie algebra automorphism if and only if*

$$\psi \cdot Z = cZ \qquad \textit{and} \qquad \Omega(\psi \cdot A, \psi \cdot B) = c\,\Omega(A,B)$$

for some $c \neq 0$.

Proof. Suppose ψ is an automorphism. It preserves the center of $\mathfrak{h}_n$ and therefore $\psi \cdot Z = cZ$ for some $c \neq 0$. For $A,B \in \mathfrak{h}_n$, we have

$$\Omega(\psi \cdot A, \psi \cdot B)Z = \psi \cdot \Omega(A,B)Z \quad \text{and so} \quad \Omega(\psi \cdot A, \psi \cdot B) = c\,\Omega(A,B).$$

Conversely, suppose ψ is a linear isomorphism such that the given conditions hold. For $A,B \in \mathfrak{h}_n$, we have

$$[\psi \cdot A, \psi \cdot B] = \Omega(\psi \cdot A, \psi \cdot B)Z = c\,\Omega(A,B)Z = \psi \cdot \Omega(A,B)Z = \psi \cdot [A,B]. \qquad \square$$

Next, we give a matrix representation for the group of automorphisms. We shall make use of the ordered basis

$$(Z, X_1, X_2, \ldots, X_n, Y_1, Y_2, \ldots, Y_n).$$

The bilinear form Ω takes the form

$$\Omega = \begin{bmatrix} 0 & 0 \\ 0 & J \end{bmatrix}, \quad \text{where} \quad J = \begin{bmatrix} 0 & I_n \\ -I_n & 0 \end{bmatrix}.$$

We denote by ρ the involution

$$\rho = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 0 & I_n \\ 0 & I_n & 0 \end{bmatrix}$$

which is clearly an automorphism.

Proposition 1 (cf. [17])**.** *The group of automorphisms $\mathsf{Aut}(\mathfrak{h}_n)$ is given by*

$$\left\{ \begin{bmatrix} r^2 & \mathbf{v} \\ 0 & rg \end{bmatrix}, \rho \begin{bmatrix} r^2 & \mathbf{v} \\ 0 & rg \end{bmatrix} \; : \; r > 0, \mathbf{v} \in \mathbb{R}^{2n}, g \in \mathsf{Sp}(n,\mathbb{R}) \right\}$$

where

$$\mathsf{Sp}\,(n,\mathbb{R}) = \left\{ g \in \mathbb{R}^{2n\times 2n} \,:\, g^\top J g = J \right\}$$

is the $n(2n+1)$*-dimensional symplectic group over* $\mathbb{R}$*.*

Proof. It is easy to show (by use of the lemma) that the given maps are automorphisms. Suppose ψ is an automorphism. Then $\psi \cdot Z = cZ$ for some $c \neq 0$. We assume $c > 0$. (If $c < 0$, then $\rho\psi$ is of the required form.) Thus

$$\psi = \begin{bmatrix} r^2 & \mathbf{v} \\ 0 & M \end{bmatrix}$$

for some $r > 0$, $\mathbf{v} \in \mathbb{R}^{2n}$ and $M \in \mathsf{GL}\,(2n,\mathbb{R})$. It then follows that $M^\top J M = r^2 J$. For $g = \frac{1}{r}M$, we get $g^\top J g = J$. Thus

$$\psi = \begin{bmatrix} r^2 & \mathbf{v} \\ 0 & rg \end{bmatrix}$$

for some $r > 0$, $\mathbf{v} \in \mathbb{R}^{2n}$ and $g \in \mathsf{Sp}\,(n,\mathbb{R})$. □

Remark. Each automorphism decomposes as a (semidirect) product of

- a translation or inner automorphism $\begin{bmatrix} 1 & \mathbf{v} \\ 0 & I_{2n} \end{bmatrix}$, $\mathbf{v} \in \mathbb{R}^{2n}$
- a dilation $\begin{bmatrix} r^2 & 0 \\ 0 & rI_{2n} \end{bmatrix}$, $r > 0$
- a symplectic transformation $\begin{bmatrix} 1 & 0 \\ 0 & g \end{bmatrix}$, $g \in \mathsf{Sp}\,(n,\mathbb{R})$
- and possibly the involution ρ.

Indeed, we have the following decomposition as semidirect products:

$$\mathsf{Aut}(\mathfrak{h}_n) \cong \mathbb{R}^{2n} \rtimes \mathbb{R} \rtimes \mathsf{Sp}(n,\mathbb{R}) \rtimes \{\mathbf{1},\rho\}.$$

3 Classification

Diffeomorphisms that are compatible with the Lie group structure (in the sense that they preserve left-invariant vector fields) are automorphisms. For the purposes of this paper, we consider only those isometries that are also Lie group automorphisms. We shall refer to such isometries as $\mathfrak{L}$-isometries. For a given left-invariant sub-Riemannian manifold $(\mathsf{G},\mathscr{D},\mathbf{g})$ on a Carnot group G, it turns out that the group of isometries $\phi : (\mathsf{G},\mathscr{D},\mathbf{g}) \to (\mathsf{G},\mathscr{D},\mathbf{g})$ decomposes as a semidirect product of the left translations (normal) and the $\mathfrak{L}$-isometries ([14]). We say that two left-invariant sub-Riemannian (resp. Riemannian) structures are $\mathfrak{L}$*-isometric* if there exists a $\mathfrak{L}$-isometry between them. We classify, under this equivalence relation, the left-invariant sub-Riemannian and Riemannian manifolds on H_n. By left invariance, we have the following simple characterization for $\mathfrak{L}$-isometries.

Proposition 2. *Suppose* G *and* G' *are simply connected.* $(\mathsf{G}, \mathscr{D}, \mathbf{g})$ *and* $(\mathsf{G}', \mathscr{D}', \mathbf{g}')$ *are* $\mathfrak{L}$-isometric *if and only if there exists a Lie algebra isomorphism* $\psi : \mathfrak{g} \to \mathfrak{g}'$ *such that* $\psi \cdot \mathscr{D}(\mathbf{1}) = \mathscr{D}'(\mathbf{1})$ *and* $\mathbf{g_1}(A,B) = \mathbf{g}'_\mathbf{1}(\psi \cdot A, \psi \cdot B)$.

We consider the sub-Riemannian case first; we start by normalizing the distribution.

Lemma 2. *For any (bracket-generating) left-invariant distribution* $\mathscr{D}$ *there exists an inner automorphism* $\phi \in \mathsf{Aut}(\mathsf{H}_n)$ *such that* $\phi_* \mathscr{D} = \bar{\mathscr{D}}$, *where* $\bar{\mathscr{D}}$ *is the left-invariant distribution specified by* $\bar{\mathscr{D}}(\mathbf{1}) = \operatorname{span}(X_1, \ldots, X_n, Y_1, \ldots, Y_n)$.

Proof. It suffices to show that there exists a inner automorphism $\psi \in \mathsf{Aut}(\mathfrak{h}_n)$ such that $\psi \cdot \mathscr{D}(\mathbf{1}) = \bar{\mathscr{D}}(\mathbf{1})$. For any subspace $\mathfrak{s} \subseteq \mathfrak{h}_n$, we have $\mathsf{Lie}(\mathfrak{s}) \leq \operatorname{span}(\mathfrak{s}, Z)$. Therefore, if $\mathsf{Lie}(\mathfrak{s}) = \mathfrak{h}_n$ and $\mathfrak{s} \neq \mathfrak{h}_n$, then $\mathfrak{s}$ has codimension one and takes the form

$$\mathfrak{s} = \operatorname{span}(X_1 + v_1 Z, \ldots, X_n + v_n Z, Y_1 + v_{n+1} Z, \ldots, Y_n + v_{2n} Z).$$

Accordingly,

$$\psi = \begin{bmatrix} 1 & -\mathbf{v} \\ 0 & I_{2n} \end{bmatrix}, \quad \mathbf{v} = \begin{bmatrix} v_1 & v_2 & \cdots & v_{2n} \end{bmatrix}$$

is an inner automorphism such that $\psi \cdot \mathfrak{s} = \operatorname{span}(X_1, \ldots, X_n, Y_1, \ldots, Y_n)$. □

We now proceed to normalise the sub-Riemannian metric and so obtain a classification of the sub-Riemannian structures. We shall make use of Williamson's theorem, which states that positive definite matrices are diagonalizable by symplectic matrices (see [10], Chapter 8.3: "Symplectic Spectrum and Williamson's Theorem"). More precisely,

Lemma 3. *If* M *is a positive definite* $2n \times 2n$ *matrix, then there exists* $S \in \mathsf{Sp}(n, \mathbb{R})$ *such that*

$$S^\top M S = \begin{bmatrix} \Lambda & 0 \\ 0 & \Lambda \end{bmatrix}, \qquad \Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0$.

The array $\operatorname{Spec}(M) = (\lambda_1, \ldots, \lambda_n)$ is called the *symplectic spectrum* of M. (The matrix JM has eigenvalues values $\pm i\lambda_j$.) $\operatorname{Spec}(M)$ is a symplectic invariant, i.e., $\operatorname{Spec}(S^\top M S) = \operatorname{Spec}(M)$ for $S \in \mathsf{Sp}(n, \mathbb{R})$.

Theorem 1. *Any left-invariant sub-Riemannian structure* $(\mathscr{D}, \mathbf{g})$ *on* H_n *is* $\mathfrak{L}$-*isometric to exactly one of the structures* $(\bar{\mathscr{D}}, \bar{\mathbf{g}}^\lambda)$ *specified by*

$$\begin{cases} \bar{\mathscr{D}}(\mathbf{1}) = \operatorname{span}(X_1, \ldots, X_n, Y_1, \ldots, Y_n) \\ \quad \bar{\mathbf{g}}^\lambda_\mathbf{1} = \begin{bmatrix} \Lambda & 0 \\ 0 & \Lambda \end{bmatrix}, \quad \Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n). \end{cases} \tag{1}$$

Here $1 = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0$ *parametrize a family of (non-equivalent) class representatives.*

Proof. By lemma 2, $(\mathscr{D},\mathbf{g})$ is $\mathfrak{L}$-isometric to $(\bar{\mathscr{D}},\mathbf{g}')$ for some left-invariant metric $\mathbf{g}'$. The automorphisms

$$\begin{bmatrix} r^2 & 0 \\ 0 & rI_{2n} \end{bmatrix},\ r>0 \quad \text{and} \quad \begin{bmatrix} 1 & 0 \\ 0 & g \end{bmatrix},\ g \in \mathsf{Sp}(n,\mathbb{R})$$

preserve the subspace $\bar{\mathscr{D}}(\mathbf{1})$, in the sense that $\psi\cdot\bar{\mathscr{D}}(\mathbf{1}) = \bar{\mathscr{D}}(\mathbf{1})$. Let Q be the matrix of the inner product $\mathbf{g}'_{\mathbf{1}}$ on $\mathrm{span}(X_1,\ldots,X_n,Y_1,\ldots,Y_n)$. There exists $g \in \mathsf{Sp}\,(n,\mathbb{R})$ such that

$$g^\top Q g = \begin{bmatrix} \Lambda & 0 \\ 0 & \Lambda \end{bmatrix}$$

where $\Lambda = \mathrm{diag}(\lambda_1,\ldots,\lambda_n)$ and $(\lambda_1,\ldots,\lambda_n) = \mathrm{Spec}(Q)$. Hence

$$\left(\tfrac{1}{\sqrt{\lambda_1}}g\right)^\top Q\left(\tfrac{1}{\sqrt{\lambda_1}}g\right) = \begin{bmatrix} \Lambda' & 0 \\ 0 & \Lambda' \end{bmatrix}$$

where $\Lambda' = \mathrm{diag}(1,\frac{\lambda_2}{\lambda_1},\ldots,\frac{\lambda_n}{\lambda_1})$. Therefore

$$\psi = \begin{bmatrix} 1 & 0 \\ 0 & g \end{bmatrix} \begin{bmatrix} \frac{1}{\lambda_1} & 0 \\ 0 & \frac{1}{\sqrt{\lambda_1}}I_{2n} \end{bmatrix}$$

is an automorphism such that $\mathbf{g}''_{\mathbf{1}}(A,B) = \mathbf{g}'_{\mathbf{1}}(\psi\cdot A,\psi\cdot B)$, where $\mathbf{g}''_{\mathbf{1}}$ has matrix $\begin{bmatrix} \Lambda' & 0 \\ 0 & \Lambda' \end{bmatrix}$. Consequently (relabelling $\frac{\lambda_i}{\lambda_1}$ as λ_i), the result follows by proposition 2.

It remains to be shown that no two class representatives are equivalent. Suppose $(\bar{\mathscr{D}},\bar{\mathbf{g}}^{\lambda})$ and $(\bar{\mathscr{D}},\bar{\mathbf{g}}^{\lambda'})$ are $\mathfrak{L}$-isometric, i.e., there exists an automorphism

$$\psi = \begin{bmatrix} r^2 & \mathbf{v} \\ 0 & rg \end{bmatrix} \quad \text{or} \quad \psi = \rho \begin{bmatrix} r^2 & \mathbf{v} \\ 0 & rg \end{bmatrix}$$

such that $\psi\cdot\bar{\mathscr{D}}(\mathbf{1}) = \bar{\mathscr{D}}(\mathbf{1})$ and $\bar{\mathbf{g}}^{\lambda}_{\mathbf{1}}(A,B) = \bar{\mathbf{g}}^{\lambda'}_{\mathbf{1}}(\psi\cdot A,\psi\cdot B)$. The former condition implies $\mathbf{v} = 0$ and so the latter implies $\Lambda = r^2 g^\top \Lambda' g$, where $\Lambda = \mathrm{diag}(\lambda_1,\ldots,\lambda_n)$ and $\Lambda' = \mathrm{diag}(\lambda'_1,\ldots,\lambda'_n)$. Thus, by symplectic invariance, we have $\mathrm{Spec}(\Lambda) = r^2\,\mathrm{Spec}(\Lambda')$. However for both $\mathrm{Spec}(\Lambda)$ and $\mathrm{Spec}(\Lambda')$, the dominant value is one; so $r = 1$. Consequently $\Lambda = \Lambda'$. That is to say, $(\bar{\mathscr{D}},\bar{\mathbf{g}}^{\lambda})$ and $(\bar{\mathscr{D}},\bar{\mathbf{g}}^{\lambda'})$ are $\mathfrak{L}$-isometric only if $\lambda = \lambda'$. □

Corollary. *Any left-invariant sub-Riemannian structure $(\mathscr{D},\mathbf{g})$ on H_n is $\mathfrak{L}$-isometric to a structure with*

$$(\nu_1 X_1,\, \nu_2 X_2,\, \ldots,\, \nu_n X_n,\, \nu_1 Y_1,\, \nu_2 Y_2,\, \ldots,\, \nu_n Y_n)$$

as orthonormal basis. Here $1 = \nu_1 \le \nu_2 \le \ldots \le \nu_n$ parametrize a family of (non-equivalent) class representatives.

We have the following coordinate-free version of Williamson's theorem. Let μ and μ' be scalar products on a symplectic vector space $(\mathbb{R}^{2n},\omega)$. The symplectic spectrum of μ (resp. μ') is the set of moduli of eigenvalues of the unique linear transformation κ defined by $\omega(\mathbf{x},\kappa\cdot\mathbf{y}) = \mu(\mathbf{x},\mathbf{y})$. A symplectic transformation is a linear isomorphism σ such that $\omega(\sigma\cdot\mathbf{x},\sigma\cdot\mathbf{y}) = \omega(\mathbf{x},\mathbf{y})$.

Lemma 4. *There exists a symplectic transformation* σ *such that*

$$\mu(\mathbf{x},\mathbf{y}) = \mu'(\sigma\cdot\mathbf{x},\sigma\cdot\mathbf{y})$$

if and only if the symplectic spectra of μ *and* μ' *are identical.*

Proof. There exists a basis for $\mathbb{R}^{2n}$ such that ω has matrix J. (A linear map σ is then a symplectic transformation if and only if its matrix is a symplectic matrix.) Let K and M be the matrices of κ and μ, respectively. We have $K = -JM$. Hence the symplectic spectrum of μ is the same as the symplectic spectrum of M (only, every value for M is repeated twice for μ). If $\mu(\mathbf{x},\mathbf{y}) = \mu'(\sigma\cdot\mathbf{x},\sigma\cdot\mathbf{y})$, then $M = S^\top M'S$ (here $S \in \mathsf{Sp}(n,\mathbb{R})$ is the matrix of σ) and so the symplectic spectra of M and M' (resp. μ and μ') match. Conversely, if μ and μ' have identical symplectic spectra, then there exists symplectic matrices $S, S' \in \mathsf{Sp}(n,\mathbb{R})$ such that $S^\top MS = S'^\top M'S'$. Consequently, $M = (S'S^{-1})^\top M'(S'S^{-1})$ and so $\mu(\mathbf{x},\mathbf{y}) = \mu'(\sigma\cdot\mathbf{x},\sigma\cdot\mathbf{y})$ where σ is the unique symplectic transformation with matrix $S'S^{-1}$. □

The Lie algebra $\mathfrak{h}_n$ (as a vector space) can be decomposed as the direct sum of a symplectic vector space $(\mathbb{R}^{2n},\omega)$ and $\mathbb{R}$; the Lie bracket of two elements is given by

$$[(\mathbf{v},z),(\mathbf{v},z)] = (0,\omega(\mathbf{v},\mathbf{v}')) \quad \text{for} \quad (\mathbf{v},z),(\mathbf{v}',z) \in \mathbb{R}^{2n} \oplus \mathbb{R}.$$

By lemma 2, any sub-Riemannian structure $(\mathscr{D},\mathbf{g})$ is $\mathfrak{L}$-isometric to one for which $\mathscr{D}(\mathbf{1}) = \mathbb{R}^{2n}$. Hence the metric $\mathbf{g_1}$ is a scalar product on $\mathbb{R}^{2n}$. The normalized symplectic spectrum of a scalar product is the symplectic spectrum normalized by the dominant value: $\{1, \frac{\lambda_2}{\lambda_1}, \frac{\lambda_3}{\lambda_1}, \ldots, \frac{\lambda_n}{\lambda_1}\}$. Accordingly, by the foregoing considerations, we get the following coordinate-free characterization of the sub-Riemannian structures.

Theorem 2. *Suppose* $(\mathscr{D},\mathbf{g})$ *and* $(\mathscr{D}',\mathbf{g}')$ *are two left-invariant sub-Riemannian structures on* H_n *such that* $\mathscr{D}(\mathbf{1}) = \mathscr{D}'(\mathbf{1}) = \mathbb{R}^{2n}$. *Then* $(\mathscr{D},\mathbf{g})$ *and* $(\mathscr{D}',\mathbf{g}')$ *are* $\mathfrak{L}$-*isometric if and only if the normalized symplectic spectra of* $\mathbf{g_1}$ *and* $\mathbf{g}'_\mathbf{1}$ *are identical.*

Next, we consider the Riemannian case; the classification result is similar to the sub-Riemannian case.

Theorem 3. *Any left-invariant Riemannian structure* $\mathbf{g}$ *on* H_n *is* $\mathfrak{L}$-*isometric to exactly one of the structures*

$$\bar{\mathbf{g}}^\lambda_\mathbf{1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \Lambda & 0 \\ 0 & 0 & \Lambda \end{bmatrix}, \quad \Lambda = \operatorname{diag}(\lambda_1,\lambda_2,\ldots,\lambda_n). \tag{2}$$

Here $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0$ *parametrize a family of (non-equivalent) class representatives.*

Proof. Let R be the matrix of the inner product $\mathbf{g_1}$ on $\mathfrak{h}_n$. We have

$$R = \begin{bmatrix} \frac{1}{r^4} & \mathbf{v} \\ \mathbf{v}^\top & Q \end{bmatrix}$$

for some $r > 0$, $\mathbf{v} \in \mathbb{R}^{2n}$ and $Q \in \mathbb{R}^{2n\times 2n}$. Hence we get

$$\psi = \begin{bmatrix} r^2 & -r^5\mathbf{v} \\ 0 & rI_{2n} \end{bmatrix} \in \mathsf{Aut}(\mathfrak{h}_n) \quad \text{and} \quad \psi^\top R\,\psi = \begin{bmatrix} 1 & 0 \\ 0 & Q' \end{bmatrix}$$

for some positive definite matrix Q'. Accordingly, there exists an automorphism $\psi' = \begin{bmatrix} 1 & 0 \\ 0 & g \end{bmatrix}$, $g \in \mathsf{Sp}\,(n,\mathbb{R})$ such that

$$(\psi \circ \psi')^\top R\,(\psi \circ \psi') = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \Lambda & 0 \\ 0 & 0 & \Lambda \end{bmatrix}$$

where $\Lambda = \mathrm{diag}(\lambda_1,\dots,\lambda_n)$ and $(\lambda_1,\dots,\lambda_n) = \mathrm{Spec}(Q')$. Consequently, the result follows by proposition 2. As in the sub-Riemannian case, it is a simple matter to show that none of these structures are $\mathfrak{L}$-isometric. □

Corollary. *Any left-invariant Riemannian structure* $\mathbf{g}$ *on* H_n *is* $\mathfrak{L}$*-isometric to a structure with*

$$(Z,\, \nu_1X_1,\, \nu_2X_2,\, \dots,\, \nu_nX_n,\, \nu_1Y_1,\, \nu_2Y_2,\, \dots,\, \nu_nY_n)$$

as orthonormal basis. Here $0 < \nu_1 \leq \nu_2 \leq \cdots \leq \nu_n$ *parametrize a family of (non-equivalent) class representatives.*

Any Riemannian structure on H_n is $\mathfrak{L}$-isometric to one for which the scalar product $\mathbf{g_1}$ on $(\mathbb{R}^{2n} \oplus \mathbb{R})$ decomposes as

$$\mathbf{g_1}((\mathbf{v},z),(\mathbf{v}',z')) = \mu_{\mathbf{g}}(\mathbf{v},\mathbf{v}') + zz'$$

where $\mu_{\mathbf{g}}$ is a scalar product on $\mathbb{R}^{2n}$. Accordingly, we have the following coordinate-free characterization the Riemannian structures.

Theorem 4. *Suppose* $\mathbf{g}$ *and* $\mathbf{g}'$ *define two left-invariant Riemannian structures on* H_n *such that*

$$\mathbf{g_1}((\mathbf{v},z),(\mathbf{v}',z')) = \mu_{\mathbf{g}}(\mathbf{v},\mathbf{v}') + zz' \quad \textit{and} \quad \mathbf{g_1}((\mathbf{v},z),(\mathbf{v}',z')) = \mu_{\mathbf{g}'}(\mathbf{v},\mathbf{v}') + zz'.$$

Then $\mathbf{g}$ *and* $\mathbf{g}'$ *are* $\mathfrak{L}$*-isometric if and only if the symplectic spectra of* $\mu_{\mathbf{g}}$ *and* $\mu_{\mathbf{g}'}$ *are identical.*

4 Invariant Optimal Control

Invariant control systems on Lie groups were first considered in 1972 by Brockett [6] and by Jurdjevic and Sussmann [13]. A left-invariant control affine system on a (real, finite-dimensional) Lie group G is a collection of left-invariant vector fields $\Xi(\cdot,u)$ on G, affinely parametrized by controls. In classical notation, a drift-free system $\Sigma = (\mathsf{G},\Xi)$ is written as

$$\dot{g} = g\,\Xi\,(\mathbf{1},u) = g\,(u_1B_1 + \cdots + u_\ell B_\ell)\,, \qquad g \in \mathsf{G},\, u \in \mathbb{R}^\ell.$$

Here the parametrization map $\Xi(\mathbf{1},\cdot) : \mathbb{R}^\ell \to \mathfrak{g}$ is an injective affine map (i.e., $B_1,\dots,B_\ell$ are linearly independent). The "product" $g\,\Xi\,(\mathbf{1},u)$ is given by $g\,\Xi\,(\mathbf{1},u) = T_\mathbf{1}L_g \cdot \Xi\,(\mathbf{1},u)$, where $L_g : \mathsf{G} \to \mathsf{G}$, $h \mapsto gh$ is the left translation by g. The dynamics $\Xi : \mathsf{G} \times \mathbb{R}^\ell \to T\mathsf{G}$ are invariant under left translations, i.e., $\Xi\,(g,u) = g\,\Xi\,(\mathbf{1},u)$. An *admissible control* is a piecewise continuous map $u(\cdot) : [0,T] \to \mathbb{R}^\ell$. A *trajectory* corresponding to an admissible control $u(\cdot)$ is a absolutely continuous curve $g(\cdot) : [0,T] \to \mathsf{G}$ such that $\dot{g}(t) = \Xi(g(t),u(t))$ almost everywhere. A system is said to be *controllable* if any two states can be joined by a trajectory. For more details about invariant control systems see, e.g., [13, 18, 2, 12].

An *invariant optimal control problem* is defined by the specification of (*i*) a left-invariant control system, (*ii*) a positive definite quadratic cost function $L : \mathbb{R}^\ell \to \mathbb{R}$ and (*iii*) boundary data, consisting of an initial state $g_0 \in \mathsf{G}$, a terminal state $g_1 \in \mathsf{G}$ and a (fixed) terminal time $T > 0$. Explicitly, we wish to minimize the functional $\mathscr{J} = \int_0^T L(u(t))dt$ over trajectory-control pairs, subject to the boundary data $g(0) = g_0, \quad g(T) = g_1$. We associate to such a problem, the *cost-extended* system (Σ, L) consisting of a controllable system Σ and a cost function L. Two cost-extended systems $(\Sigma = (\mathsf{G},\Xi),L)$ and $(\Sigma' = (\mathsf{G}',\Xi'),L')$ are *cost-equivalent* ([3]) if there exist a Lie group isomorphism $\phi : \mathsf{G} \to \mathsf{G}'$ and a linear isomorphism $\varphi : \mathbb{R}^\ell \to \mathbb{R}^\ell$ such that

$$T_g\phi \cdot \Xi(g,u) = \Xi'(\phi(g),\varphi(u)) \quad \text{and} \quad rL = L' \circ \varphi$$

for some $r > 0$. The automorphism ϕ establishes a one-to-one correspondence between the optimal trajectories (or corresponding minimising geodesics) of (Σ, L) and (Σ', L'). By left invariance, we have that (Σ, L) and (Σ', L') are cost-equivalent if and only if there exist a Lie group isomorphism $\phi : \mathsf{G} \to \mathsf{G}'$ and an affine isomorphism $\varphi : \mathbb{R}^\ell \to \mathbb{R}^{\ell'}$ such that $T_\mathbf{1}\phi \cdot \Xi(\mathbf{1},u) = \Xi'(\mathbf{1}',\varphi(u))$ and $L' \circ \varphi = rL$ for some $r > 0$.

Analogous to theorems 1 and 3, we get the following classification of cost-extended systems on H_n.

Theorem 5. *Any cost-extended system on H_n is cost-equivalent to exactly one of the following cost-extended systems:*

$$(\Sigma^{2n}, L^{2n}_\lambda) \ : \quad \begin{cases} \Xi^{2n}(\mathbf{1},u) = \displaystyle\sum_{i=1}^n \left(u_{X_i}X_i + u_{Y_i}Y_i\right) \\ L^{2n}_\lambda(u) = \displaystyle\sum_{i=1}^n \lambda_i \left(u^2_{X_i} + u^2_{Y_i}\right) \end{cases}$$

$$(\Sigma^{2n+1}, L^{2n+1}_\lambda) \ : \quad \begin{cases} \Xi^{2n+1}(\mathbf{1},u) = u_Z Z + \displaystyle\sum_{i=1}^n \left(u_{X_i}X_i + u_{Y_i}Y_i\right) \\ L^{2n+1}_\lambda(u) = u_Z^2 + \displaystyle\sum_{i=1}^n \lambda_i \left(u^2_{X_i} + u^2_{Y_i}\right). \end{cases}$$

Here $1 = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0$ parametrize families of (non-equivalent) class representatives.

Remark. The associated optimal control problems are:

$$\begin{cases} \dot{g} = g \displaystyle\sum_{i=1}^{n} \left(u_{X_i} X_i + u_{Y_i} Y_i \right), \qquad g \in \mathsf{H}_n, (u_{X_1}, \ldots, u_{Y_n}) \in \mathbb{R}^{2n} \\ g(0) = g_0 \qquad g(T) = g_1 \\ \displaystyle\int_0^T \sum_{i=1}^{n} \lambda_i \left(u_{X_i}(t)^2 + u_{Y_i}(t)^2 \right) dt \longrightarrow \min \end{cases}$$

$$\begin{cases} \dot{g} = g \Big(u_Z Z + \displaystyle\sum_{i=1}^{n} \left(u_{X_i} X_i + u_{Y_i} Y_i \right) \Big), \qquad g \in \mathsf{H}_n, (u_Z, u_{X_1}, \ldots, u_{Y_n}) \in \mathbb{R}^{2n+1} \\ g(0) = g_0 \qquad g(T) = g_1 \\ \displaystyle\int_0^T \Big(u_Z(t)^2 + \sum_{i=1}^{n} \lambda_i \left(u_{X_i}(t)^2 + u_{Y_i}(t)^2 \right) \Big) dt \longrightarrow \min . \end{cases}$$

Solutions of these optimal control problems are minimising geodesics for the corresponding sub-Riemannian (resp. Riemannian) structures.

Conclusions

We have obtained an explicit classification of the sub-Riemannian (and Riemannian) structures on H_n; an analogous classification of cost-extended control systems was also exhibited. In particular, we found that the Riemannian structures on H_n can be parametrized (up to an $\mathfrak{L}$-isometry) by n parameters, whereas the sub-Riemannian structures can be parametrized by $n-1$ parameters. Agrachev and Barilari [1] classified the invariant sub-Riemannian structures on three-dimensional Lie groups; in particular, they show that all left-invariant sub-Riemannian structures on H_1 are locally isometric. We have shown that all left-invariant sub-Riemannian structures on H_1 are in fact (globally) isometric. Topics for future research include the calculation of the isometry groups and geodesics as well as extensions to Finsler structures.

Acknowledgement

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 317721. Also, the first author would like to acknowledge the financial assistance of the National Research Foundation (DAAD-NRF) and Rhodes University towards this research.

References

[1] A. Agrachev and D. Barilari, Sub-Riemannian structures on 3D Lie groups, *J. Dynam. Control Syst.* **18** (2012), 21–44

[2] A.A. Agrachev and Y.L. Sachkov, *Control Theory from the Geometric Viewpoint*, Springer-Verlag, 2004

[3] R. Biggs and C.C. Remsing, *On the equivalence of cost-extended control systems on Lie groups*, Proc. 8th WSEAS Int. Conf. Dyn. Syst. & Control, Porto, Portugal (2012), 60–65

[4] J. Berndt, F. Tricceri and L. Vanhecke, *Generalized Heisenberg Groups and Damek-Ricci Harmonic Spaces*, Lecture Notes in Math. Vol. 1598, Springer-Verlag, 1995

[5] A.M. Bloch, *Nonholonomic Mechanics and Control*, Springer-Verlag, 2003

[6] R.W. Brockett, System theory on group manifolds and coset spaces, *SIAM J. Control* **10** (1972), 265–284

[7] O. Calin and D.-C. Chang, *Sub-Riemannian Geometry: General Theory and Examples*, Cambridge University Press, 2009

[8] O. Calin, D.-C. Chang and S.S.T. Yau, Nonholonomic systems and sub-Riemannian geometry, *Commun. Inf. Syst.* **10** (2010), 293–316

[9] P. Eberlein, *Geometry of 2-step nilpotent Lie groups*, in Modern Dynamical Systems and Applications, ed. by M. Brin, B. Hasselblatt and Y. Pesin, Cambridge Universtiy Press, 2004, pp. 67–101

[10] M. de Gosson, *Symplectic Geometry and Quantum Mechanics*, volume 166 of Operator Theory: Advances and Applications, Birkhäuser, 2006

[11] V. V. Gorbatsevich, A. L. Onishchik, and E. B. Vinberg. *Foundations of Lie theory and Lie transformation groups*, Springer-Verlag, 1997

[12] V. Jurdjevic, *Geometric Control Theory*, Cambridge University Press, 1997

[13] V. Jurdjevic and H.J. Sussmann, Control systems on Lie groups, *J. Diff. Equations* **12** (1972), 313–329

[14] I. Kishimoto, Geodesics and isometries of Carnot groups, *J. Math. Kyoto Univ.* **43** (2003), 509–522

[15] J. Lauret, Homogeneous nilmanifolds attached to representations of compact Lie groups, *Manuscripta Math.* **99** (1999), 287–309

[16] R. Montgomery, *A Tour of Subriemannian Geometries, Their Geodesics and Applications*, American Mathematical Society, 2002

[17] L. Saal, The automorphism group of a Lie algebra of Heisenberg type, *Rend. Sem. Mat. Univ. Politec. Torino* **54** (1996), 101–113

[18] Y.L. Sachkov, Control theory on Lie groups, *J. Math. Sci.* **156** (2009), 381–439

[19] K.-H. Tan and X.-P. Yang, Characterisation of the sub-Riemannian isometry groups of H-type groups, *Bull. Austral. Math. Soc.* **70** (2004), 87–100

Inequalities for the one-dimensional analogous of the Coulomb potential

Árpád Baricz[1,2] **and Tibor K. Pogány**[1,3]

[1] Óbuda University, John von Neumann Faculty of Informatics, Institute of Applied Mathematics, Bécsi út 96/b, Budapest, Hungary

[2] Babeş–Bolyai University, Department of Economics, 400591 Cluj–Napoca, Romania. e-mail: arpad.baricz@econ.ubbcluj.ro

[3] University of Rijeka, Faculty of Maritime Studies, Studentska 2, 51000 Rijeka, Croatia. e-mail: poganj@pfri.hr

Abstract: *In this paper our aim is to present some monotonicity and convexity properties for the one dimensional regularization of the Coulomb potential, which has applications in the study of atoms in magnetic fields and which is in fact a particular case of the Tricomi confluent hypergeometric function. Moreover, we present some Turán type inequalities for the function in the question and we deduce from these inequalities some new tight upper bounds for the Mills ratio of the standard normal distribution.*

Keywords: *Gaussian integral; regularization of the Coulomb potential; Mills' ratio; Turán type inequalities; functional inequalities; bounds; log-convexity and geometrical convexity.*

MSC (2010): 33E20, 26D15, 60E15.

1 Introduction

Consider the integral

$$V_q(x) = \frac{2e^{x^2}}{\Gamma(q+1)} \int_x^\infty e^{-t^2}(t^2 - x^2)^q dt,$$

where $q > -1$ and $x > 0$. This integral can be regarded as the one dimensional regularization of the Coulomb potential, which has applications in the study of atoms in magnetic fields, see [10] for more details. Recently, Ruskai and Werner [10], and later Alzer [1] studied intensively the properties of this integral. In [1, 10] the authors derived a number of monotonicity and convexity properties for the function V_q, as well as many functional inequalities.

It is important to mention that V_q in particular when $q = 0$ becomes

$$m(x) = \frac{1}{\sqrt{2}} V_0 \left(\frac{x}{\sqrt{2}} \right) = e^{x^2/2} \int_x^\infty e^{-t^2/2} dt,$$

which is the so-called Mills ratio of the standard normal distribution, and appears frequently in economics and statistics. See for example [3] and the references therein for more details on this function.

The purpose of the present study is to make a contribution to the subject and to deduce some new monotonicity and convexity properties for the function V_q, as well as some new functional inequalities. The paper is organized as follows. In section 2 we present the convexity results concerning the function V_q together with some Turán type inequalities. Note that the convexity results are presented in three equivalent formulations. Section 3 is devoted for concluding remarks. In this section we point out that V_q is in fact a particular case of the Tricomi confluent hypergeometric function, and we deduce some other functional inequalities for V_q. In this section we also point out that the Turán type inequalities obtained in section 2 are particular cases of the recent results obtained by Baricz and Ismail [5] for Tricomi confluent hypergeometric functions, however, the proofs are different. Finally, in section 3 we use the Turán type inequalities for the function V_q to derive some new tight upper bounds for the Mills ratio m of the standard normal distribution.

2 Functional inequalities for the function V_q

The first main result of this paper is the following theorem. Parts **a** and **b** of this theorem are generalizations of parts **b** and **d** of [3, Theorem 2.5].

Theorem 1. *The next assertions are true:*

- **a.** *The function $x \mapsto xV_q'(x)/V_q(x)$ is strictly decreasing on $(0,\infty)$ for $q > -1$.*
- **b.** *The function $x \mapsto x^2 V_q'(x)$ is strictly decreasing on $(0,\infty)$ for $q > -1$.*
- **c.** *The function $x \mapsto x^{-1} V_q'(x)$ is strictly increasing on $(0,\infty)$ for $q \geq 0$.*
- **d.** *The function $x \mapsto V_q'(x)/(xV_q(x))$ is strictly increasing on $(0,\infty)$ for $q \geq 0$.*

Proof. **a.** Observe that $V_q(x)$ can be rewritten as [1, Lemma 1]

$$V_q(x) = \frac{x^{q+1/2}}{\Gamma(q+1)} \int_0^\infty e^{-xs} \frac{s^q}{(x+s)^{1/2}} ds.$$

By using the change of variable $s = ux$ we obtain

$$V_q(x) = \frac{x^{2q+1}}{\Gamma(q+1)} \int_0^\infty e^{-x^2 u} \frac{u^q}{(1+u)^{1/2}} du, \tag{2.1}$$

and differentiating with respect to x both sides of this relation we get

$$V_q'(x) = \frac{(2q+1)x^{2q}}{\Gamma(q+1)} \int_0^\infty e^{-x^2 u} \frac{u^q}{(1+u)^{1/2}} du - \frac{2x^{2q+2}}{\Gamma(q+1)} \int_0^\infty e^{-x^2 u} \frac{u^{q+1}}{(1+u)^{1/2}} du.$$

Thus, for $q > -1$ and $x > 0$ we obtain the differentiation formula

$$xV_q'(x) = (2q+1)V_q(x) - 2(q+1)V_{q+1}(x), \tag{2.2}$$

which in turn implies that

$$\frac{xV_q'(x)}{V_q(x)} = 2q+1-2(q+1)\frac{V_{q+1}(x)}{V_q(x)}.$$

Now, recall that [1, Theorem 7] if $p > q > -1$, then the function $x \mapsto V_p(x)/V_q(x)$ is strictly increasing on $(0,\infty)$. In particular, the function $x \mapsto V_{q+1}(x)/V_q(x)$ is strictly increasing on $(0,\infty)$ for $q > -1$, and by using the above relation we obtain that indeed the function $x \mapsto xV_q'(x)/V_q(x)$ is strictly decreasing on $(0,\infty)$ for $q > -1$.

b. According to [1, p. 429] we have

$$V_q'(x) = -\frac{x}{\Gamma(q+1)}\int_0^\infty e^{-t}\frac{t^q}{(x^2+t)^{3/2}}dt. \tag{2.3}$$

Observe that for $q > -1$ and $x > 0$ we have

$$\begin{aligned}\left[-\Gamma(q+1)xV_q'(x)\right]' &= \left[x^2\int_0^\infty e^{-t}\frac{t^q}{(x^2+t)^{3/2}}dt\right]' \\ &= \int_0^\infty e^{-t}\frac{xt^q}{(x^2+t)^{3/2}}\left(2-\frac{3x^2}{x^2+t}\right)dt \\ &> -x\int_0^\infty e^{-t}\frac{t^q}{(x^2+t)^{3/2}}dt = \Gamma(q+1)V_q'(x).\end{aligned}$$

In other words, we proved that for $x > 0$ and $q > -1$ the differential inequality

$$-(xV_q'(x))' > V_q'(x),$$

that is,

$$xV_q''(x) + 2V_q'(x) < 0$$

is valid. Consequently

$$(x^2V_q'(x))' = x(2V_q'(x) + xV_q''(x)) < 0$$

for all $x > 0$ and $q > -1$, which means that indeed the function $x \mapsto x^2V_q'(x)$ is strictly decreasing on $(0,\infty)$ for $q > -1$.

c. Recall the following differentiation formula [10, p. 439]

$$V_q'(x) = 2x(V_q(x) - V_{q-1}(x)), \tag{2.4}$$

which holds for $q \geq 0$ and $x > 0$. Here by convention $V_{-1}(x) = 1/x$, see [10, p. 435]. On the other hand, it is known [1, Theorem 7] that if $p > q > -1$, then $x \mapsto V_p(x) - V_q(x)$ is strictly increasing on $(0,\infty)$. Consequently,

$$x \mapsto x^{-1}V_q'(x) = 2(V_q(x) - V_{q-1}(x))$$

is strictly increasing on $(0,\infty)$ for all $q \geq 0$.

d. Using again the fact that [1, Theorem 7] if $p > q > -1$, then the function $x \mapsto V_p(x)/V_q(x)$ is strictly increasing on $(0,\infty)$, we get that

$$x \mapsto \frac{V_q'(x)}{xV_q(x)} = 2\left(1 - \frac{V_{q-1}(x)}{V_q(x)}\right)$$

is strictly increasing on $(0,\infty)$ for all $q \geq 0$. □

Now, we recall the definition of convex functions with respect to Hölder means or power means. For $a \in \mathbb{R}$, $\alpha \in [0,1]$ and $x,y > 0$, the power mean H_a of order a is defined by

$$H_a(x,y) = \begin{cases} (\alpha x^a + (1-\alpha)y^a)^{1/a}, & a \neq 0 \\ x^\alpha y^{1-\alpha}, & a = 0 \end{cases}.$$

We consider the continuous function $\varphi : I \subset (0,\infty) \to (0,\infty)$, and let $H_a(x,y)$ and $H_b(x,y)$ be the power means of order a and b of $x > 0$ and $y > 0$. For $a,b \in \mathbb{R}$ we say that φ is H_aH_b-convex or just simply (a,b)-convex, if for $a,b \in \mathbb{R}$ and for all $x,y \in I$ we have

$$\varphi(H_a(x,y)) \leq H_b(\varphi(x),\varphi(y)).$$

If the above inequality is reversed, then we say that φ is H_aH_b-concave or simply (a,b)-concave. It is worth to note that $(1,1)$-convexity means the usual convexity, $(1,0)$ is the logarithmic convexity and $(0,0)$-convexity is the geometrical (or multiplicative) convexity. Moreover, we mention that if the function f is differentiable, then (see [4, Lemma 3]) it is (a,b)-convex (concave) if and only if

$$x \mapsto x^{1-a}\varphi'(x)[\varphi(x)]^{b-1}$$

is increasing (decreasing).

For the sake of completeness we recall here also the definitions of log-convexity and geometrical convexity. A function $f\colon (0,\infty) \to (0,\infty)$ is said to be logarithmically convex, or simply log-convex, if its natural logarithm $\ln f$ is convex, that is, for all $x,y > 0$ and $\lambda \in [0,1]$ we have

$$f(\lambda x + (1-\lambda)y) \leq [f(x)]^\lambda \, [f(y)]^{1-\lambda}.$$

A similar characterization of log-concave functions also holds. By definition, a function $g\colon (0,\infty) \to (0,\infty)$ is said to be geometrically (or multiplicatively) convex if it is convex with respect to the geometric mean, that is, if for all $x,y > 0$ and all $\lambda \in [0,1]$ the inequality

$$g(x^\lambda y^{1-\lambda}) \leq [g(x)]^\lambda [g(y)]^{1-\lambda}$$

holds. The function g is called geometrically concave if the above inequality is reversed. Observe that, actually the geometrical convexity of a function g means that the function $\ln g$ is a convex function of $\ln x$ in the usual sense. We also note that the differentiable function f is log-convex (log-concave) if and only if $x \mapsto f'(x)/f(x)$ is increasing (decreasing), while the differentiable function g is geometrically convex (concave) if and only if the function $x \mapsto xg'(x)/g(x)$ is increasing (decreasing).

The next result is a reformulation of Theorem 1 in terms of power means.

Theorem 2. *The next assertions are true:*

- **a.** *V_q is strictly $(0,0)$-concave on $(0,\infty)$ for $q > -1$.*
- **b.** *V_q is strictly $(-1,1)$-concave on $(0,\infty)$ for $q > -1$.*
- **c.** *V_q is strictly $(2,1)$-convex on $(0,\infty)$ for $q \geq 0$.*
- **d.** *V_q is strictly $(2,0)$-convex on $(0,\infty)$ for $q \geq 0$.*

In particular, for all $q \geq 0$ and $x, y > 0$ the next inequalities

$$V_q\left(\sqrt{\frac{x^2+y^2}{2}}\right) < \sqrt{V_q(x)V_q(y)} < V_q(\sqrt{xy}) \tag{2.5}$$

$$\frac{V_q(x)+V_q(y)}{2} < V_q\left(\frac{2xy}{x+y}\right) \tag{2.6}$$

are valid. Moreover, the second inequality in (2.5) *is valid for all $q > -1$, as well as the inequality* (2.6)*. In each of the above inequalities we have equality if and only if $x = y$.*

Now, we extend some of the results of the above theorem to (a,b)-convexity with respect to power means. We note that in the proof of the next theorem we used the corresponding results of Theorem 1. Moreover, it is easy to see that parts **a**, **b**, **c** and **d** of Theorem 3 in particular reduce to the corresponding parts of Theorem 1. Thus, in fact the corresponding parts of Theorem 1 and 3 are equivalent.

Theorem 3. *The following assertions are true:*

- **a.** *V_q is strictly (a,b)-concave on $(0,\infty)$ for $a, b \leq 0$ and $q > -1$.*
- **b.** *V_q is strictly (a,b)-concave on $(0,\infty)$ for $b \leq 1$ and $q > -1 \geq a$.*
- **c.** *V_q is strictly (a,b)-convex on $(0,\infty)$ for $a \geq 2$, $b \geq 1$ and $q \geq 0$.*
- **d.** *V_q is strictly (a,b)-convex on $(0,\infty)$ for $a \geq 2$, $b \geq 0$ and $q \geq 0$.*
- **e.** *V_q is strictly (a,b)-concave on $(0,\infty)$ for $a \leq 1$, $b \leq -1$ and $q \geq 0$.*

Proof. **a.** We consider the functions $u_1, v_1, w_1 : (0,\infty) \to \mathbb{R}$, which are defined by

$$u_1(x) = \frac{xV_q'(x)}{V_q(x)}, \quad v_1(x) = x^{-a}, \quad w_1(x) = V_q^b(x).$$

For $a, b \leq 0$ and $q > -1$ the functions v_1 and w_1 are increasing on $(0,\infty)$, and by using part **a** of Theorem 1, we obtain that the function

$$x \mapsto M_q(x) = u_1(x)v_1(x)w_1(x) = x^{1-a}V_q'(x)V_q^{b-1}(x)$$

is strictly decreasing on $(0,\infty)$. Here we used that $u_1(x) < 0$ for all $x > 0$ and $q > -1$. According to [4, Lemma 3] we obtain that indeed the function V_q is strictly (a,b)-concave on $(0,\infty)$ for $a,b \leq 0$ and $q > -1$.

b. Similarly, if we consider the functions $u_2, v_2, w_2 : (0,\infty) \to \mathbb{R}$, defined by

$$u_2(x) = x^{-a-1}, \quad v_2(x) = x^2 V_q'(x), \quad w_2(x) = V_q^{b-1}(x),$$

then for $a \leq -1 < q$ and $b \leq 1$ the function

$$x \mapsto M_q(x) = u_2(x)v_2(x)w_2(x) = x^{1-a}V_q'(x)V_q^{b-1}(x)$$

is strictly decreasing on $(0,\infty)$. Here we used part **b** of Theorem 1.

c. Analogously, if we consider the functions $u_3, v_3, w_3 : (0,\infty) \to \mathbb{R}$, defined by

$$u_3(x) = x^{2-a}, \quad v_3(x) = x^{-1}V_q'(x), \quad w_3(x) = V_q^{b-1}(x),$$

then for $a \geq 2$, $b \geq 1$ and $q \geq 0$ the function

$$x \mapsto M_q(x) = u_3(x)v_3(x)w_3(x) = x^{1-a}V_q'(x)V_q^{b-1}(x)$$

is strictly increasing on $(0,\infty)$. Here we used part **c** of Theorem 1.

d. If we consider the functions $u_4, v_4, w_4 : (0,\infty) \to \mathbb{R}$, defined by

$$u_4(x) = x^{2-a}, \quad v_4(x) = x^{-1}V_q'(x)V_q^{-1}(x), \quad w_4(x) = V_q^b(x),$$

then for $a \geq 2$, $b \leq 1$, $q \geq 0$, the function

$$x \mapsto M_q(x) = u_4(x)v_4(x)w_4(x) = x^{1-a}V_q'(x)V_q^{b-1}(x)$$

is strictly increasing on $(0,\infty)$. Here we used part **d** of Theorem 1.

e. If we consider the functions $u_4, v_4, w_4 : (0,\infty) \to \mathbb{R}$, defined by

$$u_5(x) = x^{1-a}, \quad v_5(x) = V_q'(x)V_q^{-2}(x), \quad w_5(x) = V_q^{b+1}(x),$$

then for $a \leq 1$, $b \leq -1$, $q \geq 0$, the function

$$x \mapsto M_q(x) = u_5(x)v_5(x)w_5(x) = x^{1-a}V_q'(x)V_q^{b-1}(x)$$

is strictly decreasing on $(0,\infty)$. Here we used the fact that for $q \geq 0$ the function $1/V_q$ is strictly convex (see [1, Theorem 2]) on $(0,\infty)$, which is equivalent to the fact that V_q is strictly $(1,-1)$-concave on $(0,\infty)$, or to that the function v_5 is strictly decreasing on $(0,\infty)$. □

The following theorem presents some Turán type inequalities for the function V_q. These kind of inequalities are named after the Hungarian mathematician Paul Turán who proved a similar inequality for Legendre polynomials. For more details on Turán type inequalities we refer to the papers [2, 5] and to the references therein.

Theorem 4. *For $x > 0$ the function $q \mapsto \Gamma(q+1)V_q(x)$ is strictly log-convex on $(-1,\infty)$, and if $q > -1/2$ and $x > 0$, then the next Turán type inequalities hold*

$$\frac{(q+2)(2q+1)}{(q+1)(2q+3)}V_q(x)V_{q+2}(x) < V_{q+1}^2(x) < \frac{q+2}{q+1}V_q(x)V_{q+2}(x). \tag{2.7}$$

Moreover, the right-hand side of (2.7) *is valid for $q > -1$ and $x > 0$. The left-hand side of* (2.7) *is sharp as x tends to* 0.

Proof. We use the notation $f(q) = \Gamma(q+1)V_q(x)$. Since [1, p. 426]

$$V_q(x) = \frac{1}{\Gamma(q+1)}\int_0^\infty e^{-t}\frac{t^q}{(x^2+t)^{1/2}}dt$$

it follows that

$$f(q) = \int_0^\infty e^{-t}\frac{t^q}{(x^2+t)^{1/2}}dt.$$

By using the Hölder-Rogers inequality for integrals we obtain that for all $q_1, q_2 > -1$, $q_1 \neq q_2$, $\alpha \in (0,1)$ and $x > 0$ we have

$$\begin{aligned} f(\alpha q_1 + (1-\alpha)q_2) &= \int_0^\infty e^{-t}\frac{t^{\alpha q_1+(1-\alpha)q_2}}{(x^2+t)^{1/2}}dt \\ &= \int_0^\infty \left(e^{-t}\frac{t^{q_1}}{(x^2+t)^{1/2}}\right)^\alpha \left(e^{-t}\frac{t^{q_2}}{(x^2+t)^{1/2}}\right)^{1-\alpha} dt \\ &< \left(\int_0^\infty e^{-t}\frac{t^{q_1}}{(x^2+t)^{1/2}}dt\right)^\alpha \left(\int_0^\infty e^{-t}\frac{t^{q_2}}{(x^2+t)^{1/2}}dt\right)^{1-\alpha} \\ &= (f(q_1))^\alpha (f(q_2))^{1-\alpha}, \end{aligned}$$

that is, the function f is strictly log-convex on $(-1,\infty)$ for $x > 0$. Now, choosing $\alpha = 1/2$, $q_1 = q$ and $q_2 = q+2$ in the above inequality we obtain the Turán type inequality

$$f^2(q+1) < f(q)f(q+2)$$

which is equivalent to the inequality

$$V_{q+1}^2(x) < \frac{\Gamma(q+3)\Gamma(q+1)}{\Gamma^2(q+2)}V_q(x)V_{q+2}(x),$$

valid for $q > -1$ and $x > 0$. After simplifications we get the right-hand side of (2.7).

Now, we focus on the left-hand side of (2.7). First observe that from (2.3) it follows that $V_q'(x) < 0$ for all $x > 0$ and $q > -1$. In view of the differentiation formula (2.2) this implies that for $x > 0$ and $q > -1$ we have

$$(2q+1)V_q(x) < 2(q+1)V_{q+1}(x). \tag{2.8}$$

On the other hand, recall that the function $x \mapsto V_{q+1}(x)/V_q(x)$ is strictly increasing on $(0,\infty)$ for $q > -1$, that is, the inequality

$$\left(V_{q+1}(x)/V_q(x)\right)' > 0$$

is valid for $x > 0$ and $q > -1$. By using (2.2) it can be shown that the above assertion is equivalent to the Turán type inequality

$$(q+1)V_{q+1}^2(x) - (q+2)V_q(x)V_{q+2}(x) > -V_q(x)V_{q+1}(x), \tag{2.9}$$

where $x > 0$ and $q > -1$. Combining (2.8) with (2.9) for $q > -1/2$ and $x > 0$ we have

$$(q+1)V_{q+1}^2(x) - (q+2)V_q(x)V_{q+2}(x) > -\frac{2(q+1)}{2q+1}V_{q+1}^2(x),$$

which is equivalent to the left-hand side of (2.7).

Finally, since

$$V_q(0) = \frac{\Gamma(q+1/2)}{\Gamma(q+1)},$$

it follows that

$$\frac{V_{q+1}^2(0)}{V_q(0)V_{q+2}(0)} = \frac{(q+2)(2q+1)}{(q+1)(2q+3)},$$

and thus indeed the left-hand side of (2.7) is sharp as x tends to 0. □

3 Concluding remarks and further results

3.1 Connection with Tricomi confluent hypergeometric functions and Turán type inequalities

First consider the Tricomi confluent hypergeometric function, called also sometimes as the confluent hypergeometric function of the second kind, $\psi(a,c,\cdot)$, which is a particular solution of the so-called confluent hypergeometric differential equation

$$xw''(x) + (c-x)w'(x) - aw(x) = 0$$

and its value is defined in terms of the usual Kummer confluent hypergeometric function $\Phi(a,c,\cdot)$ as

$$\psi(a,c,x) = \frac{\Gamma(1-c)}{\Gamma(a-c+1)}\Phi(a,c,x) + \frac{\Gamma(c-1)}{\Gamma(a)}x^{1-c}\Phi(a-c+1,2-c,x).$$

For $a,x > 0$ this function possesses the integral representation

$$\psi(a,c,x) = \frac{1}{\Gamma(a)}\int_0^\infty e^{-xt}t^{a-1}(1+t)^{c-a-1}dt,$$

and consequently we have

$$V_q(x) = \frac{x^{2q+1}}{\Gamma(q+1)}\int_0^\infty \mathrm{e}^{-x^2u}\,\frac{u^q}{\sqrt{1+u}}\,\mathrm{d}u = x^{2q+1}\psi(q+1,q+3/2,x^2)\,. \tag{3.1}$$

Thus, the Turán type inequality (2.7) can be rewritten as

$$\frac{(a+1)(2a-1)}{a(2a+1)} < \frac{\psi^2(a+1,a+3/2,x)}{\psi(a,a+1/2,x)\psi(a+2,a+5/2,x)} < \frac{a+1}{a}, \tag{3.2}$$

where $a > 1/2$ and $x > 0$ on the left-hand side, and $a > 0$ and $x > 0$ on the right-hand side. Now, applying the Kummer transformation

$$\psi(a,c,x) = x^{1-c}\psi(1+a-c,2-c,x),$$

the above Turán type inequality becomes

$$\frac{c(2c-3)}{(c-1)(2c-1)} < \frac{\psi^2(1/2,c,x)}{\psi(1/2,c-1,x)\psi(1/2,c+1,x)} < \frac{2c-3}{2c-1}, \tag{3.3}$$

where $x > 0 > c$ on the left-hand side, and $c < 1/2$ and $x > 0$ on the right-hand side. It is important to mention here that the right-hand side of (3.3) is not sharp when $c < 0$. Namely, in [5, Theorem 4] it was proved that the sharp Turán type inequality

$$\psi^2(a,c,x) - \psi(a,c-1,x)\psi(a,c+1,x) < 0$$

is valid for $a > 0 > c$ and $x > 0$ or $a > c-1 > 0$ and $x > 0$. This implies that

$$\frac{\psi^2(1/2,c,x)}{\psi(1/2,c-1,x)\psi(1/2,c+1,x)} < 1$$

holds for $c < 0$ and $x > 0$ or $c \in (1,3/2)$ and $x > 0$, and the constant 1 on the right-hand side of this inequality is the best possible. The above Turán type inequality clearly improves the right-hand side of (3.3) when $c < 0$, and this means that for $q > -1$ and $x > 0$ the right-hand side of (2.7) can be improved as follows

$$V_{q+1}^2(x) < V_q(x)V_{q+2}(x). \tag{3.4}$$

Note also that very recently Baricz and Ismail in [5, Theorem 4] proved the sharp Turán type inequality

$$\frac{a}{c(a-c+1)}\psi^2(a,c,x) < \psi^2(a,c,x) - \psi(a,c-1,x)\psi(a,c+1,x),$$

which is valid for $a > 0 > c$ and $x > 0$. This inequality can be rewritten as

$$\frac{c(a-c+1)}{(c-1)(a-c)} < \frac{\psi^2(a,c,x)}{\psi(a,c-1,x)\psi(a,c+1,x)},$$

which for $a = 1/2$ reduces to the left-hand side of (3.3). It is important to note here that according to [5, Theorem 4] in the above Turán type inequalities the constants

$$a(c(a-c+1))^{-1} \quad \text{and} \quad (c(a-c+1))/((c-1)(a-c))^{-1}$$

are best possible, and so is the constant

$$c(2c-3)/((c-1)(2c-1))^{-1}$$

in (3.3).

We also mention that the method of proving (2.7) is completely different than of the proof of [5, Theorem 4]. Note also that the sharp Turán type inequality (3.4) is in fact related to the following open problem [2, p. 87]: is the function $q \mapsto V_q(x)$ log-convex on $(-1,\infty)$ for $x>0$ fixed? If this result were be true then would improve Alzer's result [1, Theorem 3], which states that the function $q \mapsto V_q(x)$ is convex on $(-1,\infty)$ for all $x>0$ fixed.

Recently, for $x>0$ Simon [11] proved the next Turán type inequalities

$$\psi(a-1,c-1,x)\psi(a+1,c+1,x)-\psi^2(a,c,x) \le \frac{1}{x}\psi^2(a,c,x)\psi(a+1,c+1,x), \tag{3.5}$$

$$\psi(a,c-1,x)\psi(a,c+1,x)-\psi^2(a,c,x) \le \frac{1}{x}\psi(a,c,x)\psi(a,c-1,x). \tag{3.6}$$

In (3.5) it is supposed that $a>1$ and $c<a+1$, while in (3.6) it is assumed that $a\ge 1$ or $a>0$ and $c\le a+2$. By using (3.1) the inequality (3.5) in particular reduces to

$$V_q(x)V_{q+2}(x) \le V_{q+1}^2(x)\left(1+x^{-2(q+3)}V_{q+2}(x)\right),$$

where $q>-1$ and $x>0$. Now, observe that by using the above mentioned Kummer transformation in (3.1) we obtain

$$V_q(x)=\psi(1/2,1/2-q,x^2),$$

and by using this, (3.6) in particular reduces to

$$V_q(x)V_{q+2}(x)-V_{q+1}^2(x) \le \frac{1}{x}V_{q+1}(x)V_{q+2}(x),$$

where $q>-1$ and $x>0$. Combining this inequality with (3.4) for $q>-1$ and $x>0$ we obtain

$$-\frac{1}{x}V_{q+1}(x)V_{q+2}(x) \le V_{q+1}^2(x)-V_q(x)V_{q+2}(x)<0.$$

3.2 Connection with Mills ratio and some new bounds for this function

In this subsection we would like to show that the inequalities presented above for the function V_q can be used to obtain many new results for the Mills ratio m. For this, first recall that Mills' ratio m satisfies the differential equation [3, p. 1365] $m'(x)=xm(x)-1$ and hence

$$V_0'(x)=(\sqrt{2}\cdot m(x\sqrt{2}))'=2(\sqrt{2}x\cdot m(x\sqrt{2})-1)=2(xV_0(x)-1). \tag{3.7}$$

Note that this differentiation formula can be deduced also from (2.4). Observe that by using (2.2) and (3.7) we get

$$2V_1(x)=(1-2x^2)V_0(x)+2x, \tag{3.8}$$

$$8V_2(x)=(4x^4-4x^2+3)V_0(x)+2x(3-2x^2).$$

Now, if $q \to -1$ in (3.4), then we get that the Turán type inequality

$$V_0^2(x) < V_{-1}(x)V_1(x)$$

is valid for $x > 0$, and this is equivalent to

$$2xV_0^2(x) + (2x^2 - 1)V_0(x) - 2x < 0.$$

From this we obtain that for $x > 0$ the inequality

$$V_0(x) < \frac{1 - 2x^2 + \sqrt{4x^4 + 12x^2 + 1}}{4x}$$

is valid, and rewriting in terms of Mills ratio we get

$$m(x) < \frac{1 - x^2 + \sqrt{x^4 + 6x^2 + 1}}{4x}.$$

Similarly, if we take $q = 0$ in the left-hand side of (2.7), then we get

$$2V_0(x)V_2(x) < 3V_1^2(x),$$

which can be rewritten as

$$4x(x^2 - 1)V_0^2(x) + (3 - 10x^2)V_0(x) + 6x > 0.$$

From this for $x > 0$ we obtain

$$V_0(x) < \frac{10x^2 - 3 - \sqrt{4x^4 + 36x^2 + 9}}{8x(x^2 - 1)},$$

which in terms of Mills ratio can be rewritten as

$$m(x) < \frac{5x^2 - 3 - \sqrt{x^4 + 18x^2 + 9}}{4x(x^2 - 2)},$$

where $x > 0$. As far as we know these upper bounds on Mills ratio m are new. We note that many other results of this kind can be obtained by using for example (3.4) for $q = 0$ or by using the other Turán type inequalities in the previous subsection.

Finally, we mention that if we take in (2.4) the value $q = 0$ and we take into account that V_0 is strictly decreasing on $(0, \infty)$, we get the inequality $V_0(x) < 1/x$, which in terms of Mills ratio can be rewritten as $m(x) < 1/x$. This inequality is the well-known Gordon inequality for Mills' ratio, see [7] for more details. Note that the inequality $V_0(x) < 1/x$ can be obtained also from (2.8), just choosing $q = 0$ and taking into account the relation (3.8) between V_0 and V_1. It is important to mention here that Gordon's inequality $m(x) < 1/x$ is in fact a particular Turán type inequality for the parabolic cylinder function, see [5, p. 199] for more details.

3.3 Other results on Mills ratio and their generalizations

It is worth to mention that it is possible to derive other inequalities for V_q and its particular case m by using the recurrence relations for this function. Namely, from (2.4) we get that

$$V_q(x) < V_{q-1}(x)$$

for $q \geq 0$ and $x > 0$, and by using (2.2) we obtain

$$(xV_q(x))' = 2(q+1)(V_q(x) - V_{q+1}(x)) > 0,$$

where $x > 0$ and $q > -1$. On the other hand, by using (2.4) it follows

$$(xV_q(x))' = (2x^2+1)V_q(x) - 2x^2V_{q-1}(x),$$

and from the previous inequality we get the inequality

$$\frac{V_q(x)}{V_{q-1}(x)} > \frac{2x^2}{2x^2+1}, \tag{3.9}$$

which holds for all $q \geq 0$ and $x > 0$. Now, if we take $q = 0$ and $q = 1$ in (3.9) we obtain the inequalities

$$\frac{2x}{2x^2+1} < V_0(x) < \frac{2x(2x^2+1)}{4x^4+4x^2-1},$$

where $x > 0$ on the left-hand side, and $x\sqrt{2} > \sqrt{\sqrt{2}-1}$ on the right-hand side. This inequality in terms of Mills ratio can be rewritten as

$$\frac{x}{x^2+1} < m(x) < \frac{x(x^2+1)}{x^4+2x^2-1}, \tag{3.10}$$

where $x > 0$ on the left-hand side and $x > \sqrt{\sqrt{2}-1}$ on the right-hand side. Observe that the right-hand side of (3.10) is better than Gordon's inequality $m(x) < 1/x$ when $x > 1$. We also note that the left-hand side of (3.10) is known and it was deduced by Gordon [7].

Now, let us consider the functions $f_1, f_2, f_3, f_4, f_5 : (0, \infty) \to \mathbb{R}$, defined by

$$f_1(x) = \frac{x}{x^2+1},\ f_2(x) = \frac{1}{x},\ f_3(x) = \frac{x(x^2+1)}{x^4+2x^2-1},$$

$$f_4(x) = \frac{1 - x^2 + \sqrt{x^4+6x^2+1}}{4x},\ f_5(x) = \frac{5x^2 - 3 - \sqrt{x^4+18x^2+9}}{4x(x^2-2)}.$$

Figure 1 shows that the above new upper bounds (for the Mills ratio of the standard normal distribution) are quite tight.

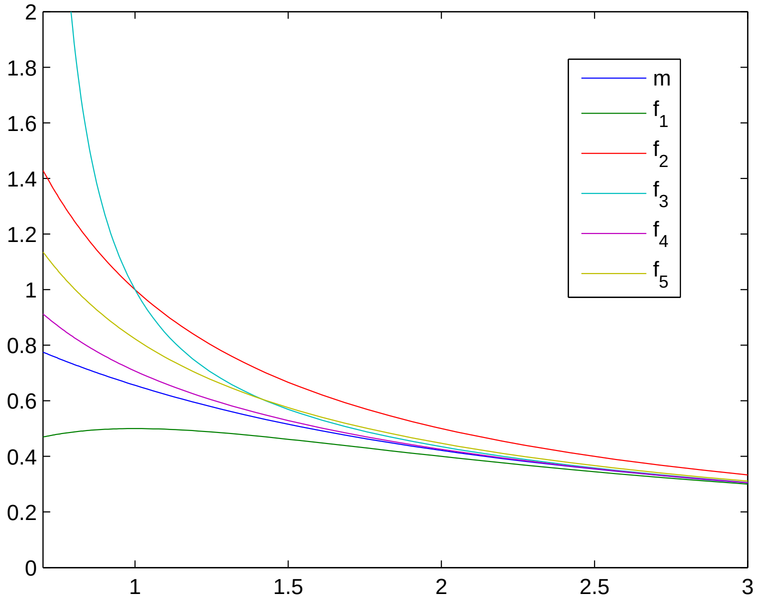


Figure 1
The graph of Mills' ratio m of the standard normal distribution and of the bounds f_1, f_2, f_3, f_4 and f_5 on $[0.7,3]$.

3.4 Connection with Gaussian hypergeometric functions

Here we would like to show that V_q can be expressed in terms of Gaussian hypergeometric functions. For this, we consider the following integral [9, p. 18, Eq. 2.29]

$$\int_0^\infty \frac{x^{p-1}\,\mathrm{d}x}{(c+bx)^\nu\,(a+dx)^\mu} = \frac{c^{-\nu+p}}{b^p d^\mu}\,\mathrm{B}(p,\mu+\nu-p)\,{}_2F_1\left(\mu,p;\mu+\nu;1-\frac{ac}{bd}\right)$$
$$= \frac{d^{-\mu+p}}{c^\nu a^p}\,\mathrm{B}(p,\mu+\nu-p)\,{}_2F_1\left(\nu,p;\mu+\nu;1-\frac{bd}{ac}\right),$$

valid in both cases for all $0<p<\mu+\nu$. Specifying inside

$$a=c=d=1, \quad b=\frac{x^2}{n}, \quad \nu=n,\mu=\frac{1}{2}, \quad p=q+1,$$

we conclude

$$\begin{aligned}V_q(x) &= \frac{x^{2q+1}}{\Gamma(q+1)} \int_0^\infty \lim_{n\to\infty} \left(1+\frac{x^2 t}{n}\right)^{-n} \frac{t^q \,\mathrm{d}t}{\sqrt{1+t}} \\ &= \frac{1}{x} \lim_{n\to\infty} n^{q+1} \frac{\Gamma(n+\frac{1}{2}-q)}{\Gamma(n+\frac{3}{2})} {}_2F_1\left(\frac{1}{2}, q+1; n+\frac{1}{2}; 1-\frac{n}{x^2}\right) \\ &= \frac{x^{2q+1}}{\Gamma(q+1)} \lim_{n\to\infty} \frac{\Gamma(n+\frac{1}{2}-q)}{\Gamma(n+\frac{3}{2})} {}_2F_1\left(n, q+1; n+\frac{1}{2}; 1-\frac{x^2}{n}\right).\end{aligned}$$

3.5 Lower and upper bounds for the function V_q

It is of considerable interest to find lower and upper bounds for the function $x \mapsto V_q(x)$ itself. Therefore remarking the obvious inequality $1+a \le \mathrm{e}^a$, $a \in \mathbb{R}$, we conclude the following. Having in mind the integral expression (2.1), and specifying $a = u$, we get

$$\begin{aligned}V_q(x) &= \frac{x^{2q+1}}{\Gamma(q+1)} \int_0^\infty \mathrm{e}^{-x^2 u} \frac{u^q}{\sqrt{1+u}} \,\mathrm{d}u \\ &\ge \frac{x^{2q+1}}{\Gamma(q+1)} \int_0^\infty \mathrm{e}^{-(x^2+\frac{1}{2})u} u^q \,\mathrm{d}u \\ &= \frac{2^{q+1} x^{2q+1}}{(1+2x^2)^{q+1}}.\end{aligned}$$

Similarly, transforming the integrand of (2.1) by the arithmetic mean–geometric mean inequality $1+u \ge 2\sqrt{u}$, $u \ge 0$, we get

$$V_q(x) \le \frac{x^{2q+1}}{\sqrt{2}\,\Gamma(q+1)} \int_0^\infty \mathrm{e}^{-x^2 u} u^{q-\frac{1}{4}} \,\mathrm{d}u = \frac{\Gamma(q+\frac{3}{4})}{\sqrt{2x}\,\Gamma(q+1)}.$$

Finally, choosing $a = x^2 t^{-1}$, we get

$$\begin{aligned}V_q(x) &= \frac{1}{\Gamma(q+1)} \int_0^\infty \mathrm{e}^{-u} \frac{u^q}{\sqrt{x^2+u}} \,\mathrm{d}u \\ &\ge \frac{1}{\Gamma(q+1)} \int_0^\infty \mathrm{e}^{-u-\frac{x^2}{2u}} u^{q-\frac{1}{2}} \,\mathrm{d}u \\ &= \frac{1}{\Gamma(q+1)} Z_1^{q+\frac{1}{2}}\left(\frac{x^2}{2}\right),\end{aligned}$$

where

$$Z_\rho^\nu(t) = \int_0^\infty u^{\nu-1} \mathrm{e}^{-u^\rho - \frac{t}{u}} \,\mathrm{d}u,$$

stands for the so-called Krätzel function, see [8], and also [6]. Note that further consequent inequalities have been established for the Krätzel function in [6], compare [6, Theorem 1].

References

[1] H. Alzer, On a convexity theorem of Ruskai and Werner and related results, *Glasgow Math. J.* 47 (2005) 425–438.

[2] Á. Baricz, Turán type inequalities for some special functions, PhD Thesis, University of Debrecen, Hungary, 2008.

[3] Á. Baricz, Mills' ratio: Monotonicity patterns and functional inequalities, *J. Math. Anal. Appl.* 340(2) (2008) 1362–1370.

[4] Á. Baricz, Geometrically concave univariate distributions, *J. Math. Anal. Appl.* 363(1) (2010) 182–196.

[5] Á. Baricz, M.E.H. Ismail, Turán type inequalities for Tricomi confluent hypergeometric functions, *Constr. Approx.* 37(2) (2013) 195–221.

[6] Á. Baricz, D. Jankov, T.K. Pogány, Turán type inequalities for Krätzel functions, *J. Math. Anal. Appl.* 388(2) (2012) 716–724.

[7] R.D. Gordon, Values of Mills' ratio of area bounding ordinate and of the normal probability integral for large values of the argument, *Ann. Math. Statistics* 12 (1941) 364–366.

[8] E. Krätzel, Integral transformations of Bessel type, in: Generalized Functions and Operational Calculus, Proc. Conf. Varna 1975, Bulg. Acad. Sci, Sofia, 1979, 148–155.

[9] F. Oberhettinger, Tables of Mellin Transforms, Springer–Verlag, Berlin, 1974.

[10] M.B. Ruskai, E. Werner, Study of a class of regularizations of $1/|x|$ using Gaussian integrals, *SIAM J. Math. Anal.* 32(2) (2000) 435–463.

[11] T. Simon, Produit beta-gamma et regularité du sign, arXiv.1207.6464.

Followers' Strategy in Stackelberg Equilibrium Problems on Curved Strategy Sets

Alexandru Kristály

Óbuda University, Institute of Applied Mathematics, Budapest, Hungary
Babeş-Bolyai University, 400591 Cluj-Napoca, Romania
Email address: alexandru.kristaly@econ.ubbcluj.ro

Szilárd Nagy

Babeş-Bolyai University, 400591 Cluj-Napoca, Romania
Email address: szilard.nagy@econ.ubbcluj.ro

Abstract: We study the existence of Stackelberg equilibrium points on strategy sets which are geodesic convex in certain Riemannian manifolds by using metric projection arguments. The present results extend those obtained in Nagy [J. Global Optimization (2013)] in the Euclidean context.

Keywords: Stackelberg model; curved spaces; variational analysis

1 Introduction

Recently, the second author obtained certain existence and location results for the Stackelberg equilibria in the Euclidean framework, see [9]. More precisely, the existence of solutions for the leader-follower game has been obtained via the study of certain variational inequalities defined on the strategy sets by using the variational backward induction method.

The purpose of the present study is to extend the analytical results from [9] to games defined on strategy sets which are embedded in a geodesic convex manner into certain Riemannian manifolds. Similar studies can be found in the literature, where certain variational arguments are applied to study equilibrium problems on Riemannian manifolds, see [4], [7], [11], [10] and references therein.

For simplicity, in the present paper we shall consider only two players although our arguments can be extended to several players as well. Let $K_1 \subset M_1$ and $K_2 \subset M_2$ be two sets in the Riemannian manifolds (M_1, g_1) and (M_2, g_2), respectively, and let $h_1, h_2 : M_1 \times M_2 \to \mathbf{R}$ be the payoff functions for the two players. As we already know from the backward induction method, the first step (for the follower) is to find the response set

$$\mathscr{R}_{SE}(x_1) = \{x_2 \in K_2 : h_2(x_1, y) - h_2(x_1, x_2) \geq 0,\ \forall y \in K_2\}$$

for every fixed $x_1 \in K_1$. If $\mathscr{R}_{SE}(x_1) \neq \emptyset$ for every $x_1 \in K_1$, the next step (for the leader) is to minimize the map $x \mapsto h_1(x, r(x))$ on K_1 where r is a fixed selection function of the set-valued map $x \mapsto \mathscr{R}_{SE}(x)$; more precisely, the objective of the first player is to determine the set

$$\mathscr{S}_{SE} = \{x_1 \in K_1 : h_1(x, r(x)) - h_1(x_1, r(x_1)) \geq 0,\ \forall x \in K_1\}.$$

Since the location of the sets $\mathscr{R}_{SE}(x_1)$ and $\mathscr{S}_{SE}$ is not an easy task, we shall introduce further sets related to them by variational inequalities defined on the Riemannian manifolds. Let us assume that $h_2 : M_1 \times M_2 \to \mathbf{R}$ is a function of class C^1; for every $x_1 \in K_1$, we introduce the set

$$\mathscr{R}_{SV}(x_1) = \left\{x_2 \in K_2 : g_2\left(\frac{\partial h_2}{\partial x_2}(x_1, x_2), \exp_{x_2}^{-1}(y)\right) \geq 0,\ \forall y \in K_2\right\}.$$

Here and in the sequel, exp denotes the usual exponential function in Riemannian geometry. According to [4] and [5], it is more easier to determine the set $\mathscr{R}_{SV}(x_1)$ than $\mathscr{R}_{SE}(x_1)$. Moreover, usually we have that $\mathscr{R}_{SE}(x_1) \subset \mathscr{R}_{SV}(x_1)$, thus we shall choose the appropriate Stackelberg equilibrium candidates from the elements of the latter set. Finally, by imposing further curvature assumptions on the Riemannian manifolds we are working on, we are able to characterize the elements of the set $\mathscr{R}_{SV}(x_1)$ by the fixed points of a suitable set-valued map which involves the metric projection map into the set K_2. In fact, we shall assume that the strategy sets are embedded into non-positively curved Riemannian manifolds where two basic properties of the metric projection will be deeply exploited; namely, the non-expansiveness and the so-called Moskovitz-Dines property (see [8]); for further details, see Section 2. Having this fixed-point characterization, we will be able to apply various fixed point theorems on (acyclic) metric spaces in order to find elements of the set $\mathscr{R}_{SV}(x_1)$. We emphasize that projection-like methods for Nash equilibria have been developed in the Euclidean context in [1], [15], [16].

We assume finally that $h_1 : M_1 \times M_2 \to \mathbf{R}$ is a function of class C^1 and for every $x_1 \in K_1$ we have that $\mathscr{R}_{SV}(x_1) \neq \emptyset$. If we are able to choose a C^1-class selection $r : K_1 \to K_1$ of the set-valued map $\mathscr{R}_{SV}$, we also introduce the set

$$\mathscr{S}_{SV} = \left\{x_1 \in K_1 : g_1\left(\frac{\partial h_1}{\partial x_1}(x_1, r(x_1)), \exp_{x_1}^{-1}(y)\right) \geq 0,\ \forall y \in K_1\right\}.$$

In particular, $\mathscr{S}_{SV}$ contains the optimal strategies of the leader, i.e., the minimizers for the map $x \mapsto h_1(x, r(x))$ on K_1.

Section 2 contains some basic notions and results from Riemannian geometry which are needed for our investigations: geodesics, curvature, metric projections, Moskovitz-Dines property, etc. Finally, in Section 3 we present the main results of the paper concerning the strategy of the follower.

2 Preliminaries

2.1 Elements from Riemannian manifolds

Let (M,g) be a connected m-dimensional Riemannian manifold, $m \geq 2$, and let $TM = \cup_{p\in M}(p,T_pM)$ and $T^*M = \cup_{p\in M}(p,T_p^*M)$ be the tangent and cotangent bundles to M. If $\xi \in T_p^*M$ then there exists a unique $W_\xi \in T_pM$ such that

$$\langle \xi, V\rangle_{g,p} = g_p(W_\xi, V) \text{ for all } V \in T_pM. \tag{1}$$

Due to (1), the elements ξ and W_ξ are identified. The norms on T_pM and T_p^*M are defined by

$$\|\xi\|_g = \|W_\xi\|_g = \sqrt{g(W_\xi, W_\xi)}.$$

It is clear that for every $V \in T_pM$ and $\xi \in T_p^*M$,

$$|\langle \xi, V\rangle_g| \leq \|\xi\|_g \|V\|_g. \tag{2}$$

Let $h : M \to \mathbf{R}$ be a C^1 function at $p \in M$; the differential of h at p, denoted by $dh(p)$, belongs to T_p^*M and is defined by

$$\langle dh(p), V\rangle_g = g(\text{grad}h(p), V) \text{ for all } V \in T_pM.$$

Let $\gamma : [0,r] \to M$ be a C^1 path, $r > 0$. The length of γ is defined by $L_g(\gamma) = \int_0^r \|\dot\gamma(t)\|_g dt$. For any two points $p,q \in M$, let

$$d_g(p,q) = \inf\{L_g(\gamma) : \gamma \text{ is a } C^1 \text{ path joining } p \text{ and } q \text{ in } M\}.$$

The function $d_g : M \times M \to \mathbf{R}$ clearly verifies the properties of the metric function. For every $p \in M$ and $r > 0$, the open ball of center $p \in M$ and radius $r > 0$ is defined by

$$B_g(p,r) = \{q \in M : d_g(p,q) < r\}.$$

A C^∞ parameterized path γ is a geodesic in (M,g) if its tangent $\dot\gamma$ is parallel along itself, i.e., $\nabla_{\dot\gamma}\dot\gamma = 0$. Here, ∇ is the Levi-Civita connection. The geodesic segment $\gamma : [a,b] \to M$ is called minimizing if $L_g(\gamma) = d_g(\gamma(a),\gamma(b))$. From the theory of ODE we have that for every $V \in T_pM$, $p \in M$, there exists an open interval $I_V \ni 0$ and a unique geodesic $\gamma_V : I_V \to M$ with $\gamma_V(0) = p$ and $\dot\gamma_V(0) = V$. On account of [2, p. 64], we introduce the exponential map $\exp_p : T_pM \to M$ as $\exp_p(V) = \gamma_V(1)$. Moreover,

$$d\exp_p(0) = \text{id}_{T_pM}.$$

In particular, for every two points $q_1, q_2 \in M$ which are close enough to each other, we have

$$\|\exp_{q_1}^{-1}(q_2)\|_g = d_g(q_1,q_2). \tag{3}$$

Let $K \subset M$ be a non-empty set. Let

$$P_K(q) = \{p \in K : d_g(q,p) = \inf_{z\in K} d_g(q,z)\}$$

be the set of *metric projections* of the point $q \in M$ to the set K. According to the theorem of Hopf-Rinow, if (M,g) is complete, then for any closed set $K \subset M$ we have that $\text{card}(P_K(q)) \geq 1$ for every $q \in M$. The map P_K is *non-expansive* if

$$d_g(p_1,p_2) \leq d_g(q_1,q_2) \text{ for all } q_1,q_2 \in M \text{ and } p_1 \in P_K(q_1), p_2 \in P_K(q_2).$$

In particular, when P_K is non-expansive, then K is a Chebishev set, i.e., $\text{card}(P_K(q)) = 1$ for every $q \in M$.

The set $K \subset M$ is *geodesic convex* if every two points $q_1,q_2 \in K$ can be joined by a unique minimizing geodesic whose image belongs to K. Clearly, relation (3) holds for every $q_1,q_2 \in K$ in a geodesic convex set K since $\exp_{q_i}^{-1}$ is well-defined on K, $i \in \{1,2\}$. The function $f : K \to \mathbf{R}$ is *convex*, if $f \circ \gamma : [0,1] \to \mathbf{R}$ is convex in the usual sense for every geodesic $\gamma : [0,1] \to K$ once $K \subset M$ is a geodesic convex set.

A non-empty closed set $K \subset M$ verifies the *Moskovitz-Dines property* if for fixed $q \in M$ and $p \in K$ the following two statements are equivalent:

(MD_1) $p \in P_K(q)$;

(MD_2) If $\gamma : [0,1] \to M$ is the unique minimal geodesic from $\gamma(0) = p \in K$ to $\gamma(1) = q$, then for every geodesic $\sigma : [0,\delta] \to K$ $(\delta \geq 0)$ emanating from the point p, we have $g(\dot{\gamma}(0), \dot{\sigma}(0)) \leq 0$.

A Riemannian manifold (M,g) is a *Hadamard manifold* if it is complete, simply connected and its sectional curvature is non-positive. We recall that on a Hadamard manifold (M,g), if $h(p) = d_g^2(p,p_0)$, $p_0 \in M$ is fixed, then

$$\text{grad}h(p) = -2\exp_p^{-1}(p_0). \tag{4}$$

It is well-known that on a Hadamard manifold (M,g) every geodesic convex set is a Chebyshev set. Moreover, we have

Proposition 1. ([3], [13]) *Let (M,g) be a finite-dimensional Hadamard manifold, $K \subset M$ be a closed set. The following statements hold true:*

(i) *If $K \subset M$ is geodesic convex, it verifies the Moskovitz-Dines property;*

(ii) *P_K is non-expansive if and only if $K \subset M$ is geodesic convex.*

2.2 Basic properties of the response sets

In the sequel we shall establish some basic properties of the response sets by using some elements from the theory of variational inequalities on Riemannian manifolds.

Lemma 1. *Let (M_i,g_i) be Riemannian manifolds, $h_i : M_1 \times M_2 \to \mathbf{R}$ be functions of class C^1, and $K_i \subset M_i$ closed, geodesic convex sets, $i = 1,2$. Then the following assertions hold:*

(i) *$\mathscr{R}_{SE}(x_1) \subseteq \mathscr{R}_{SV}(x_1)$ for every $x_1 \in K_1$;*

(ii) *$\mathscr{R}_{SE}(x_1) = \mathscr{R}_{SV}(x_1)$ when $h_2(x_1,\cdot)$ is convex on K_2 for some $x_1 \in K_1$;*

(iii) $\mathscr{S}_{SE} \subseteq \mathscr{S}_{SV}$ *when* $x \mapsto \mathscr{R}_{SV}(x)$ *is a single-valued function which has a* C^1-*extension to an arbitrary open neighborhood* $D_1 \subset M_1$ *of* K_1.

Proof. (i) Let $x_2 \in \mathscr{R}_{SE}(x_1)$ be an arbitrarily fixed element, i.e., $h_2(x_1,y) \geq h_2(x_1,x_2)$ for all $y \in K_2$. By definition, we have that

$$g_2\left(\frac{\partial h_2}{\partial x_2}(x_1,x_2), \exp_{x_2}^{-1}(y)\right) = \lim_{t\to 0^+} \frac{h_2(x_1, \exp_{x_2}(t \exp_{x_2}^{-1}(y)) - h_2(x_1,x_2)}{t}, \ \forall y \in K_2.$$

Since K_2 is geodesic convex, the element $\exp_{x_2}(t\exp_{x_2}^{-1}(y) \in K_2$ for every $t \in [0,1]$ whenever $y \in K_2$. By the above expression one has that for every $y \in K$,

$$g_2\left(\frac{\partial h_2}{\partial x_2}(x_1,x_2), \exp_{x_2}^{-1}(y)\right) \geq 0,$$

which implies that $\mathscr{R}_{SE}(x_1) \subseteq \mathscr{R}_{SV}(x_1)$ for all $x_1 \in K_1$.

(ii) Since the function $h_2(x_1,.)$ is convex and of class C^1, one has

$$h_2(x_1,y) - h_2(x_1,x_2) \geq g_2\left(\frac{\partial h_2}{\partial x_2}(x_1,x_2), \exp_{x_2}^{-1}(y)\right)$$

for all $y \in K_2$, see [14]. Taking into account that $x_2 \in \mathscr{R}_{SV}(x_1)$, one has that

$$g_2\left(\frac{\partial h_2}{\partial x_2}(x_1,x_2), \exp_{x_2}^{-1}(y)\right) \geq 0$$

for all $y \in K_2$. Thus, one has $h_2(x_1,y) - h_2(x_1,x_2) \geq 0$ for all $y \in K_2$, i.e., $x_2 \in \mathscr{R}_{SE}(x_1)$.

(iii) The proof is similar to (i). △

In the sequel, we shall prove that the elements of the set $\mathscr{R}_{SV}(x_1)$ can be obtained as the fixed points of a carefully choosen map. More precisely, for a fixed $x_1 \in K_1$ and $\alpha > 0$, let $\mathscr{F}_\alpha^{x_1} : K_2 \to K_2$ be defined by

$$\mathscr{F}_\alpha^{x_1}(x) = P_{K_2}\left(\exp_x\left(-\alpha\frac{\partial h_2}{\partial x_2}(x_1,x)\right)\right). \tag{5}$$

Theorem 1. Let (M_1,g_1) be a Riemannian manifold, and (M_2,g_2) be a Hadamard manifold. Let $h_2 : M_1 \times M_2 \to \mathbf{R}$ be a function of class C^1 and $K_i \subset M_i$ closed, geodesic convex sets, $i = 1,2$. Let $x_1 \in K_1$. The following statements are equivalent:

(i) $x_2 \in \mathscr{R}_{SV}(x_1)$;

(ii) $\mathscr{F}_\alpha^{x_1}(x_2) = x_2$ for all $\alpha > 0$;

(iii) $\mathscr{F}_\alpha^{x_1}(x_2) = x_2$ for some $\alpha > 0$.

Proof. Let us fix $x_2 \in \mathscr{R}_{SV}(x_1)$ arbitrarily, where $x_1 \in K_1$. By definition, we have that

$$g_2\left(-\alpha\frac{\partial h_2}{\partial x_2}(x_1,x_2), \exp_{x_2}^{-1}(y)\right) \leq 0, \ \forall y \in K_2,$$

for all/some $\alpha > 0$. Let $\gamma, \sigma : [0,1] \to M_2$ be the unique minimal geodesics defined by

$$\gamma(t) = \exp_{x_2}(-t\alpha \frac{\partial h_2}{\partial x_2}(x_1,x_2))$$

and

$$\sigma(t) = \exp_{x_2}(t \exp_{x_2}^{-1}(y))$$

for any fixed $\alpha > 0$ and $y \in K_2$. Since K_2 is geodesic convex in (M_2, g_2), then $\mathrm{Im}\sigma \subset K_2$ and

$$g_2(\dot{\gamma}(0), \dot{\sigma}(0)) = g_2\left(-\alpha \frac{\partial h_2}{\partial x_2}(x_1,x_2), \exp_{x_2}^{-1}(y)\right),$$

i.e., (MD_2) holds. By the Moskovitz-Dines property, see Proposition 1, one has that

$$x_2 = \gamma(0) \in P_{K_2}(\gamma(1)) = P_{K_2}\left(\exp_{x_2}\left(-\alpha \frac{\partial h_2}{\partial x_2}(x_1,x_2)\right)\right) = \mathscr{F}_\alpha^{x_1}(x_2).$$

Since $\mathrm{card}(\mathscr{F}_\alpha^{x_1}(x_2)) = 1$, the proof is complete. △

Remark. Note that for all $\alpha > 0$,

$$\mathscr{R}_{SV}(x_1) = \left\{x_2 \in K_2 : P_{K_2}\left(\exp_{x_2}\left(-\alpha \frac{\partial h_2}{\partial x_2}(x_1,x_2)\right)\right) = x_2\right\}.$$

3 Follower strategy: existence of equilibria

3.1 Compact case

Theorem 2. **(Compact case)** Let (M_i, g_i) be Hadamard manifolds, $h_i : M_1 \times M_2 \to \mathbf{R}$ be functions of class C^1 and $K_i \subset M_i$ compact, geodesic convex sets, $i = 1,2$. Then the following statements hold:

(i) $\mathscr{R}_{SV}(x_1) \neq \emptyset$ for every $x_1 \in K_1$;

(ii) $\mathscr{S}_{SV} \neq \emptyset$, whenever $\mathscr{R}_{SV}(x_1)$ is a singleton for every $x_1 \in K_1$ and the map $x \mapsto \mathscr{R}_{SV}(x)$ has a C^1-extension to an arbitrary open neighborhood $D_1 \subset M_1$ of K_1.

Proof. (i) Fix $x_1 \in K_1$ and $\alpha > 0$. Since K_2 is a Chebishev set and P_{K_2} is globally Lipschitz, we see that $\mathscr{F}_\alpha^{x_1} : K_2 \to K_2$ is a single-valued continuous function; in particular, $\mathscr{F}_\alpha^{x_1} : K_2 \to K_2$ has a closed graph. Moreover, since K_2 is geodesic convex, it is contractible, thus an acyclic set. Now, we may apply the fixed point theorem of Begle on the compact set K_2, obtaining that $\mathscr{F}_\alpha^{x_1}$ has at least a fixed point $x_2 \in K_2$. Due to Theorem 1, $x_2 \in \mathscr{R}_{SV}(x_1)$, which concludes the proof of (i).

(ii) For some $\beta > 0$, we introduce the map $\mathscr{G}_\beta : K_1 \to K_1$ defined by

$$\mathscr{G}_\beta(x) = P_{K_1}\left(\exp_x\left(-\beta \frac{\partial h_1}{\partial x}(x, \mathscr{R}_{SV}(x))\right)\right).$$

Since $\mathrm{card}(\mathscr{R}_{SV}(x)) = 1$ for every $x \in K_1$ and the map $x \mapsto \mathscr{R}_{SV}(x)$ has a C^1-extension to an arbitrary $D_1 \subset M_1$ of K_1, the function $\mathscr{G}_\beta$ is well-defined for every $\beta > 0$. By the hypotheses, the function $\mathscr{G}_\beta$ is also continuous, thus on account of the Belge fixed point theorem, there exits at least $x_1 \in K_1$ such that $\mathscr{G}_\beta(x_1) = x_1$. Since (M_1, g_1) is a Hadamard manifold where the Moskovitz-Dines property holds, an analogous argument as in Theorem 1 shows that $\mathscr{G}_\beta(x_1) = x_1$ is equivalent to $x_1 \in \mathscr{S}_{SV}$. The proof is complete. △

3.2 Non-compact case

When the strategy sets are non-compact, certain growth assumptions are needed on the payoff functions in order to guarantee the existence of Stackelberg equilibria. We first assume that for some $x_1 \in K_1$ one has

$(H_{x_1}^{h_2})$ There exists $x_2 \in K_2$ such that

$$L_{x_1,x_2} = \limsup_{d_{g_2}(x,x_2)\to\infty,\ x\in K_2} \frac{g_2\left(\frac{\partial h_2}{\partial x_2}(x_1,x), \exp_x^{-1}(x_2)\right) + g_2\left(\frac{\partial h_2}{\partial x_2}(x_1,x_2), \exp_{x_2}^{-1}(x)\right)}{d_{g_2}(x,x_2)} <$$

$$< -\left\| \frac{\partial h_2}{\partial x_2}(x_1,x_2) \right\|_{g_2}.$$

Theorem 3. Let (M_1, g_1) be a Riemannian manifold, and (M_2, g_2) be a Hadamard manifold. Let $h_2 : M_1 \times M_2 \to \mathbf{R}$ be a function of class C^1 and $K_i \subset M_i$ closed, geodesic convex sets, $i = 1, 2$. Let $x_1 \in K_1$ and assume that hypothesis $(H_{x_1}^{h_2})$ holds true. Then $\mathscr{R}_{SV}(x_1) \neq \emptyset$.

Proof. Let $E_0 \in \mathbf{R}$ such that

$$L_{x_1,x_2} < -E_0 < -\left\| \frac{\partial h_2}{\partial x_2}(x_1,x_2) \right\|_{g_2}.$$

On account of hypothesis$(H_{x_1}^{h_2})$ there exists $R > 0$ large enough such that for every $x \in K_2$ with $d_{g_2}(x,x_2) \geq R$, we have

$$g_2\left(\frac{\partial h_2}{\partial x_2}(x_1,x), \exp_x^{-1}(x_2)\right) + g_2\left(\frac{\partial h_2}{\partial x_2}(x_1,x_2), \exp_{x_2}^{-1}(x)\right) \leq -E_0 d_{g_2}(x,x_2).$$

Clearly, one may assume that $K_2 \cap \overline{B}_{g_2}(x_2,R) \neq \emptyset$. In particular, from (3) and (2), for every $x \in K_2$ with $d_{g_2}(x,x_2) \geq R$, the above relation yields

$$\begin{aligned} g_2\left(\frac{\partial h_2}{\partial x_2}(x_1,x), \exp_x^{-1}(x_2)\right) &\leq -E_0 d_{g_2}(x,x_2) \\ &\quad + \left\| \frac{\partial h_2}{\partial x_2}(x_1,x_2) \right\|_{g_2} \|\exp_{x_2}^{-1}(x)\|_{g_2} \qquad (6) \\ &= \left(-E_0 + \left\| \frac{\partial h_2}{\partial x_2}(x_1,x_2) \right\|_{g_2}\right) d_{g_2}(x,x_2) \\ &< 0. \end{aligned}$$

Let $K_R = K_2 \cap \overline{B}_{g_2}(x_2,R)$. It is clear that K_R is a geodesic convex, compact subset of (M_2,g_2). Due to Theorem 2, we immediately have that there exists $\tilde{x}_2 \in K_R$ such that

$$g_2\left(\frac{\partial h_2}{\partial x_2}(x_1,\tilde{x}_2),\exp_{\tilde{x}_2}^{-1}(y)\right) \geq 0 \text{ for all } y \in K_R. \tag{7}$$

Note that $d_{g_2}(\tilde{x}_2,x_2) < R$. By assuming the contrary, from (6) with $x = \tilde{x}_2$ we have that

$$g_2\left(\frac{\partial h_2}{\partial x_2}(x_1,\tilde{x}_2),\exp_{\tilde{x}_2}^{-1}(x_2)\right) < 0,$$

by contradicting relation (7).

Let us choose $z \in K_2$ arbitrarily. From the fact that $d_{g_2}(\tilde{x}_2,x_2) < R$, for $\varepsilon > 0$ small enough, the element $y = \exp_{\tilde{x}_2}(\varepsilon \exp_{\tilde{x}_2}^{-1}(z))$ belongs both to $K_2 \cap \overline{B}_{g_2}(x_2,R) = K_R$. By replacing y into (7), we obtain that

$$g_2\left(\frac{\partial h_2}{\partial x_2}(x_1,\tilde{x}_2),\exp_{\tilde{x}_2}^{-1}(z)\right) \geq 0.$$

Since $z \in K_2$ is arbitrarily fixed, one has that $\tilde{x}_2 \in \mathscr{R}_{SV}(x_1)$, which ends the proof. $\triangle$

In the sequel, we are dealing with another class of functions. For a fixed $x_1 \in K_1$, $\alpha > 0$ and $0 < \rho < 1$ we introduce the hypothesis:

$$(H_{x_1}^{\alpha,\rho}): \qquad d_{g_2}\left(\exp_x\left(-\alpha\frac{\partial h_2}{\partial x_2}(x_1,x)\right),\exp_y\left(-\alpha\frac{\partial h_2}{\partial x_2}(x_1,y)\right)\right) \leq$$
$$\leq (1-\rho)d_{g_2}(x,y) \text{ for all } x,y \in K_2.$$

For fixed $x_1 \in K_1$ and $\alpha > 0$, we consider the following two dynamical systems:

(a) let $(DDS)_{x_1}$ be the discrete differential system in the form

$$\begin{cases} y_{n+1} = \mathscr{F}_\alpha^{x_1}(P_{K_2}(y_n)), \quad n \geq 0, \\ y_0 \in M_2; \end{cases}$$

(b) Let $(CDS)_{x_1}$ be the continuous differential system in the form

$$\begin{cases} \frac{dy}{dt} = \exp_{y(t)}^{-1}(\mathscr{F}_\alpha^{x_1}(P_{K_2}(y(t)))), \\ y(0) = x_2 \in M_2. \end{cases}$$

The main result of the present section is the following theorem.

Theorem 4 (Non-compact case). Let (M_1,g_1) be a Riemannian manifold, and (M_2,g_2) be a Hadamard manifold. Let $h_2 : M_1 \times M_2 \to \mathbf{R}$ be a function of class C^1 and $K_i \subset M_i$ closed, geodesic convex sets, $i = 1,2$. Let $x_1 \in K_1$ and assume that hypothesis $(H_{x_1}^{\alpha,\rho})$ holds true. Then $\mathscr{R}_{SV}(x_1)$ is a singleton and both dynamical systems, $(DDS)_{x_1}$ and $(CDS)_{x_1}$, exponentially converge to the unique element of $\mathscr{R}_{SV}(x_1)$.

Proof. Since (M_2,g_2) is a Hadamard manifold, for the geodesic convex set $K_2 \subset M_2$ we have that P_{K_2} is non-expansive. Therefore, by $(H_{x_1}^{\alpha,\rho})$, one has for every $x,y\in K_2$ that

$$\begin{aligned}
&d_{g_2}(\mathscr{F}_\alpha^{x_1}(x),\mathscr{F}_\alpha^{x_1}(y))\\
&= d_{g_2}\left(P_{K_2}\left(\exp_x\left(-\alpha\frac{\partial h_2}{\partial x_2}(x_1,x)\right)\right),P_{K_2}\left(\exp_y\left(-\alpha\frac{\partial h_2}{\partial x_2}(x_1,y)\right)\right)\right)\\
&\le d_{g_2}\left(\exp_x\left(-\alpha\frac{\partial h_2}{\partial x_2}(x_1,x)\right),\exp_y\left(-\alpha\frac{\partial h_2}{\partial x_2}(x_1,y)\right)\right)\\
&\le (1-\rho)d_{g_2}(x,y).
\end{aligned}$$

Consequently, the function $\mathscr{F}_\alpha^{x_1}$ is a $(1-\rho)-$contraction on K_2.

(a) *The system* $(DDS)_{x_1}$. We shall apply the Banach fixed point theorem to the function $\mathscr{F}_\alpha^{x_1}: K_2\to K_2$, by guaranteeing the existence of the unique fixed point of $\mathscr{F}_\alpha^{x_1}$ for every $x_1\in K_1$. Moreover, every iterated sequence in the dynamical system $(DDS)_{x_1}$ converges exponentially to the unique fixed point $x_2\in K_2$ of $\mathscr{F}_\alpha^{x_1}$. Due to Theorem 1 the set $\mathscr{R}_{SV}(x_1)$ is a singleton with the element x_2. Moreover, for all $k\in\mathbf{N}$ we have that

$$d_{g_2}(y_k,x_2)\le\frac{(1-\rho)^k}{\rho}d_{g_2}(y_1,y_0).$$

(b) *The system* $(CDS)_{x_1}$. First of all, standard ODE theory shows that $(CDS)_{x_1}$ has a (local) solution in $[0,T)$. We actually prove that $T=+\infty$. To see this fact, we assume that $T<+\infty$, and we introduce the Lyapunov function which has the form

$$h_{x_1}(t)=\frac{1}{2}d_{g_2}(y(t),x_2)^2.$$

Note that for a.e. $t\in[0,T)$, we have

$$\begin{aligned}
\frac{d}{dt}h_{x_1}(t) &= -g_2\left(\exp_{y(t)}^{-1}(x_2),\frac{dy}{dt}\right)\\
&= -g_2\left(\exp_{y(t)}^{-1}(x_2),\exp_{y(t)}^{-1}(\mathscr{F}_\alpha^{x_1}(P_{K_2}(y(t))))\right)\\
&= -g_2\left(\exp_{y(t)}^{-1}(x_2),\exp_{y(t)}^{-1}(\mathscr{F}_\alpha^{x_1}(P_{K_2}(y(t))))-\exp_{y(t)}^{-1}(x_2)\right)\\
&\quad -g_2\left(\exp_{y(t)}^{-1}(x_2),\exp_{y(t)}^{-1}(x_2))\right)\\
&\le \|\exp_{y(t)}^{-1}(\mathscr{F}_\alpha^{x_1}(P_{K_2}(y(t))))-\exp_{y(t)}^{-1}(x_2)\|_{g_2}\|\exp_{y(t)}^{-1}(x_2)\|_{g_2}\\
&\quad -\|\exp_{y(t)}^{-1}(x_2)\|_{g_2}^2.
\end{aligned}$$

By using the fact that (M_2,g_2) is a Hadamard manifold, a Rauch comparison theorem and further straightforward estimates show that

$$\|\exp_{y(t)}^{-1}(\mathscr{F}_\alpha^{x_1}(P_{K_2}(y(t))))-\exp_{y(t)}^{-1}(x_2)\|_{g_2}\le d_{g_2}(\mathscr{F}_\alpha^{x_1}(P_{K_2}(y(t))),x_2).$$

Therefore, by (3) and the non-expansiveness of P_{K_2}, we have

$$\begin{aligned}
\frac{d}{dt}h_{x_1}(t) &\leq d_{g_2}(\mathscr{F}_\alpha^{x_1}(P_{K_2}(y(t))),x_2)d_{g_2}(y(t),x_2)-d_{g_2}(y(t),x_2)^2\\
&= d_{g_2}(\mathscr{F}_\alpha^{x_1}(P_{K_2}(y(t))),\mathscr{F}_\alpha^{x_1}(x_2))d_{g_2}(y(t),x_2)-d_{g_2}(y(t),x_2)^2\\
&\leq (1-\rho)d_{g_2}(P_{K_2}(y(t)),x_2)d_{g_2}(y(t),x_2)-d_{g_2}(y(t),x_2)^2\\
&\leq (1-\rho)d_{g_2}(y(t),x_2)^2-d_{g_2}(y(t),x_2)^2\\
&= -\rho d_{g_2}(y(t),x_2)^2\\
&= -2\rho h_{x_1}(t), \text{ a.e. } t\in[0,T).
\end{aligned}$$

Therefore, one has

$$\frac{d}{dt}[h_{x_1}(t)e^{2\rho t}]=\left(\frac{d}{dt}h_{x_1}(t)+2\rho h_{x_1}(t)\right)e^{2\rho t}\leq 0.$$

In particular, the function $t\mapsto h_{x_1}(t)e^{2\rho t}$ is non-increasing; therefore, for all $t\in[0,T)$ one has that $h_{x_1}(t)e^{2\rho t}\leq h_{x_1}(0)$. Consequently, $t\mapsto y(t)$ can be extended beyond T, contradicting our assumption. Therefore, $T=+\infty$.

The above estimate gives that for every $t\geq 0$, $h_{x_1}(t)\leq h_{x_1}(0)e^{-2\rho t}$. In particular, it yields that

$$d_{g_2}(y(t),x_2)\leq d_{g_2}(y_0,x_2)e^{-\rho t}.$$

The proof is concluded. △

Remark. Assume that $M_i=\mathbf{R}^{m_i}$, $i=1,2$ and $\frac{\partial f_2}{\partial x_2}(x_1,\cdot)$ is an $\lambda-$Lipschitz and $\sigma-$strictly monotone function for some $x_1\in K_1$, i.e.,

- $\|\frac{\partial f_2}{\partial x_2}(x_1,x)-\frac{\partial f_2}{\partial x_2}(x_1,y)\|\leq\lambda\|x-y\|$,
- $\langle\frac{\partial f_2}{\partial x_2}(x_1,x)-\frac{\partial f_2}{\partial x_2}(x_1,y),x-y\rangle\geq\sigma\|x-y\|^2$, $\forall x,y\in\mathbf{R}^{m_2}$.

In this case, $(H_{x_1}^{\varepsilon,\rho})$ holds true with

$$0<\varepsilon<\frac{\sigma-\sqrt{(\sigma^2-\lambda^2)_+}}{\lambda^2}$$

and

$$\rho=1-\sqrt{1-2\varepsilon\sigma+\varepsilon^2\lambda^2}\in(0,1).$$

Remark. Very recently, Kristály and Repovs [6] proved that the Moskovitz-Dines property on a generic Riemannian manifold implies the non-positiveness of the sectional curvature. Consequently, in order to develop the aforementioned results on 'curved' spaces, the non-positiveness of the sectional curvature seems to be a natural requirement.

Remark. By following the non-smooth critical point theory of Szulkin [12], it would be interesting to guarantee not only the existence of Stackelberg equilibrium points but also some multiplicity results. Here, the indicator function of geodesic convex sets as well as the Fréchet subdifferential of the indicator function (as the normal cone to the geodesic convex set) seem to play crucial roles which will be investigated in a forthcoming paper.

Acknowledgement

The work of Szilárd Nagy is supported by the grant PCCE-55/2008.

References

[1] E. Cavazzuti, M. Pappalardo, M. Passacantando, *Nash equilibria, variational inequalities, and dynamical systems,* J. Optim. Theory Appl. 114 (2002), no. 3, 491–506.

[2] M. P. do Carmo, *Riemannian Geometry*, Birkhäuser, Boston, 1992.

[3] S. Grognet, *Théorème de Motzkin en courbure négative,* Geom. Dedicata 79 (2000), 219–227.

[4] A. Kristály, *Location of Nash equilibria: a Riemannian geometrical approach,* Proc. Amer. Math. Soc. 138 (2010), 1803-1810.

[5] A. Kristály, V. Rădulescu, Cs. Varga, *Variational Principles in Mathematical Physics, Geometry, and Economics,* Cambridge University Press, Encyclopedia of Mathematics and its Applications, No. 136, 2010.

[6] A. Kristály, D. Repovs, *Metric Projections versus Non-Positive Curvature,* preprint, 2013.

[7] C. Li, G. López, V. Martín-Márquez, *Monotone vector fields and the proximal point algorithm on Hadamard manifolds,* J. Lond. Math. Soc. (2) 79 (2009), no. 3, 663–683.

[8] D. Moskovitz, L.L. Dines, *Convexity in a linear space with an inner product,* Duke Math. J. 5 (1939) 520–534.

[9] Sz. Nagy, *Stackelberg equilibria via variational inequalities and projections,*J. Global Optimization, in press (2013).

[10] S.Z. Németh, *Variational inequalities on Hadamard manifolds,* Nonlinear Analysis 52 (2003), 1491-1498.

[11] J.-S. Pang, M. Fukushima, *Quasi-variational inequalities, generalized Nash equilibria, and multi-leader-follower games,* CMS 2(2005), 21-56. DOI: 10.1007/s10287-004-0010-0

[12] A. Szulkin, *Minimax prin- ciples for lower semicontinuous functions and applications to nonlinear boundary value problems.* Ann. Inst. H. Poincaré Anal. Non Linéaire 3 (1986), no. 2, 77-109.

[13] R. Walter, *On the metric projection onto convex sets in Riemannian spaces,* Arch. Math. (Basel) 25 (1974), 91–98.

[14] C. Udrişte, *Convex Functions and Optimization Methods on Riemannian Manifolds,* Mathematics and its Applications, 297. Kluwer Academic Publishers Group, Dordrecht, 1994.

[15] Y.S. Xia, J. Wang, *On the stability of globally projected dynamical systems,* J. Optim. Theory Appl. 106 (2000), no. 1, 129–150.

[16] J. Zhang, B. Qu, N. Xiu, *Some projection-like methods for the generalized Nash equilibria,* Comput. Optim. Appl. 45 (2010), 89-109.

Solving Integer and Mixed Integer Linear Problems with ABS Method

József Abaffy

John von Neumann Faculty of Informatics, Óbuda University
Bécsi út 96/b, 1034 Budapest, Hungary
abaffy.jozsef@nik.uni-obuda.hu

Szabina Fodor

Department of Computer Science, Corvinus University of Budapest
Fővám tér 13-15, 1093 Budapest, Hungary
szabina.fodor@uni-corvinus.hu

Abstract: Solving mixed integer linear programming (MILP) problems is a difficult task due to the parallel use of both integer and non-integer values. One of the most widely used solution is to solve the problem in the real space and they apply additional iteration steps (so-called cutting-plane algorithms or Gomory's cuts) to narrow down the solution to the optimal integer solution. The ABS class of algorithms is a generalized class of algorithms which, with appropriate selection of parameters, is suitable for the solution of both integer and non-integer linear problems. Here we provide for the first time a complete ABS-based algorithm for MILP problems by adaptation of the ABS approach to Gomory's cutting-plane algorithm. We also provide a numerical example demonstrating the working principle of our algorithm.

Keywords: linear programming; ABS methods; mixed integer problem; cutting-plane methods; Gomory's cuts

1 Introduction

Mixed Integer Linear Programming (MILP) problems are linear programming problems in which some but not all elements of the solution vector are integer. They can be formulated in the following general form:

$$\min\{c^T x \ : \ Ax \geq b, x \geq 0, \ x_j \in Z \ \forall j \in I\}, \tag{1}$$

where no assumption related to the structure of the matrix A is made but the set **I** of the **x** variables need to be integer. MILP problems arise during everyday life whenever continuous and discrete parameters need to be optimized, ranging from basic business decisions through traffic control to guiding unmanned aerial vehicles. Despite their major importance, there is no perfect solution for MILP problems. In particular, because of the integer nature of some elements of the solution vector, general MILP algorithms are nondeterministic polynomial time (NP) hard algorithms which are not very effective in practice. Therefore, various approaches have been developed to solve MILP problems in polynomial or quasi-polynomial time.

One of the possible approaches to overcome the NP-hard nature of MILP problems is to first solve the same problem in the real space, i. e. without any constraints on whether elements of the solution vector need to be integer or not. Such modified problems are called the LP relaxation of the original problem and can be described in the following general form:

$$\min\{c^T x \,:\, Ax \geq b,\ x \geq 0\}, \tag{2}$$

The advantage of this approach is that (2) can be solved by general linear programming (LP) applications in polynomial time which are much more effective in practice. However, those solutions contain both integer and non-integer solution values which need to be separated (so-called separation problem). Since the condition of integer nature has to be met, the optimum solution has to be identified in a second step where the optimum solution vector is narrowed down to values that meet the integer requirement. This is performed by establishing a new condition that is only satisfied if the solution matrix meets the relevant integer requirement. With other words, solutions that do not satisfy the integer requirements are "cut out" of the resulting solution matrix. Therefore, such algorithms called "cutting-plane algorithms", "cuts" or, according to its first description, "Gomory's cuts". During a cutting-plane algorithm, several iteration steps are used to refine the solution matrix in order to find the optimum integer solution (a solution to the separation problem). Cutting planes are inequalities that solve the separation problem and. Such cutting planes serve to tighten the so-called LP relaxation resulting in better approximation of the convex hull of the original MILP problem. All current commercial MILP problem solving algorithms apply Gomory's cutting-plane algorithm to find the optimal integer solution.

Historically, Gomory first described an algorithm that finds the optimal solution in finite iteration steps for Integer Linear Programming (IP) in 1958 [1]. Such an algorithm solves the separation problem when $\mathbf{x}^*$ is an optimal basis of the LP relaxation. In 1960, Gomory introduced the Gomory Mixed Integer (GMI) cuts to deal with the mixed-integer case [2]. However, he never emphasized the practical use of this method, since the cutting plane algorithms converge very slowly to the optimum solution and the resulting large number of cuts results in very large LPs with corresponding numerical difficulties. However, a major improvement came from Balas et al. [3] who re-analyzed the original Gomory mixed-integer cuts [2] and overruled the common belief that these cuts had no practical importance. In fact, a series of improvements eventually led the same group to show that Gomory ‘s cuts are fundamental tools for the solution of 0-1 MILP problems [4]. However, given the major importance of MILP problems, additional approaches to solving such problems or performing Gomory’s cuts are still actively needed.

The ABS class of algorithms are generalized algorithms which, with appropriate selection of parameters, can be used to solve diverse mathematical problems. They were initially developed by Abaffy, Broyden and Spedicato [5-7] to solve linear systems of equations over the real space. The class was later generalized (so-called scaled ABS class) and applied also to the solution of various additional linear and nonlinear problems [8]. The ABS algorithm was applied to mathematical optimization problems such as LP problems via a certain subclass of ABS (called implicit LX) by reformulating the simplex method [9, 10]. However, those studies did not address the problem of finding an initial basis and an initial feasible solution.

It is theoretically possible that ABS-based algorithms may also be able to provide suitable solutions for MILP problems. If so, then the algorithm could take advantage of various unique features of the ABS class. For example, ABS algorithms have n inherent capability of finding cutting planes that are linearly dependent, which is a major obstacle in the algorithm presented by [4]. However, no ABS-based algorithms have yet been reported that are capable of solving an entire MILP problem.

In this paper we present a new method for solving mixed-integer problems by applying the ABS approach to Gomory’s cutting plane algorithm. Since no ABS-based methods to finding the initial basis and initial feasible solution for the simplex method have yet been described, we first present an ABS-based solution for finding those parameters. In parallel, we construct the projection matrix H of the ABS class. Together with the above mentioned LX method for the ABS-based reformulation of the simplex method, these results now allow the ABS-based solution of LP problems. Next we describe a new method by applying the ABS class to Gomory’s cutting plane methods. Those components are placed into a frame allowing the solution of MILP problems. Finally, we provide a numerical example to illustrate the working principles of our algorithm.

2 ABS Algorithm for Solving LP Problems

Let us consider the following modified system

$$\begin{aligned} &\min\{e^T u\} \\ &Ax + Iu = b \\ &x, u \geq 0 \end{aligned} \tag{3}$$

where **e** is the vector of all ones, and **b≥0**¸ (if not, then we can multiply the constraints by **-1** to achieve this) and the **u** are artificial (slack) variables. Define $\tilde{x} = \left[x^T \ u^T\right]^T$ and $\tilde{A} = [A\ I]$ so that the constraints of the modified can be written as $\tilde{A}\tilde{x} = b, \tilde{x} \geq 0$.

Let **B** be the indices corresponding to the artificial variables. Then **B** is a basis, since the corresponding columns of $\tilde{A}$ are I, the identity, and thus linearly independent. The corresponding basic feasible solution is $x = 0$, $u = b$. We use this to initialize the necessary parameters (i.e. the projection matrix) for the ABS-based simplex algorithm.

Algorithm 1: ***Finding an initial feasible solution***

(A1) Let $\tilde{x}_1 \in R^n$ be arbitrary, $\tilde{x}_1 = 0$, **i=1**, and $H_1 = I$, where $I \in R^{n,n}$ unit matrix.

(B1) Calculate the following vectors

$$s_i = H_i \tilde{a}_i, \tag{4}$$

$$p_i = \tilde{a}_i^T \tilde{x}_i - b_i. \tag{5}$$

If $s_i \neq 0$, then go to C1.

If $s_i = 0$ and $p_i = 0$ then $\tilde{x}_{i+1} = \tilde{x}_i$, $H_{i+1} = H_i$ go to F1. (The *ith* equation linearly depends on the previous ones.)

(C1) Compute the search vector

$p_i = H_i^T e_{m+i}$, where e_{m+i} is the **m+ith** unit vector.

(D1) Update the approximation of the solution by

$\tilde{x}_{i+1} = \tilde{x}_i - \alpha_i p_i$, where $\alpha_i = \dfrac{r_i}{\tilde{a}_i^T p_i}$.

(E1) Update the H_i matrix by

$$H_{i+1} = H_i - \frac{s_i p_i^T}{p_i^T \tilde{a}_i}$$

(F1) If **i=m**, then STOP. $\tilde{x}_{m+1}$ is a solution of the system.

If **i≠m**, then increment the index **i** by one and go to B1.

Remark 2.1 The algorithm is well-defined as the conditions $e_{m+i}^T s_i \neq 0$ and the $\tilde{a}_i^T p_i \neq 0$ are trivially true.

Remark 2.2 The original ABS algorithm contains a case $s_i = 0$ and $r_i \neq 0$ in step B1, which means the incompatibility of the system of equations. This never happens in our case as our system has the obvious solution $x^T = \left[\mathbf{0},\ldots,\mathbf{0},\mathbf{b}^\mathbf{T}\right]$.

Remark 2.3 The H_i projection matrices, generated by Algorithm 1, are Hermitian and they have the following special structure

$$\begin{bmatrix} \mathbf{0} & & \mathbf{\ldots} & & \mathbf{0} \\ \cdot & & & & \cdot \\ \mathbf{0} & & \mathbf{\ldots} & & \mathbf{0} \\ * & \mathbf{\ldots} & * & \mathbf{1} & \mathbf{\ldots} & \mathbf{1} \\ \cdot & & \cdot & \cdot & & \cdot \\ * & \mathbf{\ldots} & * & \mathbf{0} & \mathbf{..} & \mathbf{1} \end{bmatrix},$$

where * indicates possible non-zero elements. Furthermore, the indexes of the zero rows are the basis elements, and the indexes of the non-zero rows are the non-basis ones. [10]

Remark 2.4 In general, finding an initial basis for the standard problem is as difficult as finding an optimal solution for the original problem. Please refer to Abaffy et al. [11] for finding an initial feasible solution, where the initial bases and the **H** projection matrix are parallely calculated saving a number of operations.

Let's use the following notation. The indexes of the non-identically zero rows of the $\mathbf{H_i}$ matrix is $\mathbf{B_i}$ and the indexes of the zero rows of $\mathbf{H_i}$ is $\mathbf{N_i}$.

The simplex algorithm performs successive iteration steps (pivot operations) to gradually improve the feasible intermediate solution.

Once the pivot column has been selected, the choice of pivot row is largely determined by the requirement that the resulting solution is feasible. This means using the ABS terminology that we need to minimize the expression $\frac{x^T e_k}{e_k^T H_i^T e_{N^*}} \mid k \in B_i$ such that $e_k^T H_i^T e_{N^*} > 0$ [10].

The ABS formulation of the simplex method is defined by the following procedure.

Algorithm 2: ***ABS based simplex method (Finding the optimal solution in real space)***

(A2) Let $x_1 \in R^n$ be a feasible solution of problem (2). H_1 is the projection matrix for this feasible solution, and **i=1**.

(B2) Compute the search vector $\mathbf{p_i}$ by

$p_i = H_i^T e_{N^*}$, where e_{N^*} is unit vector, where $c^T H_i^T e_{N^*} = \min\{c^T H_i^T e_j \mid j \in N_i\}$.

(C2) Update the solution

$x_{i+1} = x_i - \alpha_i p_i$, where $\alpha_i = \frac{-x_{B^*}}{e_{B^*}^T H_i^T e_{N^*}} = \min\{\frac{-x_k}{e_k^T H_i^T e_{N^*}} \mid k \in B_i \; and \; e_k^T H_i^T e_{N^*} > 0\}$.

(D2) Update the projection matrix

$$H_{i+1} = H_i - \frac{(H_i e_{B^*} - e_{B^*}) e_{N^*}^T H_i}{e_{N^*}^T H e_{B^*}}$$

(E2) If $c^T H_{i+1}^T > 0$ then STOP. $\mathbf{x_{i+1}}$ is the optimal solution.

If $c^T H_{i+1}^T$ vector has negative element, then the $\mathbf{x_{i+1}}$ solution is not optimal, $\mathbf{H_i}$=$\mathbf{H_{i+1}}$, $\mathbf{x_i}$=$\mathbf{x_{i+1}}$and go to step B2.

Remark 2.5 Computing the $\mathbf{p_i}$ vector means that we determine the entering variable into the basis in step B5 and updating the solution with the selected $\mathbf{e_B}^*$ means that we select the leaving variable from the basis.

Remark 2.6 The selection of the entering variable in step B2 is taken as the column with least relative cost. In the ABS approach it corresponds to the minimization of the expression$c^T H_i^T e_j$. However, the minimization can be changed to maximization or other selection strategy.

Remark 2.7 Residual cost vector is $r = H_i c$ in every iteration step [10].

There are two conceptually different approaches to solving LP problems in real space. The simplex algorithm first finds a basic solution that is feasible (i. e. the solution is nonnegative) and the following iteration steps refine this solution towards the optimum solution while maintaining feasibility of the intermediate solutions throughout the entire procedure. In contrast, the dual simplex algorithm first finds a basic solution that is primal infeasible (there are certain negative values) but dual feasible and the following iteration steps are similarly feasible in the dual but not in the primal case except for the last iteration step in which the final solution will be both primal and dual feasible. With other words, the simplex algorithm performs the entire iteration procedure in the primal feasible space whereas the dual simplex algorithm does so in the dual feasible (but primal infeasible) space and only the final step will ensure primal feasibility. Nevertheless, both algorithms are able to find the same final (optimal) solution.

An important feature of the dual simplex algorithm is that it is most suitable to solve problems where a dual feasible solution can easily be found, or when additional conditions (change of parameters, additional constraints) are set after having obtained an initial fasible solution for the original problem.

As mentioned above, Algorithm 2 finds a feasible optimum solution for problem (3) in the real space using the principles of the simplex algorithm. In the following section we re-formulate the ABS algorithm to also perform the dual simplex method in the real space (Algorithm 3). This algorithm will then be used to re-optimize the intermediate solution following the introduction of a new integer condition in Algorithm 4.

Algorithm 3: ***ABS based dual simplex method (Re-optimizing with dual simplex method)***

(A3) Let $x_1 \in R^n$ a dual feasible solution of the problem (2), H_1 is the projection matrix for this dual feasible solution, and **i=1**.

(B3) (*Selection of the leaving variable.*) Find an index (**N**$^{*)}$ with a negative right-hand-side constant. If more than one value is negative then select

$$N^* = \min\{x_j \mid x_j < 0 \;\; j \in B_i\}.$$

(C3) (*Determining the entering variable.*) Let K$_i$ = (I-H$_i$), where $I \in R^{n,n}$ unit matrix. Find the index B* where

$$B^* = \min\{\tfrac{-c^T H_i^T e_k}{e_k^T K_i^T e_{N^*}} \mid k \in N_i \; and \, e_k^T K_i^T e_{N^*} < 0\},$$ where e_{N^*} and e_k are unit vector.

(D3) (*Change the basis.*)

Update the solution

$$x_{i+1} = x_i - \alpha_i p_i \text{ , where } \alpha_i = \frac{-x_{B^*}}{e^T_{B^*} H^T_i e_{N^*}} \text{ and } p_i = H^T_i e_{N^*}$$

Update the projection matrix

$$H_{i+1} = H_i - \frac{(H_i e_{B^*} - e_{B^*}) e^T_{N^*} H_i}{e^T_{N^*} H e_{B^*}}$$

(E3) (Feasibility test)

If all entries, **x_{i+1}**>0 are nonnegative the solution is primal feasible, so STOP **x_{i+1}**,is the optimal solution.

If **x_{i+1}** vector has negative element, then the **x_{i+1}** solution is not optimal, **H_i=H_{i+1}**, **x_i=x_{i+1}**and go to step B3.

Remark 2.7 In step B3, there are several strategies for choosing the index of the leaving variable (**N^***). We select the most negative one, but selecting the first negative element has also been proposed.

3 ABS Algorithm for Solving Integer and Mixed Integer Problems

Here we will present our ABS-based algorithm to solve integer and mixed integer LP problems. Our basic idea is to apply Gomory cutting plane methods to add a linear constraint to exclude any non-integer optimal solutions. Let an MILP problem be formulated as in (2). The method proceeds by first dropping the requirement that certain **x_i** be integer and solving the associated linear programming problem. If the solution found does not satisfy to the integer condition, then we add constraints (cuts) to the already solved LP. While such constraints can make the primal solution infeasible, they do not affect feasibility of the dual solution. We can therefore simply add the constraint and continue running the dual LP algorithm from the current solution until the primal solution again becomes feasible. The process is repeated until an integer solution is found. Cut or condition generation is a crucial step in the method. Many different strategies are known to construct the condition [3]. We implemented the pure Gomory cut to illustrate the running principle of our algorithm.

$$s_1 + \sum_{j \notin B} (\lfloor \bar{a}_{i,j} \rfloor - \bar{a}_{i,j}) x_j = \lfloor \bar{b}_i \rfloor - \bar{b}_i ,$$

where **$s_1 \geq 0$** is a new slack variable, and $\bar{a}_{i,j}$ denotes entry of the optimal tableau in the **ith** row and the **jth** column.

Algorithm 4 ***Solving Integer and Mixed Integer LP.***

(A4) Initialization:

Rephrase the mixed or integer LP that we drop the integrity restriction for the variable.

(B4) Find an initial feasible solution for the new problem. (use Algorithm-1)

(C4) Solve the LP relaxation LP_0 problem (use Algorithm-2)

If the relaxation does not have optimal solution, then STOP. Denote $\mathbf{x}^*$ is an optimal vertex.

(D4) If $\mathbf{x}^*$ is integer then STOP, otherwise

Choose the first (i.e., highest) row **ith** where the optimal solution ($\mathbf{x}^*$) is not integral. (Note that this includes the **zeroth** row.)

Add the cut to the bottom of the optimal tableau, and add n+1 to the basis B.

(E4) Use the dual simplex algorithm starting with the previous optimal tableau extended by the Gomory cut to find the lexicographically largest feasible solution of the relaxation (use Algorithm-3)

(F4) Set i:= i+ 1. Go to D4

Remark 3.1 A cut is never based on a previous cut, so i≠n+1 in step D4.

Remark 3.2 The current algorithm adds just one new line to the system in every step, and the new cut uses previous ones. Cuts can also be rewritten.

Remark 3.3 The Gomory Cutting Plane Algorithm terminates in a finite number of steps. The proof strongly utilizes the fact that we choose the first row for the new cut in step D4 [1, 2].

Remark 3.4 Note that the form the implicit LX algorithm follows he special structure of the projection matrix. Therefore the number of the non-zero rows remains fix that is it does not increase with the new cuttings.

4 Numerical Results

To illustrate the numerical feature of our algorithm, we implemented it in MATLAB R2010a on a personal computer running Microsoft Windows 7. In all cases our algorithm found the optimal solution.

Below we show an example to illustrate how our algorithm finds the solution. Consider the following integer problem [12] to

$$z = (7x_1 + 4x_2 + 3x_3 + 2x_4) \rightarrow \max$$

subject to

$$\begin{aligned} 2x_1 + x_2 - x_3 \quad &\le 10 \\ x_1 + x_2 + 2x_3 + x_4 &\le 12 \\ x_1 \quad + 3x_3 + 2x_4 &\le 14,\ x_1, x_2, x_{3,} x_4 \in Z^+ \end{aligned} \qquad (6)$$

Define the following LP problem

$$\min_x c^T x$$

subject to $Ax \le b$, $x \ge 0$, where

$$A = \begin{bmatrix} \mathbf{2} & \mathbf{1} & \mathbf{-1} & \mathbf{0} \\ \mathbf{1} & \mathbf{1} & \mathbf{2} & \mathbf{1} \\ \mathbf{1} & \mathbf{0} & \mathbf{3} & \mathbf{2} \end{bmatrix},\ b = \begin{bmatrix} \mathbf{10} \\ \mathbf{12} \\ \mathbf{14} \end{bmatrix},\ \mathbf{c}^{\mathbf{T}} = [\mathbf{-7} \quad \mathbf{-4} \quad \mathbf{-3} \quad \mathbf{-2}]$$

Introduce the $\mathbf{u_1}$, $\mathbf{u_2}$, $\mathbf{u_3}$ non-negative slack variable to obtain the following standard LP problem

$$\tilde{A} = \begin{bmatrix} \mathbf{2} & \mathbf{1} & \mathbf{-1} & \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{0} \\ \mathbf{1} & \mathbf{1} & \mathbf{2} & \mathbf{1} & \mathbf{0} & \mathbf{1} & \mathbf{0} \\ \mathbf{1} & \mathbf{0} & \mathbf{3} & \mathbf{2} & \mathbf{0} & \mathbf{0} & \mathbf{1} \end{bmatrix},\ \tilde{x} = \begin{bmatrix} \mathbf{x_1} \\ \mathbf{x_2} \\ \mathbf{x_3} \\ \mathbf{x_4} \\ \mathbf{u_1} \\ \mathbf{u_2} \\ \mathbf{u_3} \end{bmatrix},$$

$$d^T = [\mathbf{-7} \quad \mathbf{-4} \quad \mathbf{-3} \quad \mathbf{-2} \quad \mathbf{0} \quad \mathbf{0} \quad \mathbf{0}]$$

We apply our Algorithm-1(***Finding an initial feasible solution - Phase 1)*** in three steps using $\mathbf{e_5}$, $\mathbf{e_6}$, $\mathbf{e_7}$ unit vectors respectively. We obtain the following projection matrix

$$H_3^T = \begin{bmatrix} \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{-2} & \mathbf{-1} & \mathbf{-1} \\ \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{-1} & \mathbf{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{1} & \mathbf{-2} & \mathbf{-3} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{-1} & \mathbf{-2} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix},$$

and an initial feasible solution

$$x_3 = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{10} \\ \mathbf{12} \\ \mathbf{14} \end{bmatrix}.$$

We can notice that the **H₃** matrix is Hermitian and the indexes of basis **B₃={5, 6, 7}**. Note that the number of the non-zero rows is the number of the elements of the bases. As the cost vector $c^T H_3 = [\mathbf{-7} \quad \mathbf{-4} \quad \mathbf{-3} \quad \mathbf{-2} \quad \mathbf{0} \quad \mathbf{0} \quad \mathbf{0}]$ has negative elements, our feasible solution is not optimal.

We apply the ABS based simplex algorithm (***Finding the optimal solution***). After four steps we obtain the optimal solution

$$x_7 = \begin{bmatrix} \mathbf{6} \\ \mathbf{\frac{2}{3}} \\ \mathbf{\frac{8}{3}} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix},$$

and our projection matrix is

$$H_7 = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ -\frac{\mathbf{1}}{\mathbf{2}} & \frac{\mathbf{1}}{\mathbf{2}} & -\frac{\mathbf{1}}{\mathbf{2}} & \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ -\frac{\mathbf{1}}{\mathbf{2}} & \frac{\mathbf{3}}{\mathbf{18}} & \frac{\mathbf{3}}{\mathbf{18}} & \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{0} \\ \frac{\mathbf{1}}{\mathbf{2}} & -\frac{\mathbf{21}}{\mathbf{18}} & -\frac{\mathbf{3}}{\mathbf{18}} & \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{0} \\ -\frac{\mathbf{1}}{\mathbf{2}} & \frac{\mathbf{5}}{\mathbf{6}} & -\frac{\mathbf{3}}{\mathbf{18}} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1} \end{bmatrix}$$

The structure of the **H$_7$** matrix clearly shows that the indexes of the basis are **B$_7$={1, 2, 3}**. As the obtained solution is not integer we need to introduce a new slack variable s_1 and add a new equation (constraint)

$$-\frac{1}{2}x_4 - \frac{15}{18}x_5 - \frac{1}{6}x_6 - \frac{1}{6}x_7 + s_1 = -\frac{2}{3} \qquad (\mathbf{c_1})$$

The basic feature of the ABS methods is that by adding a new equation to our system the algorithm finds a solution lying at the intersection of the linear varieties of the solutions of the original and the new equation within one step. We add a new line for our projection matrix

$$H_{\tilde{7}} = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ -\frac{\mathbf{1}}{\mathbf{2}} & \frac{\mathbf{1}}{\mathbf{2}} & -\frac{\mathbf{1}}{\mathbf{2}} & \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ -\frac{\mathbf{1}}{\mathbf{2}} & \frac{\mathbf{3}}{\mathbf{18}} & \frac{\mathbf{3}}{\mathbf{18}} & \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \frac{\mathbf{1}}{\mathbf{2}} & -\frac{\mathbf{21}}{\mathbf{18}} & -\frac{\mathbf{3}}{\mathbf{18}} & \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{0} \\ -\frac{\mathbf{1}}{\mathbf{2}} & \frac{\mathbf{5}}{\mathbf{6}} & -\frac{\mathbf{3}}{\mathbf{18}} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1} \end{bmatrix},$$

and we solve the new equations.

We obtain that our solution (x_8) is infeasible.

$$x_8 = \begin{bmatrix} \mathbf{6} \\ \mathbf{\frac{2}{3}} \\ \mathbf{\frac{8}{3}} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ -\mathbf{\frac{2}{3}} \end{bmatrix}.$$

We need to use our Algorithm-3 to move to a feasible solution. The projection matrix for our dual simplex method is

$$H_8 = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ -\mathbf{\frac{1}{2}} & \mathbf{\frac{1}{2}} & -\mathbf{\frac{1}{2}} & \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ -\mathbf{\frac{1}{2}} & \mathbf{\frac{3}{18}} & \mathbf{\frac{3}{18}} & \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{\frac{1}{2}} & -\mathbf{\frac{21}{18}} & -\mathbf{\frac{3}{18}} & \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{0} \\ -\mathbf{\frac{1}{2}} & \mathbf{\frac{5}{6}} & -\mathbf{\frac{3}{18}} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1} \end{bmatrix}$$

After two steps we obtain the optimal solution for the modified system.

$$x_{c_1} = \begin{bmatrix} \mathbf{\frac{16}{3}} \\ \mathbf{\frac{4}{3}} \\ \mathbf{2} \\ \mathbf{\frac{2}{3}} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix},$$

and our projection matrix is

$$H_{c_1} = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \frac{\mathbf{1}}{\mathbf{3}} & -\frac{\mathbf{2}}{\mathbf{3}} & \mathbf{1} & -\frac{\mathbf{4}}{\mathbf{3}} & \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \frac{\mathbf{2}}{\mathbf{3}} & -\frac{\mathbf{4}}{\mathbf{3}} & \mathbf{0} & -\frac{\mathbf{1}}{\mathbf{3}} & \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{0} \\ -\frac{\mathbf{1}}{\mathbf{3}} & \frac{\mathbf{2}}{\mathbf{3}} & \mathbf{0} & -\frac{\mathbf{1}}{\mathbf{3}} & \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{0} \\ -\mathbf{1} & \mathbf{1} & -\mathbf{1} & \mathbf{2} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1} \end{bmatrix}.$$

A new fractional solution has been found, we need to generate a new constraint, which is valid for the integer solution, but not for our current solution. The new cutting plane is

$$-\frac{2}{3}x_5 - \frac{1}{3}x_6 - \frac{1}{3}x_7 + s_2 = -\frac{1}{3} \qquad \textbf{(c}_2\textbf{)}$$

After solving the new equations, and using Algorithm-3 for re-optimizing the solution, we obtain a new solution.

$$x_{c_2} = \begin{bmatrix} \frac{\mathbf{11}}{\mathbf{2}} \\ \mathbf{1} \\ \frac{\mathbf{5}}{\mathbf{2}} \\ \frac{\mathbf{1}}{\mathbf{2}} \\ \frac{\mathbf{1}}{\mathbf{2}} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix},$$

and projection matrix is

$$H_{c_2}=\begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \frac{\mathbf{1}}{\mathbf{2}} & -\mathbf{1} & -\frac{\mathbf{1}}{\mathbf{2}} & \frac{\mathbf{1}}{\mathbf{2}} & -\frac{\mathbf{1}}{\mathbf{2}} & \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ -\frac{\mathbf{1}}{\mathbf{2}} & \mathbf{1} & -\frac{\mathbf{1}}{\mathbf{2}} & \frac{\mathbf{1}}{\mathbf{2}} & -\frac{\mathbf{1}}{\mathbf{2}} & \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{0} \\ \mathbf{1} & \mathbf{1} & -\mathbf{1} & \mathbf{2} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{0} \\ \frac{\mathbf{1}}{\mathbf{2}} & -\mathbf{1} & \frac{\mathbf{3}}{\mathbf{2}} & -\frac{\mathbf{5}}{\mathbf{2}} & \frac{\mathbf{5}}{\mathbf{2}} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1} \end{bmatrix}$$

The found solution is not integer, therefore we add the constraint

$$-\frac{1}{2}x_6-\frac{1}{2}x_7-\frac{1}{2}x_9+s_3=-\frac{1}{2} \qquad \textbf{(c}_3\textbf{)}$$

After solving the new equations, and using Algorithm-3, we obtain the solution

$$x_{C_3}=\begin{bmatrix} \mathbf{5} \\ \mathbf{2} \\ \mathbf{2} \\ \mathbf{1} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{1} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix},$$

and projection matrix is

$$H_{c_3} = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{1} & -\mathbf{2} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1} & -\mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ -\mathbf{1} & \mathbf{1} & -\mathbf{1} & \mathbf{2} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{0} \\ \mathbf{1} & -\mathbf{2} & \mathbf{2} & -\mathbf{3} & \mathbf{2} & \mathbf{0} & -\mathbf{1} & \mathbf{0} & \mathbf{1} & \mathbf{0} \\ -\mathbf{1} & \mathbf{2} & -\mathbf{1} & \mathbf{1} & -\mathbf{1} & \mathbf{0} & \mathbf{2} & \mathbf{0} & \mathbf{0} & \mathbf{1} \end{bmatrix}.$$

Our optimal solution is integer, therefore we found the solution for our problem (6) because the x_{C_3} is primal feasible and every components is integer.

Discussion and Conclusions

In this paper we showed that the Gomory's original cutting-plane approach can be embedded in the ABS class of algorithms. Furthermore, we implemented our new algorithm in MATLAB and an example was given to demonstrate the correctness of our method.

Though Zou and Xia described an ABS-based algorithm for solving integer LP problems [13], the published method worked only on a special case when the **A** matrix is unimodular, i.e. the determinant of **A** is 1. The constraints on the **A** matrix and the inability of that algorithm to deal with mixed integer problems strongly limited the spectrum of problems that the algorithm was able to solve.

A cucial step of our algorithm is the generation of Gomory's cuts. In the current version of our algorithm, we used Gomory's original cuts defined in the LP optimal tableau. Since a number of additional cutting strategies have also been published, we also intend to extend our algorithm t those other types of cuts. We are planning to compare them and analyze the numerical feature of them, emphasizing the possibilities of the parallelization as the ABS algorithms are suitable for parallelization.

We should mention that every cut adds a new slack variable to the system, which means that the number of columns of the projection matrix increases by one in every steps (the number of rows remains). Some results were published to avoid this problem [14] and we are planning to implement them, too.

A number of papers were published showing that ABS-based algorithms are suitable for solving integer LP and Diophantine linear systems of equations too

[15-19]. Therefore some further work should be considered including investigations of the implementation of the pure integer algorithm using Gomory's cuts [14, 20].

Acknowledgement

We thank Ferenc Forgó for helpful discussions and encouraging notes.

References

[1] R. E. Gomory, "Outline of an Algorithm for Integer Solutions to Linear Programs," Bulletin of the American Mathematical Society, Vol. 64, pp. 275-278, 1958

[2] R. Gomory, "An Algorithm for the Mixed Integer Problem," DTIC Document 1960

[3] E. Balas, S. Ceria, G. Cornuéjols, and N. Natraj, "Gomory Cuts Revisited," Operations Research Letters, Vol. 19, pp. 1-9, 1996

[4] A. Zanette, M. Fischetti, and E. Balas, "Can Pure Cutting Plane Algorithms Work?," in Integer Programming and Combinatorial Optimization, ed: Springer, 2008, pp. 416-434

[5] J. Abaffy, "A lineáris egyenletrendszerek általános megoldásának egy direkt módszerosztálya. ," Alkalmazott Matematikai Lapok, Vol. 5, 1979

[6] J. Abaffy, C. Broyden, and E. Spedicato, "A Class of Direct Methods for Linear-Systems," Numerische Mathematik, Vol. 45, pp. 361-376, 1984

[7] J. Abaffy and E. Spedicato, ABS Projection Algorithms: Mathematical Techniques for Linear and Nonlinear Equations. Chichester, West Sussex, England: Ellis Horwood, 1989

[8] J. Abaffy and E. Spedicato, "A Class of Scaled Direct Methods for Linear-Systems," Annals of the Institute of Statistical Mathematics, Vol. 42, pp. 187-201, Mar 1990

[9] E. Spedicato, Z. Xia, and L. Zhang, "The Implicit LX Method of the ABS Class," Optimization Methods and Software, Vol. 8, pp. 99-110, 1997

[10] E. Spedicato, Z. Q. Xia, and L. W. Zhang, "ABS Algorithms for Linear Equations and Optimization," Journal of Computational and Applied Mathematics, Vol. 124, pp. 155-170, Dec 1 2000

[11] J. Abaffy, L. X-J, and X. Z-Q, "A Modified Non-Simplex Active Set Method for the Standard LP Problem," Pure Mathematics and Applications, Vol. 23, pp. 1-11, 2012

[12] L. Gáspár, "Operációkutatás II.," ed. Budapest: Tankönyvkiadó, 1977

[13] M.-F. Zou and Z.-Q. Xia, "ABS Algorithms for Diophantine Linear Equations and Integer LP Problems," Journal of Applied Mathematics and Computing, Vol. 17, pp. 93-107, 2005

[14] B. Vizvári, Operációkutatási modellek: Typotex Kft, 2009

[15] H. Esmaeili, N. Mahdavi-Amiri, and E. Spedicato, "ABS Solution of a Class of Linear Integer Inequalities and Integer LP Problems," Optimization Methods & Software, Vol. 16, pp. 179-192, 2001

[16] H. Esmaeili, N. Mahdavi-Amiri, and E. Spedicato, "A Class of ABS Algorithms for Diophantine Linear Systems," Numerische Mathematik, Vol. 90, pp. 101-115, Nov 2001

[17] S. Fodor, "Symmetric and Non-Symmetric ABS Methods for Solving Diophantine Systems of Equations," Annals of Operations Research, Vol. 103, pp. 291-314, 2001

[18] S. Fodor, "ABS Class of Methods for Diophantine Systems of Equations Part I.," Quaderni del Dipartimento di Matematica, Statistica, Informatica ed Applicazioni, Universita degli Studi di Bergamo, Vol. Report No. 2000/15, pp. 1-19, 2000

[19] S. Fodor, "ABS Class of Methods for Diophantine Systems of Equations Part II.," Quaderni del Dipartimento di Matematica, Statistica, Informatica ed Applicazioni, Universita degli Studi di Bergamo, Vol. Report No. 2000/16, pp. 1-16, 2000

[20] F. Forgó, Nonconvex programming: Akadémiai Kiadó Budapest, 1988

Numerical Simulation of Anisotropic Mean Curvature of Graphs in Relative Geometry

Dieu Hung Hoang, Michal Beneš, Tomáš Oberhuber

Department of Mathematics
Faculty of Nuclear Sciences and Physical Engineering
Czech Technical University in Prague
Trojanova 13, 12000 Praha, Czech Republic
hoangdieu@fjfi.cvut.cz, michal.benes@fjfi.cvut.cz, tomas.oberhuber@fjfi.cvut.cz

Abstract: The aim of this paper is the numerical simulation of anisotropic mean curvature of graphs in the context of relative geometry, developed in [1]. We extend results in [4] to our problem; we prove an existence theorem and energy equality. The numerical scheme is based on the method of lines where the spatial derivatives are approximated by finite differences [2]. We then solve the resulting ODE system by means of the adaptive Runge-Kutta-Merson method. To show the stability of the scheme we prove the discrete version of the energy equality. Finally, we show experimental order of convergence and results of numerical experiments with various anisotropy settings.

Keywords: anisotropy; mean curvature; Finsler geometry; method of lines; FDM

1 Introduction

The paper studies the following motion law for surfaces in $\mathbb{R}^3$ denoted by Γ:

$$velocity = curvature + forcing \tag{1}$$

in a certain sense which is specified below. Both the velocity and the curvature are evaluated with respect to the direction given by a vector locally influenced by the orientation of the Euclidean normal vector to Γ.

One example of the law (1) is represented by the isotropic mean-curvature flow given by the equation

$$v_\Gamma = -\kappa_\Gamma + f \text{ on}\Gamma(t) \tag{2}$$

in the direction of $n_\Gamma$which is the Euclidean normal vector to Γ, while v_Γ the normal velocity, κ_Γ the mean curvature, and f the forcing term. The equation (2) in the form of the Gibbs-Thompson law is contained in the modified Stefan problem. For details, we refer the reader to [9, 16].

One of few anisotropic examples where the analytical solution is known considers a ball under the relative geometry which shrinks according to (1) with$f = 0$. In this case we have the initial ball with radius r_0, normal velocity $\dot{r}$, actual curvature along the ball of radius r being $\frac{1}{r}$. The equation (1) reads

$$\dot{r} = -\frac{1}{r},$$

and has the solution

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

This law has been intensively studied, see e.g. [4, 5, 13].

This paper deals with the motion by anisotropic mean curvature in relative geometry associated with the Finsler metric, developed in [1], which reads

$$v_{\Gamma,\Phi} = -\kappa_{\Gamma,\Phi} + f \text{ on } \Gamma(t) \tag{3}$$

Here, $v_{\Gamma,\phi}$ denotes the normal velocity, $\kappa_{\Gamma,\phi}$ is the anisotropic mean curvature of $\Gamma(t)$ with respect to the Finsler metric ϕ, and f is the forcing term.

Deckelnick and Dziuk proved the convergence and gave the optimal error estimates using finite element method for graph [4, 7] and parametric case [8]. Haußer and Voigt [11] presented a parametric finite element approximation for a regularized version. Pozzi studied the anisotropic mean curvature flow in higher codimension in [15].

2 Anisotropy in Relative Geometry

In what follows we shall first define anisotropy by means of the Finsler geometry; then, we shall transform the motion law (3) into graph formulation. For this purpose, we assume that there is a smooth function with non-vanishing gradient $p: \mathbb{R}^{2+1} \to \mathbb{R}$ such that

$$\Gamma(t) = \{[x, y] \in \mathbb{R}^3 | y = p(t, x), x \in \Omega \subset \mathbb{R}^2\}.$$

We say that a continuous function $\phi: \mathbb{R}^3 \to \mathbb{R}_0^+$ is a Finsler metric if it satisfies the properties

1. $\phi \in C^{3+\alpha}(\mathbb{R} - \{0\})$,
2. $\phi^2$is strictly convex,
3. $\phi(t\eta) = |t|\phi(\eta),\ \ t \in \mathbb{R},\ \ \eta \in \mathbb{R}^3$,
4. $\lambda|\eta| \leq \phi(\eta) \leq \Lambda|\eta|,\ \ \eta \in \mathbb{R}^3$,
   for two suitable constants $0 < \lambda \leq \Lambda < \infty$.

Associated to ϕ we define the unit ball (also so-called Wulff shape)

$B_{\phi} = \{\eta \in \mathbb{R}^3 | \phi(\eta) \leq 1\}$.

One can prove that a dual function $\phi^0 : \mathbb{R}^3 \to \mathbb{R}_0^+$ given by

$\phi^0(\eta^*) = \sup\{\eta^* \cdot \eta | \eta \in B_{\phi}\}$

is also a Finsler metric.

For simplicity we use η instead of η^*. Then the following relations hold [3]

$$\phi_{\eta}^0(t\eta) = \frac{t}{|t|}\phi_{\eta}^0(\eta), \ \ \phi_{\eta\eta}^0(t\eta) = \frac{1}{|t|}\phi_{\eta\eta}^0(\eta), \ \ t \in \mathbb{R} - \{0\},$$

$$\zeta \cdot \phi^0(\eta)\phi_{\eta\eta}^0(\eta) \cdot \zeta \geq \gamma_0 \left| \zeta - \frac{\zeta \cdot \eta}{|\eta|^2}\eta \right|^2, \ \ \eta \neq 0, \ \ \zeta \in \mathbb{R}^3, \ \ \gamma_0 > 0,$$

where the index η means the derivative with respect to η.

We define the map $T^0 : \mathbb{R}^3 \to \mathbb{R}^3$as

$$T^0(\eta) = \left(\tilde{T}^0(\eta), T_3^0(\eta)\right) = \phi^0(\eta)\phi_{\eta}^0(\eta), \ \ \eta \neq 0,$$

$$T^0(0) = 0.$$

Then, the ϕ-normal vector, ϕ-mean curvature, and ϕ-normal velocity of Γ are defined as

$$n_{\Gamma,\phi} = \frac{T^0(\nabla p, -1)}{\phi^0(\nabla p, -1)} = \phi_{\eta}^0(\nabla p, -1), \tag{4}$$

$$\kappa_{\Gamma,\phi} = \operatorname{div} n_{\Gamma,\phi} = \nabla \cdot \frac{\tilde{T}^0(\nabla p, -1)}{\phi^0(\nabla p, -1)}, \tag{5}$$

$$v_{\Gamma,\phi} = -\frac{\partial_t p}{\phi^0(\nabla p, -1)}. \tag{6}$$

By substituting the quantities (4)-(6) into the Eq. (3), we obtain the non-linear parabolic partial differential equation

$$\partial_t p = \phi^0(\nabla p, -1)\left(\nabla \cdot \left(\frac{\tilde{T}^0(\nabla p, -1)}{\phi^0(\nabla p, -1)}\right) + f\right) \text{on} \Omega \times (0, T). \tag{7}$$

The initial and boundary conditions are given by

$$p|_{t=0} = p_0 \text{on } \overline{\Omega}, \tag{8}$$

$$p = p_0 \text{ on } \partial\Omega \times (0, T). \tag{9}$$

In our numerical experiments we use the Finsler metrics listed below. We denote $n = \frac{\eta}{|\eta|}$. The corresponding Wulff shapes are illustrated in Fig. 1.

The 4-fold anisotropy reads as

$$\phi(\eta) = |\eta|\left(1 - A_1\left(1 - (n_1^4 + n_2^4 + n_3^4)\right)\right). \tag{10}$$

The 6-fold anisotropy reads as

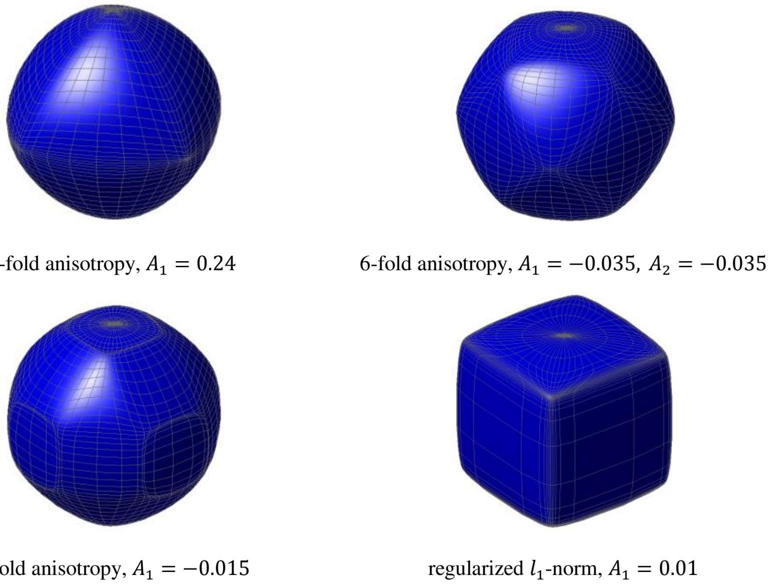
$$\begin{aligned}\phi_\eta^0(\eta) = |\eta|\Bigg(1 - A_1\left(n_1^4 + n_2^4 + n_3^4 - \frac{3}{5}\right) \\ + A_2\left(3(n_1^4 + n_2^4 + n_3^4) + 66 n_1^2 n_2^2 n_3^2 - \frac{17}{7}\right)\Bigg)\end{aligned} \tag{11}$$

The 8-fold anisotropy reads as

$$\begin{aligned}\phi_\eta^0(\eta) = |\eta|(1 - A_1(n_1^8 + n_2^8 + n_3^8 \\ - 28(n_1^6 n_2^2 + n_1^2 n_2^6 + n_2^6 n_3^2 + n_2^2 n_3^6 + n_3^6 n_1^2 + n_3^2 n_1^6) \\ + 70(n_1^4 n_2^4 + n_2^4 n_3^4 + n_3^4 n_1^4))).\end{aligned} \tag{12}$$

The regularized l_1-anisotropy reads as

$$\phi_\eta^0(\eta) = \sum_{i=1}^{3}\left(\eta_i^2 + A_1 \sum_{j=1}^{3} \eta_j^2\right)^{\frac{1}{2}}. \tag{13}$$



4-fold anisotropy, $A_1 = 0.24$

6-fold anisotropy, $A_1 = -0.035$, $A_2 = -0.035$

8-fold anisotropy, $A_1 = -0.015$

regularized l_1-norm, $A_1 = 0.01$

Figure 1

Wulff shapes for various anisotropies

3 Analytical Properties

In the following section we shall introduce some analytical results for law (7) in the context of relative geometry, which are due to [4, 11, 12]. We shall prove the energy equality and give the existence result for our problem.

Theorem 1. *For the solution of problem (7)-(9), one has the energy equality*

$$\int_\Omega p_t\left(\frac{p_t}{\phi^0(\nabla p,-1)}-f\right)+\frac{d}{dt}\int_\Omega \phi^0(\nabla p,-1)=0.$$

If $f=0$, then

$$\int_\Omega \frac{p_t^2}{\phi^0(\nabla p,-1)}+\frac{d}{dt}\int_\Omega \phi^0(\nabla p,-1)=0. \tag{14}$$

Proof. Since $p_t=0$on $\partial\Omega$, the proof is straightforward

$$\begin{aligned}\frac{d}{dt}\int_\Omega \phi^0(\nabla p,-1) &= \int_\Omega \phi_\eta^0(\nabla p,-1)\cdot[\nabla p,-1]_t=\int_\Omega [\nabla p_t,0]\cdot\frac{T^0(\nabla p,-1)}{\phi^0(\nabla p,-1)}\\ &=\int_\Omega \nabla p_t\cdot\frac{\tilde{T}^0(\nabla p,-1)}{\phi^0(\nabla p,-1)}=\int_\Omega p_t\nabla\cdot\frac{\tilde{T}^0(\nabla p,-1)}{\phi^0(\nabla p,-1)}\\ &=-\int_\Omega p_t\left(\frac{p_t}{\phi^0(\nabla p,-1)}-f\right).\end{aligned}$$

If $f=0$, we obtain the equality (14).

Lemma 1. *Let* $v(\tilde{\eta})=\phi^0(\tilde{\eta},-1)$, $a_i=\frac{T_i^0(\nabla p,-1)}{\phi^0(\nabla p,-1)}$, *and* $a_{ij}=\frac{\partial a_i}{\partial \eta_j}$. *Then for the solution of the problem (7)-(9) with* $f=0$ *one has the identity*

$$\begin{aligned}v_t &= \sum_{i,j=1}^{2}\frac{\partial}{\partial x_i}\left(a_{ij}\frac{\partial v}{\partial x_j}\right)v+\sum_{l=1}^{2}\kappa_\phi a_l\frac{\partial v}{\partial x_l}\\ &+\sum_{i,j,k=1}^{2}a_{ij}a_{kl}\frac{\partial^2 p}{\partial x_i\partial x_k}\frac{\partial^2 p}{\partial x_j\partial x_l}v.\end{aligned} \tag{15}$$

Proof. We have

$$v_t=\phi_{\tilde{\eta}}^0(\nabla p,-1)\nabla p_t=\sum_{l=1}^{2}a_l\frac{\partial(v\kappa_\phi)}{\partial x_l}=v\sum_{l=1}^{2}a_l\frac{\partial\kappa_\phi}{\partial x_l}+\kappa_\phi\sum_{l=1}^{2}a_l\frac{\partial v}{\partial x_l}.$$

Let now compute

$$\sum_{l=1}^{2} a_l \frac{\partial \kappa_\phi}{\partial x_l} = \sum_{l=1}^{2} a_l \frac{\partial}{\partial x_l} \sum_{i=1}^{2} \frac{\partial}{\partial x_i} a_i = \sum_{i,l=1}^{2} a_l \frac{\partial}{\partial x_i} \frac{\partial}{\partial x_l} a_i$$
$$= \sum_{i,l=1}^{2} \frac{\partial}{\partial x_i} \left(a_l \frac{\partial}{\partial x_l} a_i \right) - \sum_{i,l=1}^{2} \frac{\partial}{\partial x_i} a_l \frac{\partial}{\partial x_l} a_i.$$

Since

$$\frac{\partial}{\partial x_j} v = \sum_{k=1}^{2} a_k \frac{\partial}{\partial x_j} \frac{\partial}{\partial x_k} p, \quad \frac{\partial}{\partial x_l} a_i = \sum_{j=1}^{2} a_{ij} \frac{\partial}{\partial x_j} \frac{\partial}{\partial x_l} p,$$

we get the identity (15).

Theorem 2. *Let* $\partial\Omega \in C^{3+\alpha}$*and* $d(x) := dist(x, \partial\Omega)$*. We assume that*

$$\sum_{i,j=1}^{2} \phi^0_{\eta_i\eta_j}(\nabla d(x), 0) d_{x_i x_j}(x) \leq 0\,, \quad x \in \partial\Omega.$$

Let $p_0 \in C^{3+\alpha}(\overline{\Omega})$ *satisfies the compatibility condition*

$$\sum_{i,j=1}^{2} \phi^0_{\eta_i\eta_j}(\nabla p_0, -1) p_{0,x_i x_j} = 0, \quad x \in \partial\Omega.$$

Then (7)-(9) with $f = 0$ *has a solution* $p \in H^{3+\alpha,\frac{3+\alpha}{2}}(\overline{\Omega} \times [0,T])$ *with* $p_t \in L^2(0,T; H^{2,2}(\Omega))$ *for all* $T < \infty$*.*

Proof. Similarly as in [4] we are looking for a solution of the initial boundary value problem

$$p_t - \sum_{i,j=1}^{2} a_{ij}(\nabla p) p_{x_i} p_{x_j} = 0$$

but with the difference

$$a_{ij}(\tilde{\eta}) = \phi^0(\tilde{\eta}, -1) \phi^0_{\eta_i\eta_j}(\tilde{\eta}, -1).$$

Since ϕ^0 is a Finsler metric, $\phi^0 \in C^{0,1}(\mathbb{R})$ holds. Moreover, since $\phi^0 \in C^{3+\alpha}(\mathbb{R} - \{0\})$, we have $a_{ij} \in C^{0,1}(\mathbb{R})$.

Following standard lines of Theorem 4.1 in [4] and using the previous Lemma 1 we can show there is a constant K such that for every solution p^τ of

$$\begin{aligned} p_t^\tau - \tau \sum_{i,j=1}^{2} a_{ij}(\nabla p^\tau) p^\tau_{x_i x_j} - (1-\tau)\Delta p^\tau = 0 &\quad \text{in}\Omega \times (0,T), \\ p^\tau = \tau p_0 &\quad \text{on } \partial\Omega \times (0,T), \\ p^\tau(\cdot\,,0) = \tau p_0 &\quad \text{on } \Omega, \end{aligned}$$

the estimate

$$\max_{\Omega\times(0,T)} |p^\tau| + \max_{\Omega\times(0,T)} |\nabla p^\tau| \le K$$

is valid. This means (7)-(9) with $f = 0$ has a solution $p \in H^{3+\alpha,\frac{3+\alpha}{2}}(\overline{\Omega} \times [0,T])$.

4 Numerical Scheme

We employed the numerical scheme based on the method of lines. The spatial derivatives are discretized and the time variable is left continuous. After discretizing the problem by finite differences in space, we solve the resulting ODE system by the adaptive Runge-Kutta-Merson method. We consider the computational domain $\Omega = (0, L_1) \times (0, L_2)$ and introduce the following notation:

$$\begin{aligned} h_1 &= \frac{L_1}{N_1}, h_2 = \frac{L_2}{N_2} \\ \omega_h &= \{[ih_1, jh_2] | i = 1, \cdots, N_1 - 1; j = 1, \cdots, N_2 - 1\}, \\ \overline{\omega}_h &= \{[ih_1, jh_2] | i = 0, \cdots, N_1; j = 0, \cdots, N_2\}, \\ \gamma_h &= \omega_h - \overline{\omega}_h, \\ u_{ij} &= u(ih_1, jh_2), \\ u_{x_1,ij} &= \frac{u_{i+1,j} - u_{ij}}{h_1}, u_{x_2,ij} = \frac{u_{i,j+1} - u_{ij}}{h_2}, \\ u_{\bar{x}_1,ij} &= \frac{u_{ij} - u_{i-1,j}}{h_1}, u_{\bar{x}_2,ij} = \frac{u_{ij} - u_{i,j-1}}{h_2}, \\ \nabla_h u &= [u_{x_1}, u_{x_2}], \\ \overline{\nabla}_h u &= [u_{\bar{x}_1}, u_{\bar{x}_2}], \\ \mathcal{P}_h g &= g|_{\overline{\omega}_h}. \end{aligned}$$

We define the following expressions

$$(f,g)_h = \sum_{i=1}^{N_1-1}\sum_{j=1}^{N_2-1} h_1h_2f_{ij}g_{ij},\quad \|f\|_h^2 = (f,f)_h,$$
$$(f^1,g^1] = \sum_{i=1}^{N_1}\sum_{j=1}^{N_2-1} h_1h_2f_{ij}^1g_{ij}^1,$$
$$(f^2,g^2] = \sum_{i=1}^{N_1-1}\sum_{j=1}^{N_2} h_1h_2f_{ij}^2g_{ij}^2,$$
$$(\boldsymbol{f},\boldsymbol{g}] = (f^1,g^1] + (f^2,g^2],$$
$$(f,g] = \sum_{i=1}^{N_1}\sum_{j=1}^{N_2} h_1h_2f_{ij}g_{ij}.$$

We then propose a semi-discrete scheme [2]

$$\begin{aligned} p_t^h &= \phi^0(\overline{\nabla}_hp^h,-1)\left(\nabla_h\cdot\left(\frac{\tilde{T}^0(\overline{\nabla}_hp^h,-1)}{\phi^0(\overline{\nabla}_hp^h,-1)}\right)+f\right)\text{on}\omega_h\times(0,T),\\ p^h|_{t=0} &= \mathcal{P}_hp_0\text{on}\overline{\omega}_h\\ p^h &= \mathcal{P}_hp_0\text{on}\gamma_h\times(0,T). \end{aligned} \tag{16}$$

This is an ODE system and existence and uniqueness of solutions are guaranteed by the theory of ordinary differential equations (the Picard–Lindelöf theorem).

As the stability criterion we use the basic energy equality (14). For this purpose we shall now prove the discrete version of Theorem 1.

Theorem 3. For the solution of problem (16), the following energy equality holds

$$\left(p_t^h,\frac{p_t^h}{\phi^0(\overline{\nabla}_hp^h,-1)}-f\right)_h+\frac{d}{dt}(\phi^0(\overline{\nabla}_hp^h,-1),1]=0.$$

If $f=0$, then

$$\left((p_t^h)^2,\frac{1}{\phi^0(\overline{\nabla}_hp^h,-1)}\right)_h+\frac{d}{dt}(\phi^0(\overline{\nabla}_hp^h,-1),1]=0. \tag{17}$$

Proof. Applying the grid version of Green's formula as in [2], we obtain

$$\left(p_t^h,\frac{p_t^h}{\phi^0(\overline{\nabla}_hp^h,-1)}-f\right)_h=\left(p_t^h,\nabla_h\cdot\left(\frac{\tilde{T}^0(\overline{\nabla}_hp^h,-1)}{\phi^0(\overline{\nabla}_hp^h,-1)}\right)\right)_h$$
$$=\left(\overline{\nabla}_{\mathrm{h}}p_t^h,\frac{\tilde{T}^0(\overline{\nabla}_hp^h,-1)}{\phi^0(\overline{\nabla}_hp^h,-1)}\right]$$
$$=-\sum_{i=1}^{N_1}\sum_{j=1}^{N_2-1}h_1h_2p_t^h|_{\bar{x}_1,ij}\left.\frac{T_1^0(\overline{\nabla}_hp^h,-1)}{\phi^0(\overline{\nabla}_hp^h,-1)}\right|_{ij}$$

$$-\sum_{i=1}^{N_1-1}\sum_{j=1}^{N_2} h_1h_2p_t^h|_{\bar{x}_2,ij}\frac{T_2^0(\overline{\nabla}_hp^h,-1)}{\phi^0(\overline{\nabla}_hp^h,-1)}\Bigg|_{ij}$$

$$=-\sum_{i=1}^{N_1}\sum_{j=1}^{N_2} h_1h_2p_t^h|_{\bar{x}_1,ij}\frac{T_1^0(\overline{\nabla}_hp^h,-1)}{\phi^0(\overline{\nabla}_hp^h,-1)}\Bigg|_{ij}$$

$$-\sum_{i=1}^{N_1}\sum_{j=1}^{N_2} h_1h_2p_t^h|_{\bar{x}_2,ij}\frac{T_2^0(\overline{\nabla}_hp^h,-1)}{\phi^0(\overline{\nabla}_hp^h,-1)}\Bigg|_{ij} = -\frac{d}{dt}(\phi^0(\overline{\nabla}_hp^h,-1),1].$$

If $f = 0$, we get the equality (17).

5 Computational Results

We first investigate the convergence of the numerical scheme. Then, we explore the long time behaviour of the anisotropic motion law (3).

Experimental order of convergence. The computations have been performed over a range of different grid resolutions which allows quantifying the numerical convergence by the experimental order of convergence (EOC). A numerical solution computed on the finest grid is used to substitute the analytical solution. Given errors Error_1 and Error_2 for two mesh sizes h_1, h_2, respectively, the EOC is defined as

$$\text{EOC}=\frac{\log(\text{Error}_1/\text{Error}_2)}{\log(\text{h}_1/\text{h}_2)}.$$

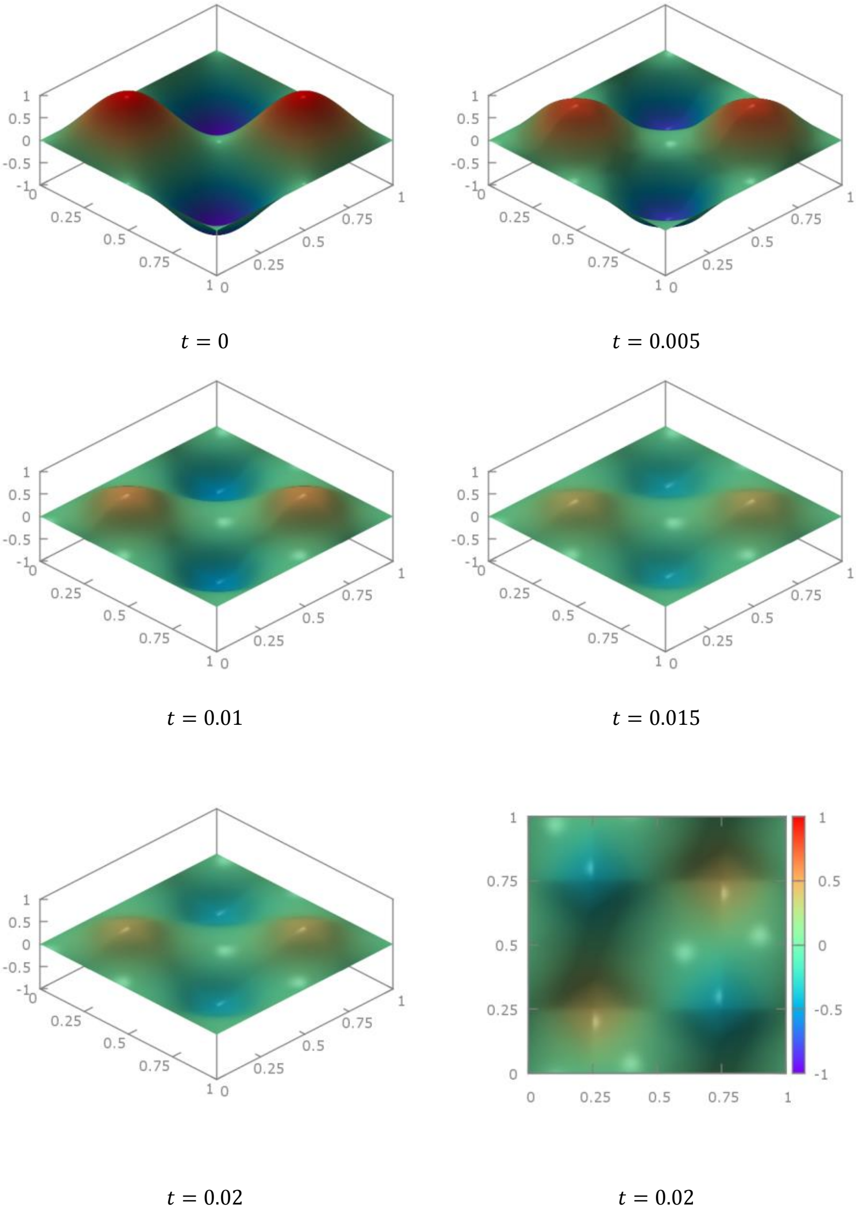
The result is shown in the following table.

Table 1

Experimental order of convergence of the scheme (16)

N	h	Error L_∞	EOC L_∞	Error L_2	EOC L_2
50	1/50	0.05924	-	0.01175	-
100	1/100	0.03676	0.68843	0.00731	1.00000
150	1/150	0.02689	0.77219	0.00511	1.00000
200	1/200	0.02058	0.92781	0.00357	1.00000

Morphology evolution. We present the solutions at different times for various anisotropies. Figs. 2-6 show surface evolutions under anisotropic mean curvature flow without the forcing term ($f = 0$). Anisotropy is shown to be crucial in the formation of different surface morphologies. The surface is first determined by symmetry of anisotropy; it then evolves towards to the flat surface. Finally, the effect of the forcing term f on the surface evolution is shown in Fig. 7.



$t = 0$

$t = 0.005$

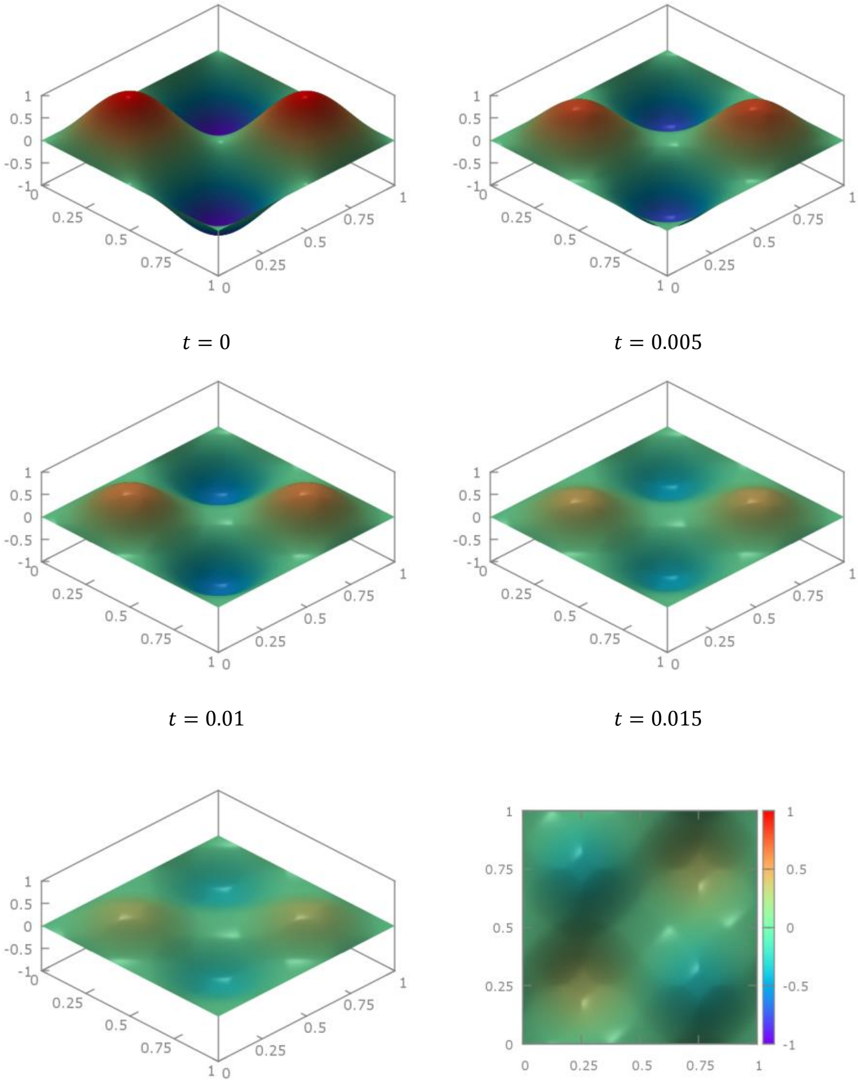
$t = 0.01$

$t = 0.015$

$t = 0.02$

$t = 0.02$

Figure 2
Morphology evolution for $f = 0$, the 4-fold anisotropy (10) with $A_1 = 0.24$,
$p_0(x_1, x_2) = \sin(2x_1)\sin(2x_2)$ at different times



$t = 0$

$t = 0.005$

$t = 0.01$

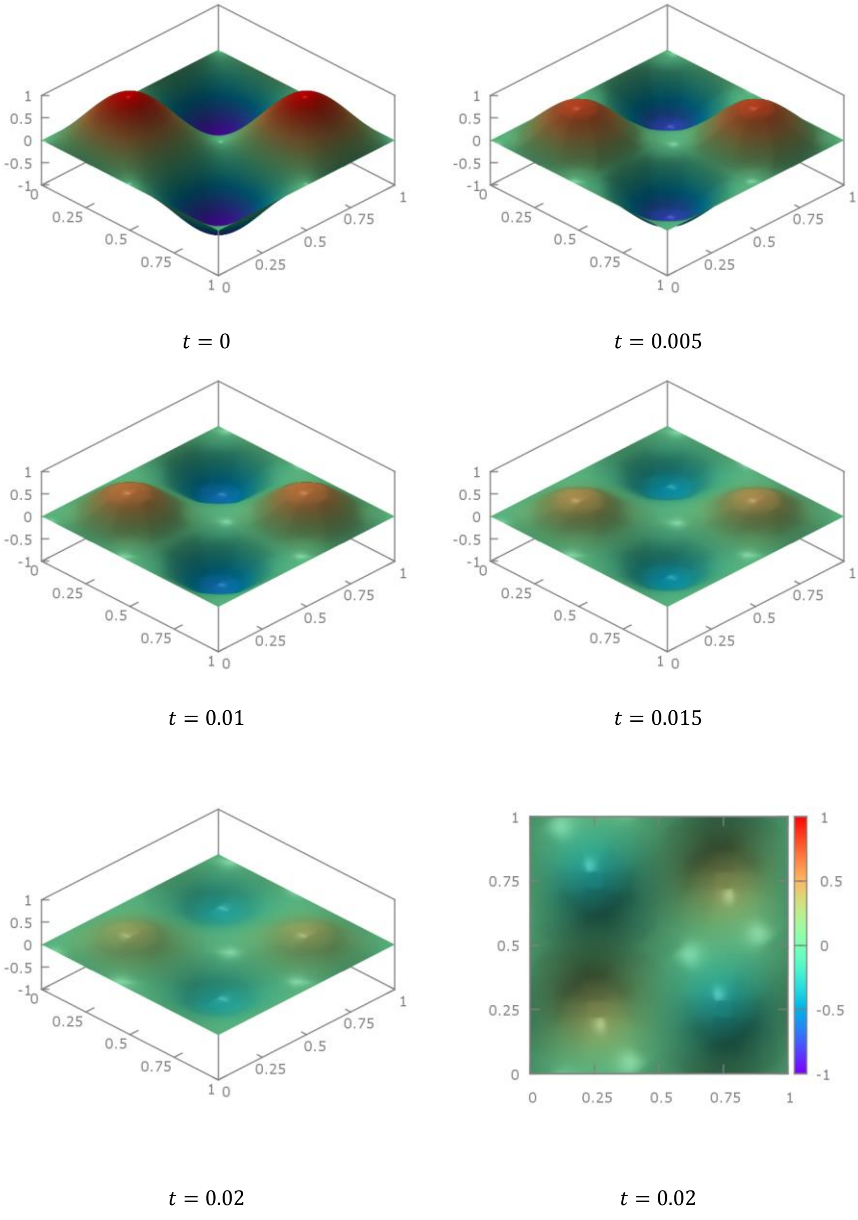
$t = 0.015$

$t = 0.02$

$t = 0.02$

Figure 3

Morphology evolution for $f = 0$, the 6-fold anisotropy (11) with $A_1 = -0.035$, $A_2 = -0.035$, $p_0(x_1, x_2) = \sin(2x_1)\sin(2x_2)$ at different times



$t = 0$

$t = 0.005$

$t = 0.01$

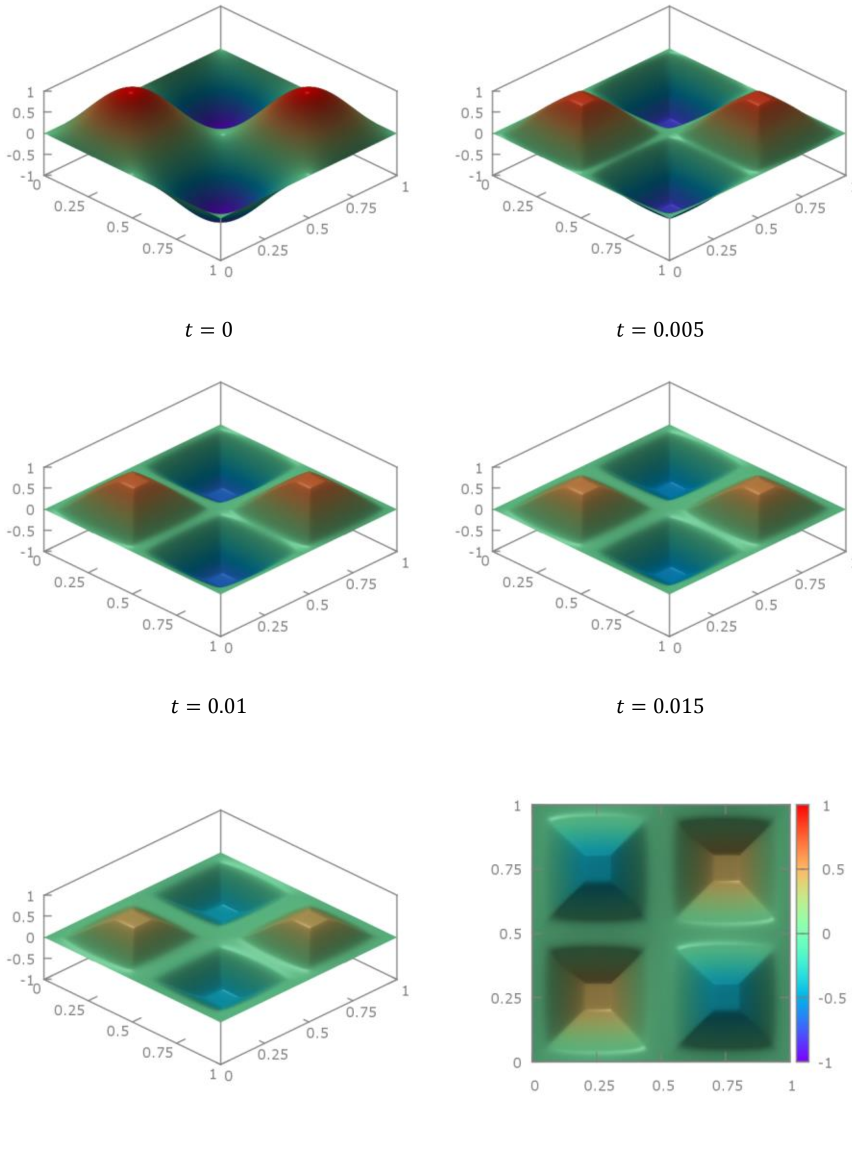
$t = 0.015$

$t = 0.02$

$t = 0.02$

Figure 4

Morphology evolution for $f = 0$, the 8-fold anisotropy (12) with $A_1 = -0.015$, $p_0(x_1, x_2) = \sin(2x_1)\sin(2x_2)$ at different times



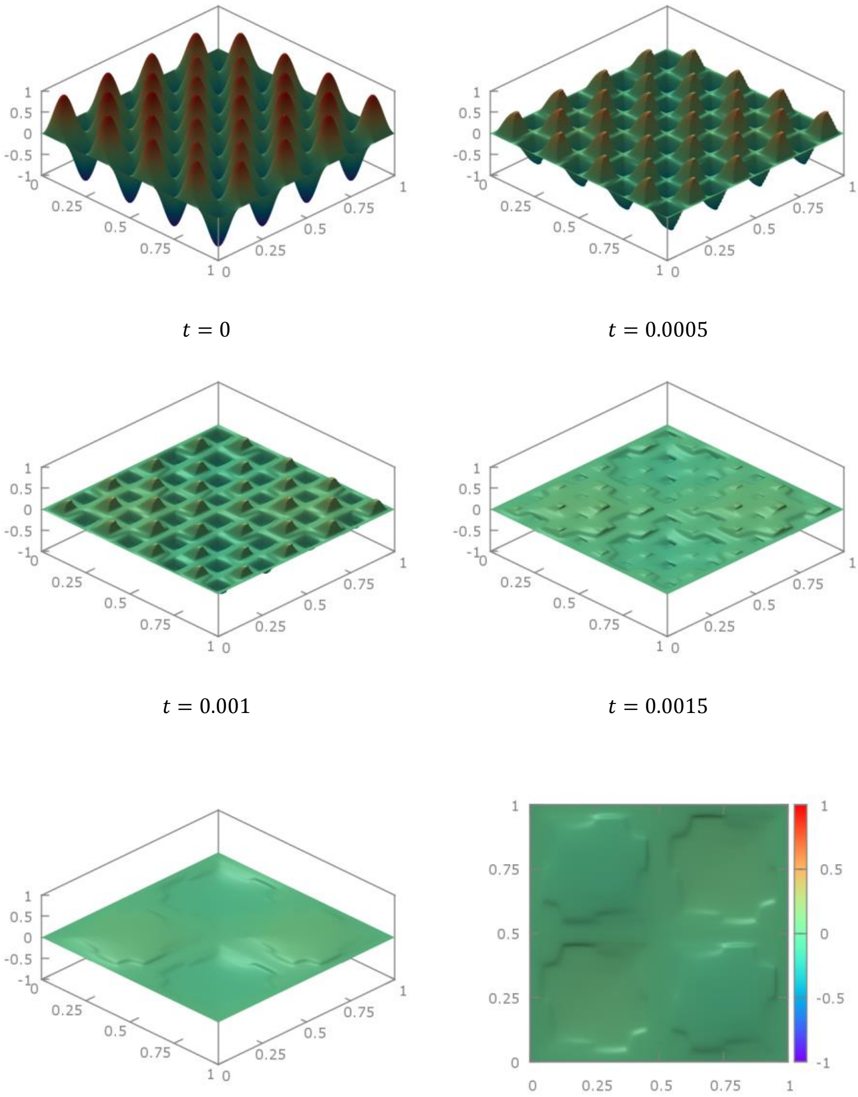
$t = 0$ $t = 0.005$

$t = 0.01$ $t = 0.015$

$t = 0.02$ $t = 0.02$

Figure 5

Morphology evolution for $f = 0$, the regularized l_1 norm (13) with $A_1 = 0.01$, $p_0(x_1, x_2) = \sin(2x_1)\sin(2x_2)$ at different times



$t = 0$

$t = 0.0005$

$t = 0.001$

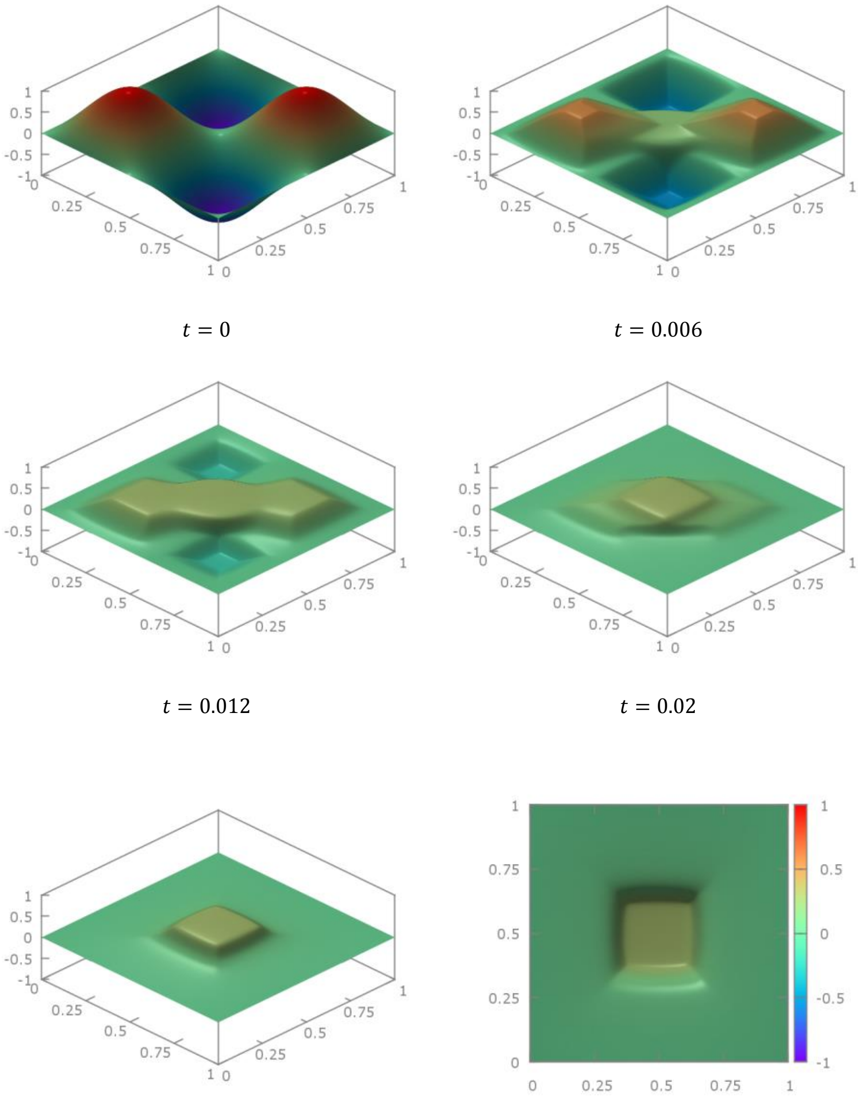
$t = 0.0015$

$t = 0.002$

$t = 0.002$

Figure 6

Morphology evolution for $f = 0$, the regularized l_1 norm (13) with $A_1 = 0.01$, $p_0(x_1, x_2) = 0.1\sin(2x_1)\sin(2x_2) + 0.9\sin(4x_1)\sin(4x_2)$ at different times



$t = 0$

$t = 0.006$

$t = 0.012$

$t = 0.02$

$t = 0.03$

$t = 0.03$

Figure 7

Morphology evolution for

$f = 10(1 - \tanh(20(\mathrm{sqrt}\big((x_1 - 0.5)(x_1 - 0.5) + (x_2 - 0.5)(x_2 - 0.5)\big) - 0.15)))$,

the regularized l_1 norm (13) with $A_1 = 0.01$, $p_0(x_1, x_2) = \sin(2x_1)\sin(2x_2)$ at different times

Conclusion

In the paper, we have studied the anisotropic mean curvature flow in relative geometry for which a global existence result has been derived. A numerical scheme based on the method of lines has been presented and analysed concerning its stability. In the numerical experiments, the influence of various anisotropy symmetries and the forcing term on the surface evolution has been addressed.

Acknowledgement

This work was partially supported by the project No. P108/12/1463 "Two scales discrete-continuum approach to dislocation dynamics" of the Grant Agency of the Czech Republic.

References

[1] Bellettini, G., Paolini, M.: Anisotropic Motion by Mean Curvature in the Context of Finsler Geometry, Hokkaido Mathematical Journal, Vol. 25, No. 3, pp. 537-566, 1996

[2] Beneš, M.: Diffuse-Interface Treatment of the Anisotropic Mean-Curvature Flow, Applications of Mathematics, Vol. 48, pp. 437-453, 2003

[3] Beneš, M., Hilhorst, D., Weidenfeld, R.: Interface Dynamics for an Anisotropic Allen-Cahn Equation, in *Nonlocal Elliptic and Parabolic Problems*, pp. 39-45, eds. Biler P., Karch G. and Nadzieja T., Banach Center Publications, Volume 66, 2004, Institute of Mathematics, Polish Academy of Sciences, Warszawa, 2004

[4] Deckelnick, K., Dziuk, G.: Discrete Anisotropic Curvature Ow of Graphs, ESAIM: Mathematical Modelling and Numerical Analysis, Vol. 33, No. 6, pp. 1203-1222, 1999

[5] Deckelnick, K., Dziuk, G.: Error Estimates for a Semi-Implicit Fully Discrete Finite Element Scheme for the Mean Curvature Flow of Graphs, Interfaces and Free Boundaries, Vol. 2, No. 4, pp. 341-359, 2000

[6] Deckelnick, K., Dziuk, G.: Convergence of Numerical Schemes for the Approximation of Level Set Solutions to Mean Curvature Flow, Numerical Methods for Viscosity Solutions and Applications, Vol. 59, pp. 77-93, 2001

[7] Deckelnick, K., Dziuk, G.: A Fully Discrete Numerical Scheme for Weighted Mean Curvature Flow, Numerische Mathematik, Vol. 91, No. 3, pp. 423-452, 2002

[8] Dziuk, G.: Discrete Anisotropic Curve Shortening Flow, SIAM J. Numer. Anal., Vol. 36, No. 6, pp. 1808-1830, 1999

[9] Gurtin, M. E.: On the Two-Phase Stefan Problem with Interfacial Energy and Entropy, Archive for Rational Mechanics and Analysis, Vol. 96, No. 3, pp. 199-241, 1986

[10] Haußer, F., Voigt, A.: A Numerical Scheme for Regularized Anisotropic Curve Shortening Flow, Applied Mathematics Letters, Vol. 19, No. 8, pp. 691-698, 2006

[11] Huisken, G.: Non-Parametric Mean-Curvature Evolution with Boundary-Conditions, Journal of Differential Equations, Vol. 77, No. 2, pp. 369-378, 1989

[12] Lieberman, G. M.: The First Initial-Boundary Value Problem for Quasilinear Second Order Parabolic Equations, Annali della Scuola Normale Superiore di Pisa - Classe di Scienze, Vol. 13, No. 3, pp. 347-387, 1986

[13] Oberman, A. M.: A Convergent Monotone Difference Scheme for Motion of Level Sets by Mean Curvature, Numer. Math., Vol. 99, No. 2, pp. 365-379, 2004

[14] Pozzi, P.: Anisotropic Mean Curvature Flow for Two-Dimensional Surfaces in Higher Codimension: a Numerical Scheme, Interfaces and Free Boundaries, Vol. 10, No. 4, pp. 539-576, 2008

[15] Visintin, A.: Models of Phase Transitions, Birkhäuser, Boston, 1996

On Irregularities of Bidegreed Graphs

Tamás Réti[1], **Darko Dimitrov**[2]

[1]Óbuda University
Bécsi út 96/B, H-1034 Budapest, Hungary
E-mail: reti.tamas@bgk.uni-obuda.hu

[2]Institut für Informatik, Freie Universität Berlin
Takustraße 9, D–14195 Berlin, Germany
& Hochschule für Technik und Wirtschaft Berlin
Wilhelminenhofstraße 75A, D–12459 Berlin, Germany
E-mail: dimdar@zedat.fu-berlin.de

Abstract: A graph is *regular* if all its vertices have the same degree. Otherwise a graph is *irregular*. To measure how irregular a graph is, several graph topological indices were proposed including: the *Collatz-Sinogowitz index* [8], the *variance of the vertex degrees* [7], the *irregularity of a graph* [4], and recently proposed the *total irregularity of a graph* [1]. Here, we compare the above mentioned irregularity measures for bidegreed graphs.

Keywords: topological graph indices; complete split graph; 2-walk linear graph

1 Introduction

All graphs considered here are simple and undirected. Let G be a graph of order $n = |V(G)|$ and size $m = |E(G)|$. For $v \in V(G)$, the degree of v, denoted by $d_G(v)$, is the number of edges incident to v. The *adjacency matrix* $A(G)$ of a graph G is a matrix with rows and columns labeled by graph vertices, with a 1 or a 0 in position (v_i, v_j) according to whether vertices v_i and v_j are adjacent or not. The *characteristic polynomial* $\phi(G,t)$ of G is defined as characteristic polynomial of $A(G)$: $\phi(G,\lambda) = \det(\lambda \mathbf{I_n} - A(G))$, where $\mathbf{I_n}$ is $n \times n$ identity matrix. The set of eigenvalues of the adjacent matrix $A(G)$ of a graph G is called a *graph spectrum.* The largest eigenvalue of $A(G)$, denoted by $\rho(G)$, is called the *spectral radius* of G. An eigenvalue of a graph G is called *main eigenvalue* if it has an eigenvector the sum of whose entries is not equal to zero.

In the sequel, we present the irregularity measures consider in this paper. Collatz-Sinogowitz [8] introduced the irregularity measure of a graph G as

$$\mathrm{CS}(G) = \rho(G) - \frac{2m}{n}. \tag{1}$$

An alternative to CS(G) is the *variance of the vertex degrees*

$$\mathrm{Var}(G) = \frac{1}{n}\sum_{i=1}^{n} d_G^2(v_i) - \frac{1}{n^2}\left(\sum_{i=1}^{n} d_G(v_i)\right)^2 . \tag{2}$$

Bell [7] was first who has compared CS(G) and Var(G) and showed that they are not always compatible. Albertson [4] defines the *imbalance* of an edge $e = uv \in E$ as $|d_G(u) - d_G(v)|$ and the *irregularity* of G as

$$\mathrm{irr}(G) = \sum_{uv \in E} |d_G(u) - d_G(v)| . \tag{3}$$

Recently, in [1] a new measure of irregularity of a simple, undirected graph, so-called the *total irregularity*, was defined as

$$\mathrm{irr}_t(G) = \frac{1}{2}\sum_{u,v \in V(G)} |d_G(u) - d_G(v)| . \tag{4}$$

More about the above presented irregularity measures, comparison studies of them, and other attempts to measure the irregularity of a graph, one can find in [3, 6, 10–12]. It is interesting that the above four irregularity measures are not always compatible for some pairs of graphs. In this paper we study the relations between the above mentioned irregularity measures for bidegreed graphs.

A *universal* vertex is the vertex adjacent to all other vertices. A set of vertices is said to be *independent* when the vertices are pairwise non-adjacent. The vertices from an independent set are *independent vertices.*

The *degree set*, denoted by $\mathcal{D}(G)$, of a simple graph G is the set consisting of the distinct degrees of vertices in G.

The *distance* between two vertices in a graph is the number of edges in a shortest path connecting them. The *eccentricity* of a vertex v in a connected graph G is the maximum graph distance between v and any other vertex of G. The *radius* of a graph G, denoted by rad(G), is the minimum graph eccentricity of any graph vertex of G. The *diameter* of a graph G, denoted by diam(G), is the maximal graph eccentricity of any graph vertex of G.

Let $m_{r,s}$ denotes the number of edges in G with end-vertex degrees r and s, and let n_r denotes the numbers of vertices n G with degree r. Numbers $m_{r,s}$ and n_r are referred as the *edge-parameters* and the *vertex-parameters* of G, respectively . The *mean degree* of a graph G is defined as $\overline{d}(G) = 2m/n$. Graphs G_1 and G_2 are said to be *edge-equivalent* if for their corresponding edge-parameters sets $\{m_{r,s}(G_1) > 0\} = \{m_{r,s}(G_2) > 0\}$ holds. Analogously, they are called *vertex-equivalent* if for their vertex-parameters sets $\{n_r(G_1) > 0\} = \{n_r(G_2) > 0\}$ is fulfilled. It is easy to see that if two graphs are edge-equivalent, then they are vertex-equivalent, as well.

For two graphs G_1 and G_2 with disjoint vertex sets $V(G_1)$ and $V(G_2)$ and disjoint edge sets $E(G_1)$ and $E(G_2)$ the *disjoint union* of G_1 and G_2 is the graph $G = G_1 \cup G_2$ with the vertex set $V(G_1) \cup V(G_2)$ and the edge set $E(G_1) \cup E(G_2)$. The *join* $G + H$ of simple undirected graphs G and H is the graph with the vertex set $V(G + H) = V(G) \cup V(H)$ and the edge set $E(G + H) = E(G) \cup E(H) \cup \{uv : u \in V(G),\ v \in V(H)\}$. Let C_n denote a cycle on n vertices. Further, let K_n denote the complete graph on n vertices, and tK_1 denote the graph with t isolated vertices and no edges.

A graph G is a *complete k-partite* graph if there is a partiton $V_1 \cup \cdots \cup V_k = V(G)$ of the vertex set, such that $uv \in E(G)$ if and only if u and v are in different parts of the partition. A connected bipartite graph G is *semiregular* if every edge of G joins a vertex of degree δ to a vertex of degree $\varDelta$. A connected graphs G is called a *balanced irregular* graph if the equality $\mathrm{irr}(G) = \mathrm{irr}_t(G)$ holds.

The rest of the paper is structured as follows. In Section 2 we present some types of bidegreed graphs and some known results related to the above mentioned irregularity measures. In Section 3 we investigate new relations between irregularity indices of bidegreed graphs. Bidegreed graphs with same irregularity indices are investigated in Section 4. We conclude with final remarks and open problems in Section 5.

2 Some types of bidegreed graphs and known results

A graph G is called *bidegreed* if its degree set $\mathcal{D}(G) = \{\varDelta, \delta\}$ with $\varDelta > \delta \geq 1$. In the sequel, we present some special types of connected bidegreed graphs that will be of interest later.

i) A bidegreed graph is called a *balanced bidegreed graph* if the equality $n_\varDelta n_\delta = m_{\varDelta,\delta}$ holds for it. It should be noted that the complete bipartite graphs, for which $m = m_{\varDelta,\delta} = n_\varDelta n_\delta = \varDelta\delta$ holds, form a subset of balanced bidegreed graphs.

ii) A balanced bidegreed graph with n vertices is called a *complete split graph* if it contains $q = n_\varDelta \geq 1$ universal vertices and $n-q$ independent vertices [5]. Thus, a complete split graph, denoted by $G_{cs}(n, q)$, can be obtained as join of $n - q$ graphs K_1 and the complete graph K_q, i.e., $G_{cs}(n, q) = (n - q)K_1 + K_q$. An existing complete split graph $G_{cs}(n, q)$ is uniquely defined by their parameters n and q. This implies that two complete split graphs with identical n, q parameters are isomorphic. For a complete split graph the equalities $m = m_{\varDelta,\delta} + m_{\varDelta,\varDelta}$ and $2m = (2n - 1)\delta - \delta^2$ hold [5].

iii) A balanced bidegreed graph is called a *complete split-like graph*, denoted by $G_{csl}(n, q, \delta)$, if it has $q \geq 1$ universal vertices. This implies that for a complete split-like graph the equality $qn_\delta = m_{\varDelta,\delta}$ holds. The complete split graphs represent a subset of complete split-like graphs. It is easy to see that if G is a complete split-like graph then the equalities $\mathrm{rad}(G) = 1$ and $\mathrm{diam}(G) = 2$ are fulfilled. In Fig. 1 non-isomorphic complete split-like graphs with 5 and 6 vertices are depicted. Note that they are not complete split graphs.

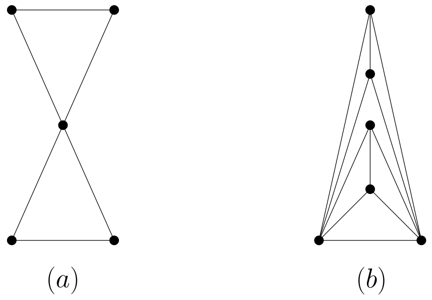


Fig. 1. Complete split-like graphs (a) $G_{csl}(5, 1, 2)$ and (b) $G_{csl}(6, 2, 3)$

Also note that since for a complete split-like graph G $qn_\delta = m_{\Delta,\delta}$, it follows that if G is not a complete bipartite graph, then G is non-bipartite and contains a triangle.

iv) In a particular case, if $q = 1$, then a complete split-like graph is called a *generalized windmill graph* and is denoted by $G_{csl}(n, 1, \delta)$. We would like to recall that the classical windmill graph, denoted by $W_d(k, p)$, can be constructed by joining p copies of the complete graph K_k with a common vertex. For a generalized windmill graph the equality $m = m_{\Delta,\delta} + m_{\delta,\delta}$ is fulfilled. It follows that the star graphs S_n with $n \geq 3$ vertices, the wheel graphs W_n with $n \geq 5$ vertices, and the classical windmill graphs $W_d(k, p)$ with $(k-1)p + 1$ vertices and $pk(k-1)/2$ edges defined for $k \geq 2$ and $p \geq 2$ positive integers, form the subsets of generalized windmill graphs. In Fig. 2 two non-isomorphic generalized windmill graphs are depicted.
Next, we state some known results that will be used afterwords.

Lemma 1 ([16]). *Let G be a connected bidegreed graph with spectral radius $\rho(G)$. Then*

$$\rho(G) = \sqrt{\frac{1}{n} \sum_{u \in V(G)} d^2(u)} = \sqrt{\Delta\delta},$$

if and only if G is a semiregular connected bipartite graph.

Lemma 2 ([15]). *Let G be a connected graph with mean degree $\overline{d}(G)] = 2m/n$, and just two main eigenvalues, ρ and $\mu < \rho$, where ρ is the spectral*

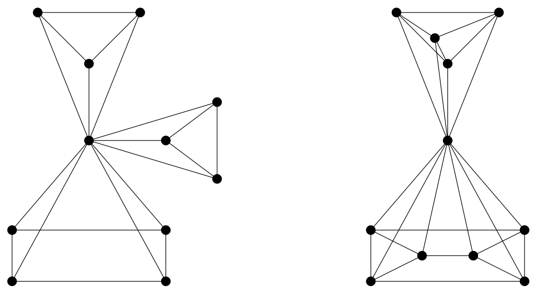


Fig. 2. Two generalized windmill graphs

radius of G. Then

$$\mathrm{Var}(G) = \frac{1}{n}\sum_{u\in V(G)} d^2(u) - \left(\frac{2m}{n}\right)^2 = \left(\rho - \frac{2m}{n}\right)\left(\frac{2m}{n} - \mu\right).$$

Lemma 3 ([15]). *Let G be a connected graph with spectral radius ρ. Then G is a semiregular bipartite graph if and only if the main eigenvalues of G are ρ and $-\rho$.*

Lemma 4 ([13]). *Let G be a connected graph with spectral radius ρ. Then*

$$\rho(G) \le \frac{\delta - 1 + \sqrt{(\delta+1)^2 + 4(2m - \delta n)}}{2}.$$

Equality holds if and only if G is regular or a bidegreed graph in which each vertex is of degree either δ or $n-1$.

3 Relations between irregularity indices - new results

In this section, we present some new results about the relations between irregularity indices of bidegreed graphs. We start with the following simple proposition.

Proposition 1. *Let $G(\Delta, \delta)$ be a connected bidegreed graph having n_Δ and n_δ vertices with degree Δ and δ, respectively. Then the following relations hold:*

$$m = m(G(\Delta,\delta)) = m_{\Delta,\Delta} + m_{\Delta,\delta} + m_{\delta,\delta} \ge m_{\Delta,\delta}, \tag{5}$$

$$\mathrm{irr}(G(\Delta,\delta)) = m_{\Delta,\delta}(\Delta - \delta), \tag{6}$$

$$\mathrm{irr}_t(G(\Delta,\delta)) = n_\Delta n_\delta(\Delta - \delta) = n_\Delta(n - n_\Delta)(\Delta - \delta), \tag{7}$$

$$\mathrm{irr}_t(G(\Delta,\delta)) = \frac{n_\Delta n_\delta}{m_{\Delta,\delta}}\mathrm{irr}(G(\Delta,\delta)). \tag{8}$$

Proof. It is obvious that for a connected bidegreed graph $G(\varDelta, \delta)$ the equality $m = m_{\varDelta,\delta}$ holds if and only if $G(\varDelta, \delta)$ is semiregular. The equalities (6), (7) and (8) follow from the definitions of irregularity indices.

Because the function $f(n_\varDelta) = n - n_\varDelta$ has a maximum value for $n_\varDelta = n/2$, we have the following corollary.

Corollary 1. *For a connected bidegreed graph $G(\varDelta, \delta)$ it holds that*

$$\mathrm{irr}_t(G(\varDelta, \delta)) = n_\varDelta(n - n_\varDelta)(\varDelta - \delta) \leq \frac{n^2}{4}(\varDelta - \delta). \tag{9}$$

Inequality (9) is sharp. There exist bidegreed graphs with n vertices for which $\mathrm{irr}_t(G(\varDelta, \delta)) = n^2(\varDelta-\delta)/4$. Such bidegreed graphs with 8 vertex and deegre set $\{3, 4\}$ are shown in Fig. 3(a). These graphs are non edge-equivalent, but only vertex equivalent, and the equality $n_3 = n_4 = n/2 = 4$ holds for them. Another example of bidegreed graphs that satisfy equality in (9) is given in Fig. 3(b). Those graphs are with 8 vertices and have deegre set $\{2, 3\}$. They are edge-equivalent, and satisfy the equality $n_2 = n_3 = n/2 = 4$. It is interesting to note that the graphs in Fig. 3(b) are not only edge-equivalent ($m_{2,3} = 8, m_{3,3} = 2$), but they have identical spectral radius $(1 + \sqrt{17})/2$, as well. Consequently, all considered irregularity indices (CS, Var, irr and irrt) are identical for them.

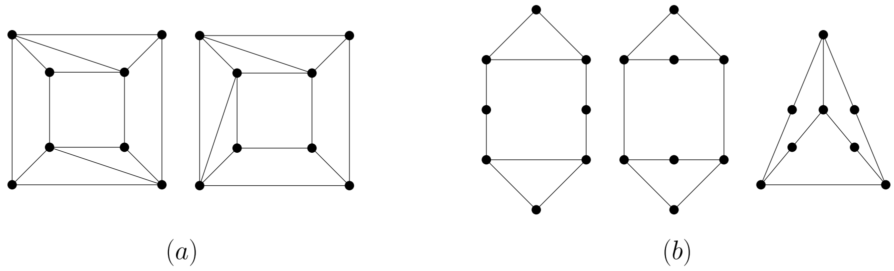


Fig. 3. Examples of non-isomorphic bidegreed graphs with 8 vertices with identical maximum total irregularity indices

Proposition 2. *Let $G(\varDelta, \delta)$ be a connected bidegreed graph, then*

$$\mathrm{irr}_t(G(\varDelta, \delta)) = \frac{\varDelta - \delta}{\varDelta\delta}\left(m^2 - (m_{\varDelta,\varDelta} - m_{\delta,\delta})^2\right) \leq \frac{\varDelta - \delta}{\varDelta\delta}m^2.$$

The equality holds if $m_{\varDelta,\varDelta} = m_{\delta,\delta}$.

Proof. For any bidegreed graph $G(\Delta, \delta)$, it holds that

$$\begin{aligned} \Delta n_{\Delta} &= m_{\Delta,\delta} + 2m_{\Delta,\Delta}, \quad \text{and} \\ \delta n_{\Delta} &= m_{\Delta,\delta} + 2m_{\delta,\delta}. \end{aligned}$$

This together with (7) implies that

$$\operatorname{irr}_t(G(\Delta, \delta)) = \frac{\Delta - \delta}{\Delta\delta}(m_{\Delta,\delta} + 2m_{\Delta,\Delta})(m_{\Delta,\delta} + 2m_{\delta,\delta}).$$

Since $m_{\Delta,\delta} = m - m_{\Delta,\Delta} - m_{\delta,\delta}$, it follows that

$$\operatorname{irr}_t(G(\Delta, \delta)) = \frac{\Delta - \delta}{\Delta\delta}\left(m^2 - (m_{\Delta,\Delta} - m_{\delta,\delta})^2\right) \leq \frac{\Delta - \delta}{\Delta\delta} m^2. \tag{10}$$

The equality in (10) is obtained when $m_{\Delta,\Delta} = m_{\delta,\delta}$. This condition holds for the bidegreed graphs with 10 vertices and 12 edges in Fig. 4. Consequently all of them have the same maximum total irregularity index $\operatorname{irr}_t = n_2 n_3 = 6 \cdot 4 = 24$.

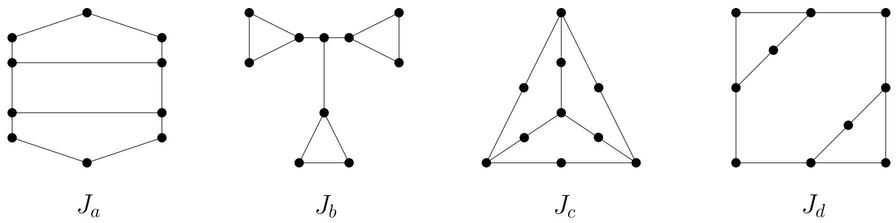

J_a J_b J_c J_d

Fig. 4. Bidegreed graphs having identical vertex degree set ($n_3 = 4, n_2 = 6$) and identical maximum total irregularity index $\operatorname{irr}_t = 24$

Among bidegreed graphs having identical vertex degree set (n_{Δ}, n_{δ}), the semiregular graphs (for which the equality $m_{\Delta,\Delta} = m_{\delta,\delta}$=0 holds) possess the maximal irregularity $\operatorname{irr}(G)$, as it is a case with graphs J_c and J_d in Fig. 4.

4 Bidegreed graphs with same irregularity indices

In the following we will show that there exists a broad class of bidegreed graphs having "similar irregularity", or in other words, there exist non-isomorphic graph pairs for which two (or more than two) irregularity indices are equal. Moreover, we will show that there are some particular classes of bidegreed graphs whose irregularity indices are considered algebraically dependent quantities.

4.1 Balanced bidegreed graphs

From the definition of balanced bidegreed graphs, it follows that

$$\mathrm{irr}(G(\mathit{\Delta},\delta)) = \mathrm{irr}_t(G(\mathit{\Delta},\delta)) = n_{\mathit{\Delta}} n_\delta(\mathit{\Delta}-\delta) = m_{\mathit{\Delta},\delta}(\mathit{\Delta}-\delta).$$

This implies that the balanced bidegreed graphs form a subset of balanced irregular graphs.

Proposition 3. *Let $G(\mathit{\Delta},\delta)$ be a balanced bidegreed graph for which $m_{\mathit{\Delta},\mathit{\Delta}} = 0$ or $m_{\delta,\delta} = 0$ hold. Then*

$$\mathrm{irr}(G(\mathit{\Delta},\delta)) = \mathrm{irr}_t(G(\mathit{\Delta},\delta)) = (2m - \mathit{\Delta}\delta)(\mathit{\Delta}-\delta).$$

Proof. For any bidegreed graph $G(\mathit{\Delta},\delta)$

$$\mathit{\Delta} n_{\mathit{\Delta}} = m_{\mathit{\Delta},\delta} + 2m_{\mathit{\Delta},\mathit{\Delta}},$$

$$\delta n_\delta = m_{\mathit{\Delta},\delta} + 2m_{\delta,\delta}.$$

Consequently, we get

$$n_{\mathit{\Delta}} n_\delta = m_{\mathit{\Delta},\delta} = \frac{(m_{\mathit{\Delta},\delta} + 2m_{\mathit{\Delta},\mathit{\Delta}})(m_{\mathit{\Delta},\delta} + 2m_{\delta,\delta})}{\mathit{\Delta}\delta}, \qquad \text{and}$$

$$m_{\mathit{\Delta},\delta}^2 + (2(m_{\mathit{\Delta},\mathit{\Delta}} + m_{\delta,\delta}) - \mathit{\Delta}\delta)m_{\mathit{\Delta},\delta} + 4m_{\mathit{\Delta},\mathit{\Delta}} m_{\delta,\delta} = 0.$$

Taking into consideration that $m_{\mathit{\Delta},\mathit{\Delta}} + m_{\delta,\delta} = m - m_{\mathit{\Delta},\delta}$, we have

$$m_{\mathit{\Delta},\delta}^2 + (\mathit{\Delta}\delta - 2m)m_{\mathit{\Delta},\delta} - 4m_{\mathit{\Delta},\mathit{\Delta}} m_{\delta,\delta} = 0.$$

Because $m_{\mathit{\Delta},\delta}$ is a positive number it is easy to see that the proper solution of the equation above is

$$n_{\mathit{\Delta}} n_\delta = m_{\mathit{\Delta},\delta} = \frac{1}{2}\left(2m - \mathit{\Delta}\delta + \sqrt{(2m-\mathit{\Delta}\delta)^2 + 16m_{\mathit{\Delta},\mathit{\Delta}} m_{\delta,\delta}}\right).$$

If as a particular case the equality $m_{\mathit{\Delta},\mathit{\Delta}} m_{\delta,\delta} = 0$ holds for graph $G(\mathit{\Delta},\delta)$, one obtains

$$n_{\mathit{\Delta}} n_\delta = m_{\mathit{\Delta},\delta} = 2m - \mathit{\Delta}\delta,$$

from which the main result follows.

Example 1. We present two infinite sequences of balanced bidegreed graphs with the property $m_{\mathit{\Delta},\mathit{\Delta}} m_{\delta,\delta} = 0$. The first infinite sequence is comprised of graphs $B(k)$, where k is a positive integer. The case $k = 2$ is depicted in Fig. $5(a)$. A graph $B(k)$ has a vertex degree distribution $n_3 = 2k$ and $n_{2k} = 2$, and edge number $m = 5k$, where $k \geq 2$ positive integer. It is easy to see that for graphs $B(k)$, the equality $m_{2k,2k} = 0$ holds.

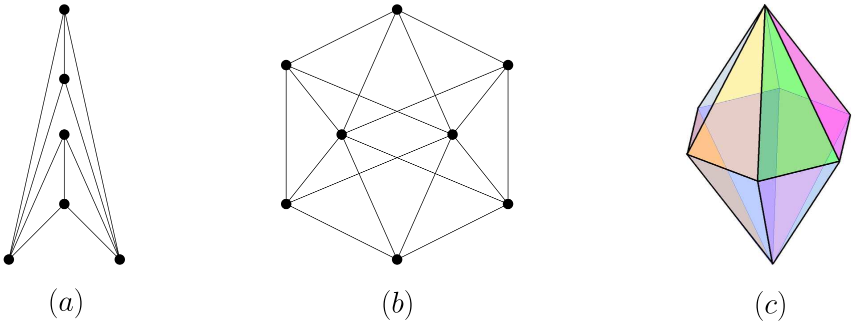


Fig. 5. Balanced planar bidegreed graphs. (a) Planar graph $B(2)$ and $(b), (c)$ Polyhedral graph $P(6)$ of 6-gonal bipyramid

The second infinite sequence is comprised of k-gonal bipyramids. A k-gonal bipyramid, with integer $k \geq 3$, is formed by joining a k-gonal pyramid and its mirror image base-to-base. It is a polyhedon having $2k$ triangular faces. The case $k = 6$ is depicted in Fig. 5(b) and redrawn in Fig. 5(c) for a better illustration. The graph $P(k)$ of a k-gonal bipiramid belongs to the family of balanced bidegreed graphs with degree 4 and k. For these graphs the equalities $n_4 n_k = m_{4k} = 2k, m = 3k$ and $m_{k,k} = 0$ hold.

4.2 Complete split graphs and complete split-like graphs

Proposition 4 ([2]). *There exist a complete split graph pairs with n vertices $G_{cs}(n,q)$ and $G_{cs}(n,q{+}1)$ with certain n and q positive integers, for which the equality* $\mathrm{irr}_t(G_{cs}(n,q)) = \mathrm{irr}_t(G_{cs}(n,q+1)) = \mathrm{irr}(G_{cs}(n,q)) = \mathrm{irr}(G_{cs}(n,q+1))$ *holds.*

Example 2. The smallest complete split graph pair with this property is the star graph on 5 vertices $G_{cs}(5,1)$, and the graph $G_{cs}(5,2)$ are depicted in Fig. 6.
For graphs $G_{cs}(5,1)$ and $G_{cs}(5,2)$ the following equality holds: $\mathrm{irr}_t(G_{cs}(5,1)) = \mathrm{irr}_t(G_{cs}(5,2)) = \mathrm{irr}(G_{cs}(5,1)) = \mathrm{irr}(G_{cs}(5,2)) = 12$.

Proposition 5. *Let $G_{csl}(n,q,\delta)$ be a complete split-like graph. Then*

$$\mathrm{irr}(G_{csl}(n,q,\delta)) = \mathrm{irr}_t(G_{csl}(n,q,\delta)) = q(n-q)(n-1-\delta).$$

Proof. Since the complete split-like graphs form a subset of balanced bidegreed graphs, it is easy to see that

$$\begin{aligned}\mathrm{irr}(G_{csl}(n,q,\delta)) &= m_{\Delta,\delta}|\Delta - \delta| = n_\Delta n_\delta |\Delta - \delta| = q(n-q)(n-1-\delta)\\ &= \mathrm{irr}_t(G_{csl}(n,q,\delta)).\end{aligned}$$

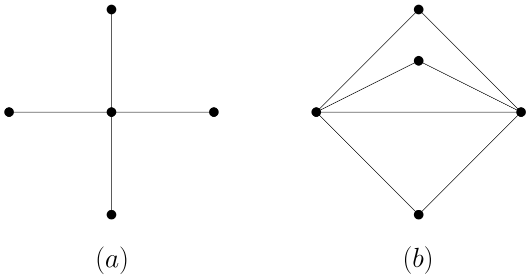


Fig. 6. Complete split graphs (a) $G_{cs}(5,1)$ and (b) $G_{cs}(5,2)$ with different degree sets

Proposition 6. *There exist complete split-like graph pairs* $G_{csl}(n_a, q_a, \delta_a)$ *and* $G_{csl}(n_b, q_b, \delta_b)$ *with different* $n_a, n_b, q_a, q_b, \delta_a$ *and* δ_b *parameters, for which the equality*

$$\begin{aligned}\mathrm{irr}_t(G_{csl}(n_a, q_a, \delta_a)) = \mathrm{irr}_t(G_{csl}(n_b, q_b, \delta_b)) &= \mathrm{irr}(G_{cs}(n_a, q_a, \delta_a)) \\ &= \mathrm{irr}(G_{cs}(n_b, q_b, \delta_b))\end{aligned}$$

holds.

Proof. A complete split-like graph pair with this property is the graph pair $G_{csl}(5,1,2)$ and $G_{csl}(6,2,4)$ depicted in Fig. 7. For these graphs, equality $\mathrm{irr}_t(G_{csl}(5,1,2)) = \mathrm{irr}_t(G_{csl}(6,2,4)) = \mathrm{irr}(G_{cs}(5,1,2)) = \mathrm{irr}(G_{cs}(6,2,4)) = 8$ holds.

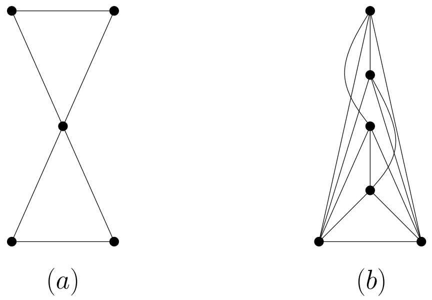


Fig. 7. Complete split-like graphs (a) $G_{csl}(5,1,2)$ and (b) $G_{csl}(6,2,4)$ with equal irr and irr_t measures

There are several ways to construct complete split-like graphs. For example, a complete split-like graph with n vertices $G_{csl}(n, q, \delta)$ can be generated using the following graph operations:

$$G_{csl}(n, q, \delta) = K_q + \left(\cup_{j=1}^{J} H(j, R)\right).$$

In the formula above, K_q is the complete graph on $q \geq 1$ vertices, $H(j,R)$ are $R \geq 1$ regular connected graphs for $j = 1, 2, \ldots, J$.

As an example, in Fig. 8 two non-isomorphic edge-equivalent complete split-like graphs are shown. These complete split-like graphs are defined as $G^1_{csl}(14,2,4) = K_2 + C_{12}$ and $G^2_{csl}(14,2,4) = K_2 + (C_3 \cup C_4 \cup C_5)$, respectively. It is easy to see that $\mathrm{irr}_t(G^1_{csl}(14,2,4)) = \mathrm{irr}_t(G^2_{csl}(14,2,4)) = \mathrm{irr}(G^1_{csl}(14,2,4)) = \mathrm{irr}(G^2_{csl}(14,2,4)) = 216$.

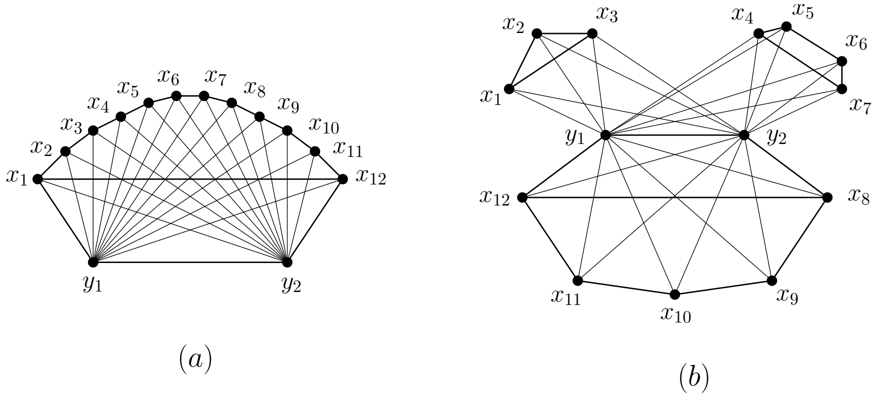


Fig. 8. Edge-equivalent complete split-like graphs, (a) $G^1_{csl}(14,2,4)$ and (b) $G^2_{csl}(14,2,4)$

From the previous considerations the following result follows.

Proposition 7. *Let G_1 and G_2 be edge-equivalent complete split-like graphs. Then the equalities $\mathrm{irr}_t(G_1) = \mathrm{irr}_t(G_2) = \mathrm{irr}(G_1) = \mathrm{irr}(G_2)$, $\mathrm{Var}(G_1) = \mathrm{Var}(G_2)$ and $\mathrm{CS}(G_1) = \mathrm{CS}(G_2)$ are fulfilled for them.*

Proof. Because G_1 and G_2 are edge-equivalent graphs, this implies that the equalities $\mathrm{irr}_t(G_1) = \mathrm{irr}_t(G_2), \mathrm{irr}(G_1) = \mathrm{irr}(G_2)$ and $\mathrm{Var}(G_1) = \mathrm{Var}(G_2)$ hold. Moreover, because G_1 and G_2 are complete split-like graphs, in which each vertex is of degree δ or $n-1$, it follows from Lemma 4 that their spectral radii are identical.

For an illustration of Proposition 7, see the complete split-like graph pair depicted in Fig. 8.

4.3 Semiregular graphs

It is important to note that except the complete bidegreed bipartite graphs, the semiregular graphs do not belong to the family of balanced bidegreed graphs.

Proposition 8. *Let $S_1(\varDelta_1, \delta_1)$ and $S_2(\varDelta_2, \delta_2)$ be semiregular graphs for which $\varDelta = \varDelta_1 = \varDelta_2, \delta = \delta_1 = \delta_2$, and $m_{\varDelta,\delta} = m(S_1) = m(S_2)$ hold. Then,*

$$\mathrm{CS}(S_1) = \mathrm{CS}(S_2) = \sqrt{\varDelta\delta} - \frac{2\varDelta\delta}{\varDelta + \delta},$$

and

$$\mathrm{Var}(S_i) = \left(\sqrt{\varDelta\delta} + \frac{2\varDelta\delta}{\varDelta + \delta}\right) \mathrm{CS}(S_i),$$

for $i = 1, 2$, where $\mathrm{CS}(G)$ is the Collatz-Sinogowitz irregularity index of a graph G.

Proof. It is easy to see that for a semiregular graph S with n vertices

$$n = n_\varDelta + n_\delta = \frac{m_{\varDelta,\delta}}{\varDelta} + \frac{m_{\varDelta,\delta}}{\delta} = \frac{\varDelta + \delta}{\varDelta\delta} m_{\varDelta,\delta}.$$

This implies that for the mean degrees $\overline{d}$ we have

$$\overline{d}(S_1) = \overline{d}(S_2) = \frac{2m_{\varDelta,\delta}}{n} = \frac{2\varDelta\delta}{\varDelta + \delta}.$$

From Lemma 1 one obtains

$$\rho = \rho(S_1) = \rho(S_2) = \sqrt{\varDelta\delta}.$$

consequently, we have

$$\mathrm{CS}(S_1) = \mathrm{CS}(S_2) = \rho - \frac{2m_{\varDelta,\delta}}{n} = \sqrt{\varDelta\delta} - \frac{2\varDelta\delta}{\varDelta + \delta}.$$

Moreover, from Lemmas 2 and 3, it follows that for a semiregular graphs S

$$\begin{aligned}\mathrm{Var}(S) &= \left(\rho - \frac{2m}{n}\right)\left(\frac{2m}{n} + \rho\right) = \rho^2 - \left(\frac{2m}{n}\right)^2 = \varDelta\delta - \left(\frac{2\varDelta\delta}{\varDelta + \delta}\right)^2 \\ &= \left(\sqrt{\varDelta\delta} + \frac{2\varDelta\delta}{\varDelta + \delta}\right) \mathrm{CS}(S).\end{aligned}$$

This implies that

$$\mathrm{Var}(S_i) = \left(\sqrt{\varDelta\delta} + \frac{2\varDelta\delta}{\varDelta + \delta}\right) \mathrm{CS}(S_i).$$

for $i = 1, 2$.

Proposition 9. *Let $S_1(\varDelta_1, \delta_1)$ and $S_2(\varDelta_2, \delta_2)$ be semiregular graphs for which $\varDelta = \varDelta_1 = \varDelta_2, \delta = \delta_1 = \delta_2$, and $m_{\varDelta,\delta} = m(S_1) = m(S_2)$ hold. Then, the equalities $\mathrm{irr}_t(S_1) = \mathrm{irr}_t(S_2), \mathrm{irr}(S_1) = \mathrm{irr}(S_2)$ are fulfilled for them.*

Proof. It is obvious that

$$\operatorname{irr}(S_1) = \operatorname{irr}(S_2) = m_{\Delta,\delta}(\Delta - \delta).$$

Moreover, because for a semiregular graph

$$n_\Delta n_\delta = \frac{m^2_{\Delta,\delta}}{\Delta\delta},$$

we get

$$\operatorname{irr}_t(S_1) = \operatorname{irr}_t(S_2) = n_\Delta n_\delta(\Delta - \delta) = \frac{\Delta - \delta}{\Delta\delta} m^2_{\Delta,\delta}.$$

As a consequence of Proposition 8 and 9, we have the following result.

Corollary 2. *Let $S_1(\Delta_1, \delta_1)$ and $S_2(\Delta_2, \delta_2)$ be semiregular graphs for which $\Delta = \Delta_1 = \Delta_2, \delta = \delta_1 = \delta_2$, and $m_{\Delta,\delta} = m(S_1) = m(S_2)$ hold. Then the equalities $\operatorname{irr}_t(S_1) = \operatorname{irr}_t(S_2), \operatorname{irr}(S_1) = \operatorname{irr}(S_2), \operatorname{Var}(S_1) = \operatorname{Var}(S_2)$ and $\operatorname{CS}(S_1) = \operatorname{CS}(S_2)$ are fulfilled for them.*

Graphs J_c and J_d depicted in Fig. 4 satisfy Corollary 2. From Proposition 9, we have the following corollary.

Corollary 3. *Let $S(\Delta, \delta)$ be a semiregular graph. Then,*

$$\operatorname{irr}_t(S(\Delta, \delta)) = \frac{\operatorname{irr}^2(S(\Delta, \delta))}{\Delta\delta(\Delta - \delta)}.$$

4.4 Bidegreed graphs with identical CS, Var, irr and irr_t indices

In Fig. 3(b), Proposition 7 and Corollary 2 examples of pairs of bidegreed graphs were presented, with the property that both graphs from a given pair have identical $\operatorname{CS}, \operatorname{Var}, \operatorname{irr}$ and irr_t. Next, we present another such pair of graphs. A 6-vertex graph pair with degree set $\{2, 3\}$ and with identical $\operatorname{CS}, \operatorname{Var}, \operatorname{irr}$ and irr_t indices is depicted in Fig. 9. These graphs are edge-equivalent ($m_{2,3} = 4, m_{3,3} = 4$), and they have identical spectral radius $1 + \sqrt{3}$.

In the sequel, we show that there exists an infinitely large family of pairs of bidegreed graphs with identical $\operatorname{CS}, \operatorname{Var}, \operatorname{irr}$ and irr_t indices . For that purpose, first we need the following definition:

Let $d_2(v)$ denote the sum of the degrees of all vertices adjacent to a vertex v in a graph G. Then, G is called 2-*walk linear* (more precisely, 2-*walk* (a, b)-*linear*)) if there exists a unique rational numbers pair (a, b) such that

$$d_2(v) = a \cdot d(v) + b$$

holds for every vertex v of G.

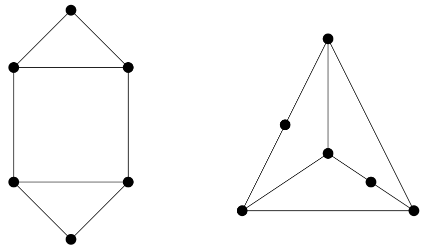


Fig. 9. Tricyclic, bidegreed, edge equivalent graph pair with identical spectral radius $1+\sqrt{3}$ [9]

Lemma 5 ([14]). *A graph G has exactly two main eigenvalues if and only if G is 2-walk linear. Moreover, if G is a 2-walk (a,b)-linear connected graph, then parameters a and b must be integers, and the spectral radius of G is*

$$\rho = \frac{1}{2}\left(a+\sqrt{a^2+4b}\right).$$

Using the above lemma we will demonstrate by examples that there are infinitely many bidegreed graph pairs having identical irregularity indices $\mathrm{CS}, \mathrm{Var}, \mathrm{irr}$ and irr_t.

Example 3. Consider the two infinite sequences of bidegreed graphs denoted by $G_a(k)$ and $G_b(k)$ (an illustration when $k=5$ is given in Fig. 10). Both $G_a(k)$ and $G_b(k)$ are of order $3k$, where $k \geq 3$. Graphs $G_a(k)$ and $G_b(k)$ are edge-equivalent, because the identities $m_{2,2}=k, m_{2,4}=2k, m_{4,4}=k, m=4k$ are fulfilled. Moreover, $G_a(k)$ and $G_b(k)$ are 2-walk $(3,0)$ linear graphs. By Lemma 5, it follows that they have identical spectral radius which is equal to 3. It is easy to show that for graphs $G_a(k)$ and $G_b(k)$ the following equalities hold: $\mathrm{CS}(G_a(k)) = \mathrm{CS}(G_b(k)) = 1/3$, $\mathrm{Var}(G_a(k)) = \mathrm{Var}(G_b(k)) = 8/9$,

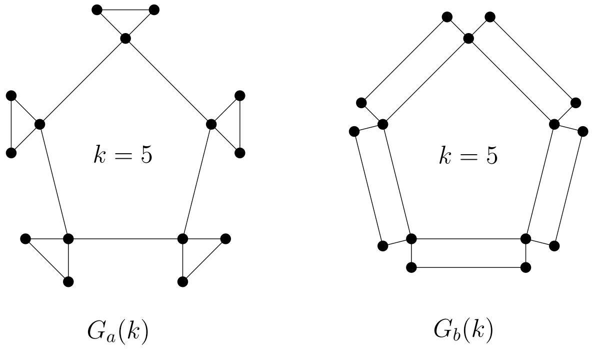


Fig. 10. Bidegreed graph pair $G_a(5)$ and $G_b(5)$

$\text{irr}(G_a(k)) = \text{irr}(G_b(k)) = 4k$, and $\text{irr}_t(G_a(k)) = \text{irr}_t(G_b(k)) = 8k^2$. It is interesting to note that $\text{irr}(G_a(k))/n = \text{irr}(G_b(k))/n = 4/3$, and $\text{irr}_t(G_a(k))/n^2 = \text{irr}_t(G_b(k))/n^2 = 4/9$, for any $k \geq 3$.

Example 4. Another infinite sequence of bidegreed graph pairs denoted by $H_a(k)$ and $H_b(k)$ is shown in Fig. 11. Each of them has $n = 4k$ vertices, where $k \geq 2$. Graphs $H_a(k)$ and $H_b(k)$ are edge-equivalent, because the identities $m_{2,3} = 4k = n$, $m_{3,3} = k$, and $m = 5k$ hold. It is easy to see that $H_a(k)$ and $H_b(k)$ are 2-walk $(1, 4)$ linear graphs. From this it follows that they have identical spectral radius which is equal to $\left(1 + \sqrt{17}\right)/2$. For graphs $H_a(k)$ and $H_b(k)$ the following equalities hold: $\text{CS}(H_a(k)) = \text{CS}(H_b(k)) = (\sqrt{17} - 4)/2$, $\text{Var}(H_a(k)) = \text{Var}(H_b(k)) = 1/4$, $\text{irr}(H_a(k))/n = \text{irr}(H_b(k))/n = 1$, and $\text{irr}_t(H_a(k))/n^2 = \text{irr}_t(H_b(k))/n^2 = 1/4)$.

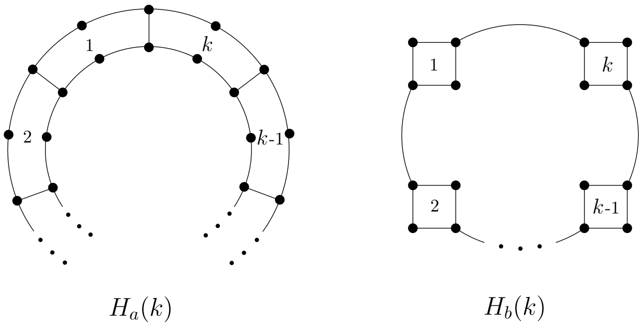


Fig. 11. Bidegreed graph pair $H_a(k)$ and $H_b(k)$

Example 5. Semi-regular bidegreed graph pairs denoted by $J_a(k)$ and $J_b(k)$ are shown in Fig. 12. Both of them are comprised of $n = 5k$ vertices, where $k \geq 2$. Graphs $J_a(k)$ and $J_b(k)$ are edge-equivalent, since the identity $m_{2,3} = 6k$ is fulfilled. Moreover, these graphs are 2-walk $(0, 6)$ linear. Consequently, they have identical spectral radius which is equal to $\sqrt{6}$. For graphs $J_a(k)$ and $J_b(k)$ the following equalities hold: $\text{CS}(J_a(k)) = \text{CS}(J_b(k)) = \sqrt{6} - 12/5$, $\text{Var}(J_a(k)) = \text{Var}(J_b(k)) = 6/25$, $\text{irr}(J_a(k))/n = \text{irr}(J_b(k))/n = 6/5$, and $\text{irr}_t(J_a(k))/n^2 = \text{irr}_t(J_b(k))/n^2 = 6/25$.

4.5 Smallest bidegreed graphs with identical irregularity indices

In this section we present pairs of smallest graphs that have identical two or more irregularity measures. The results were obtained by computer search. For two graphs of same order $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$, we said that G_1 *is smaller than* G_2 if $|E_1| < |E_2|$. Consequently, for two pairs of graphs of same order $D_1 = (G_1, G_2)$ and $D_2 = (G_3, G_4)$, we said that D_1 *is smaller than* D_2 if $|E_1| + |E_2| < |E_3| + |E_4|$.

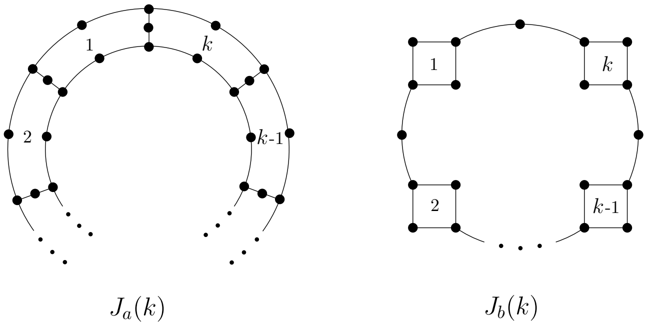


Fig. 12. Bidegreed graph pair $J_a(k)$ and $J_b(k)$

First, in Fig. 13(a) the smallest pair of graphs, that have identical all four irregularity indices $\mathrm{CS}, \mathrm{Var}, \mathrm{irr}$ and irr_t, is presented. The graphs G_1 and G_2 are of order 6 and size 7. Their $\mathrm{CS}, \mathrm{Var}, \mathrm{irr}$ and irr_t indices are 0.080880, 0.266667, 4, and 8, respectively. They also have same spectral radius which is 2.414214. We note that the pair (G_1, G_2) is at same time the smallest pair of graphs with equal CS index.

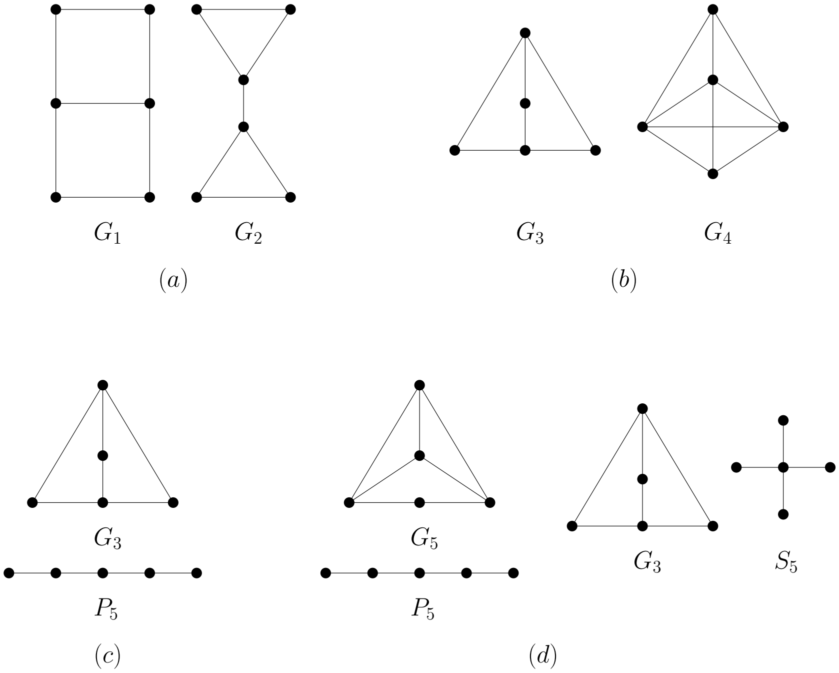


Fig. 13. Smallest bidegreed graphs with identical irregularity indices

In Fig. 13(b) the smallest pair of graphs, that have identical $\mathrm{Var}, \mathrm{irr}$ and irr_t indices is presented. This pair is also the smallest pair with the property that both graphs have equal Var and irr indices. The graphs G_3 and G_4 are of order 5 and sizes 6 and 9, respectively. Their $\mathrm{Var}, \mathrm{irr}$ and irr_t indices are 0.300000, 6, and 6, respectively.

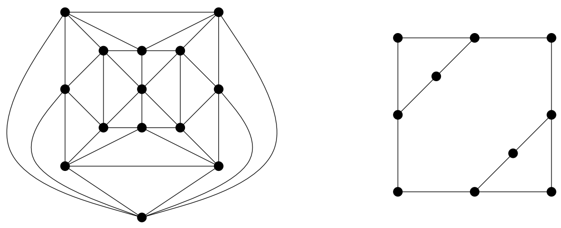
The pair (P_5, G_3), depicted in Fig. 13(c), is the smallest pair with the property that both graphs have equal Var and irr_t indices. At same time, it is the smallest pair with both graphs having equal Var index. Also, it is the smallest pair with both graphs having equal irr_t index. Their Var and irr_t indices are 0.300000 and 6, respectively.

The pair (S_5, G_3), depicted in Fig. 13(d), is the smallest pair with the property that both graphs have equal irr and irr_t indices. It holds that $\mathrm{irr}(S_5) = irr(G_3) = 12$ and $\mathrm{irr}_t(S_5) = irr_{(}G_3) = 12$. At same time, together with the pair (P_5, G_5), it is the smallest pair with both graphs having equal irr index.

5 Final remarks and open problems

In this paper we focused our investigation to the study of the relations between the irregularity indices of bidegreed connected graphs. Comparing the irregularity indices of various graphs, in the majority of cases it was supposed that the number of vertices or the corresponding degree sets are identical (see Figures 3, 4, 6, 8, 9, 10, 11, 12, 13(a)). It would be interesting to consider graphs of same order which have different degree sets, but their corresponding irregularity indices are identical (as few examples in Fig. 13(b),(c),(d)).

Another interesting problem is to estimate the maximum possible difference of vertex and edge numbers of graphs having identical irregularity indices (assuming that such positive finite integer exists.) Both cases, when graphs are of same or different order, are of interest. In Fig. 14, bidegreed graphs $B(6,5)$ and $B(3,2)$ represent an example concerning this problem. We would like to note that, the bidegreed polyhedral graph $B(6,5)$ is the dual of the



$B(6,5)$ $B(3,2)$

Fig. 14. Bidegreed graphs with identical $\mathrm{irr}_t = 24$ and $\mathrm{irr} = 12$ indices

graph of the smallest C_{24} fullerene which is composed of 12 pentagonal and 2 hexagonal faces, and graph $B(3,2)$ is a semiregular graph. It is worth noting that graph $B(6,5)$ has 14 vertices and 36 edges, while graph $B(3,2)$ has 10 vertices and 12 edges. It is surprising that there is a large difference between the corresponding edge-numbers of the two graphs, $(36 - 12 = 24)$.

References

1. H. Abdo, D. Dimitrov, *The total irregularity of a graph*, arxiv.org/abs/1207.5267, 2012.
2. H. Abdo, N. Cohen, D. Dimitrov, *Bounds and computation of irregularity of a graph*, http://arxiv.org/abs/1207.4804, 2012.
3. Y. Alavi, J. Liu, J. Wang, *Highly irregular digraphs*, Discrete Math. **111** (1993) 3–10.
4. M. O. Albertson, *The irregularity of a graph*, Ars Comb. **46** (1997) 219–225.
5. M. Aouchiche, F. K. Bell, D. Cvetković, P. Hansen, P. Rowlinson, S. K. Simić, D. Stevanović, *Variable neighborhood search for extrenmal graphs. 16. Some conjectures related to the largest eivenvalue of a graph*, Eur. J. Oper. Res. **191** (2008) 661–676.
6. F. K. Bell, *On the maximal index of connected graphs*, Linear Algebra Appl. **144** (1991) 135–151.
7. F. K. Bell, *A note on the irregularity of graphs*, Linear Algebra Appl. **161** (1992) 45–54.
8. L. Collatz, U. Sinogowitz, *Spektren endlicher Graphen*, Abh. Math. Sem. Univ. Hamburg **21** (1957) 63–77.
9. X. Fan, Y. Luo, *Tricyclic graphs with exactly two main eigenvalues*, http://arxiv.org/abs/1012.0963, 2010.
10. I. Gutman, P. Hansen, H. Mélot, *Variable neighborhood search for extremal graphs.* 10. *Comparison of irregularity indices for chemical trees*, J. Chem. Inf. Model. **45** (2005) 222–230.
11. P. Hansen, H. Mélot, *Variable neighborhood search for extremal graphs.* 9. *Bounding the irregularity of a graph*, in Graphs and Discovery, DIMACS Ser. Discrete Math. Theoret. Comput. Sci **69** (2005) 253–264.
12. M. A. Henning, D. Rautenbach, *On the irregularity of bipartite graphs*, Discrete Math. **307** (2007) 1467–1472.
13. Y. Hong, J-L. Shu, K. Fang, *A sharp upper bound of the spectral radius of graphs,* J. Combin. Theory Ser. B **81** (2001) 177–183.
14. Y. Hou, F. Tian, *Unicyclic graphs with exactly two main eigenvalues,* Appl. Math. Lett. **19** (2006) 1143–1147.
15. P. Rowlinson, *The main eigenvalues of a graph: A survey,* Appl. Anal. Discrete Math. **1** (2007) 445–471.
16. A. Yu, M. Lu, F. Tian, *On the spectral radius of graphs,* Linear Algebra Appl. **387** (2004) 41–49.

Finite-Time Stability of Continuous Time Delay Systems: Lyapunov-like Approach with Jensen's and Coppel's Inequality

Dragutin Lj. Debeljkovic(1), Sreten B. Stojanovic(2), Aleksandra M. Jovanovic(1)

(1) Faculty of Mechanical Engineering, University of Belgrade, Department of Control Engineering, 11000 Belgrade, Serbia, e-mail: ddebeljkovic@mas.bg.ac.rs

(2) Faculty of Technology, University of Nis, Department of Engineering Sciences and Mathematics, 16000 Leskovac, Serbia, e-mail: sstojanovic@tf.ni.ac.rs

Abstract: In this paper, the finite-time stability (FTS) of linear continuous time-delay systems is studied. By using suitable Lyapunov-like function and Jensen's and Coppel's inequality, a FTS condition is derived as a set of algebraic inequalities. The comparison of this method with some previous one is done and it has been showed that the numerical computation is reduced.

Keywords: time delay systems; finite-time stability; continuous systems; Jensen's integral inequality; Coppel's inequality

1 Introduction

Asymptotic stability, BIBO stability and other classical stability concepts deal with systems operating over an infinite time interval. However, in many practical cases, larger values of the state variables are not allowable in the specified (finite) time interval. Then, instead of asymptotic stability, it is preferable to use the stability defined over a finite time interval, i.e. finite-time stability (FTS). A system will be FTS if its state does not exceed some previously defined limit, for a given time interval. This concept stability dates back to the 1950s [1-3]. In references [4-11] some controllers are proposed such that the feedback system is FTS.

Many technical systems, such as pneumatic, hydraulic and electric systems, as well as process systems in the chemical industry, possess time-delay. The stability analysis of time-delay systems is more complex because the time-delay impairs the system stability. Similar to the non-delay systems, we can define FTS for time-delay systems. In references [12-18], some basic results on FTS are derived. These results are conservative, because they use the inequalities based on the norm of state vector. Recently, using new boundary technique based on the vector and matrix inequality, integral or no integral type, some less conservative results are obtained [19–22].

This article considers a novel delay-dependent FTS sufficient condition of linear continuous time-delay systems. The combination of Lyapunov-like approach and two algebraic inequalities (Jensen's integral inequality and Coppel's inequality) is used to solve this stability problem. The condition is expressed in the form of a set of algebraic inequality.

Notation. $\Re^n$ and $\Re^{n\times m}$ denote the *n*-dimensional Euclidean space and set of all $n\times m$ real matrices. $X>0$ means that X is real positive definite symmetric matrix; $X>Y$ is equivalent to $X-Y>0$. $\mu(X)=1/2\lambda_{\max}(X+X^T)$ is matrix measure of matrix X.

2 Preliminaries and Problem Formulation

Consider the following linear time-delay system:

$$\dot{x}(t)=A_0x(t)+A_1x(t-\tau) \tag{1}$$

with a initial conditions:

$$x(t)=\phi(t),\quad t\in[-\tau,0] \tag{2}$$

where $x(t)\in\Re^n$ is the state vector, $A_0\in\Re^{n\times n}$, $A_1\in\Re^{n\times n}$ and $B\in\Re^{n\times m}$ are constant matrices and τ is time-delay.

In the process of derivation of the stability condition, following definition and three lemmas are used.

Definition 1. [22] Time-delay system (1) satisfying the given initial condition (2) is said to be finite-time stable (FTS) with respect to (α,β,T) if

$$\sup_{t\in[-\tau,0]}\ \phi^T(t)\phi(t)\le\alpha\quad\Rightarrow\quad x^T(t)x(t)<\beta,\ \forall t\in[0,T] \tag{3}$$

Lemma 1. [22] (Jensen's integral inequality) For any positive symmetric constant matrix $M\in\square^{n\times n}$, scalars a, b satisfying $a<b$, a vector function $f:[a,b]\to\square^{n}$ such that the integrations concerned are well defined, then:

$$\left(\int_a^b f(\theta)d\theta\right)^T M\left(\int_a^b f(\theta)d\theta\right) \le (b-a)\int_a^b f^T(\theta)\, M\, f(\theta)d\theta \tag{4}$$

Lemma 2. [12] (Coppel's inequality) For any real square matrix $M \in \square^{n\times n}$ and scalar variable t, the following expression holds:

$$\lambda_{\max}\left(e^{Mt}\cdot e^{M^T t}\right) \le e^{2\mu(M)t} \tag{5}$$

where $\mu(M)$ is matrix measure of the matrix M.

Lemma 3. For any symmetric positive definite matrix $\Gamma = \Gamma^T > 0$, the following expressions hold:

$$2u(t)v(t) \le u^T(t)\Gamma u(t) + v^T(t)\Gamma^{-1}v(t) \tag{6}$$

$$-2u(t)v(t) \le u^T(t)\Gamma u(t) + v^T(t)\Gamma^{-1}v(t) \tag{7}$$

3 Main Result

In this section Lyapunov-like approach will be used in order to find sufficient delay dependent FTS conditions of the time-delay system (1). The following lemma, that is necessary for the design of Lyapunov-like function, is developed. We note that the new result is based on the result given in [23].

Lemma 4. Let a scalar function $V(y(t))$ be defined by:

$$V(y(t)) = y^T(t)\,y(t) \tag{8}$$

where $y(t)$ is vector which is defined by:

$$y(t) = x(t) + \int_0^\tau Q(\theta)\,x(t-\theta)\,d\theta \tag{9}$$

$x(t) \in \Re^n$ is the state vector of the system (1), $Q(t) \in \Re^{n\times n}$ is continuous and differentiable matrix function over time interval $[0,\tau]$ satisfying the following differential matrix equation:

$$\dot{Q}(\vartheta) = \left(A_0 + Q(0)\right)Q(\vartheta), \quad \vartheta \in [0,\tau] \tag{10}$$

with initial condition:

$$Q(\tau) = A_1 \tag{11}$$

Then, derivative of $V\big(y(t)\big)$ is given with:

$$\dot{V}\big(y(t)\big) = y^T(t)\,\Xi\,y(t) \tag{12}$$

where:

$$\Xi = \big(A_0 + Q(0)\big)^T + \big(A_0 + Q(0)\big) \tag{13}$$

Proof. From (8), follows:

$$\begin{aligned}\dot{V}\big(y(t)\big) = &\left(\dot{x}^T(t) + \frac{d}{dt}\int_0^\tau x^T(t-\theta)Q^T(\theta)d\theta\right) \times \left(x(t) + \int_0^\tau Q(\eta)x(t-\eta)d\eta\right) \\ &+ \left(x^T(t) + \int_0^\tau x^T(t-\theta)Q^T(\theta)d\theta\right) \times \left(\dot{x}(t) + \frac{d}{dt}\int_0^\tau Q(\eta)x(t-\eta)d\eta\right)\end{aligned} \tag{14}$$

First derivative of the integral term $\int_0^\tau Q(\theta)x(t-\theta)d\theta$ can be determined as follows. From

$$\frac{d}{d\theta}\big(Q(\theta)x(t-\theta)\big) = \dot{Q}(\theta)x(t-\theta) + Q(\theta)\frac{\partial}{\partial\theta}\big(x(t-\theta)\big) \tag{15}$$

$$\frac{\partial}{\partial\theta}\big(x(t-\theta)\big) = -\frac{\partial}{\partial t}\big(x(t-\theta)\big) \tag{16}$$

we get:

$$\frac{d}{d\theta}\big(Q(\theta)x(t-\theta)\big) = \dot{Q}(\theta)x(t-\theta) - Q(\theta)\frac{\partial}{\partial t}\big(x(t-\theta)\big) \tag{17}$$

or rearranging:

$$Q(\theta)\frac{\partial}{\partial t}\big(x(t-\theta)\big) = \dot{Q}(\theta)x(t-\theta) - \frac{d}{d\theta}\big(Q(\theta)x(t-\theta)\big) \tag{18}$$

Using the following identity:

$$\frac{d}{dt}\big(Q(\theta)x(t-\theta)\big) = Q(\theta)\frac{\partial}{\partial t}\big(x(t-\theta)\big) \tag{19}$$

one can finally have:

$$\begin{aligned}\frac{d}{dt}\int_0^{\tau} Q(\theta)x(t-\theta)d\theta &= \int_0^{\tau}\dot{Q}(\theta)x(t-\theta)d\theta - \frac{d}{d\theta}\int_0^{\tau} Q(\theta)x(t-\theta)d\theta \\ &= \int_0^{\tau}\dot{Q}(\theta)x(t-\theta)d\theta - Q(\tau)x(t-\tau)+Q(0)x(t)\end{aligned} \tag{20}$$

Employing (11), we have:

$$\frac{d}{dt}\int_0^{\tau} Q(\theta)x(t-\theta)d\theta = \int_0^{\tau}\dot{Q}(\theta)x(t-\theta)d\theta - A_1\,x(t-\tau)+Q(0)x(t) \tag{21}$$

Finally, (14) becomes:

$$\begin{aligned}&\dot{V}(y(t)) = \\ &= \begin{pmatrix} x^T(t)A_0^T + x^T(t-\tau)A_1^T + \\ +\int_0^{\tau} x^T(t-\theta)\dot{Q}^T(\theta)d\theta - x^T(t-\tau)A_1^T + x^T(t)Q^T(0)\end{pmatrix}\times \\ &\times\left(x(t)+\int_0^{\tau}Q(\eta)x(t-\eta)d\eta\right) \\ &+\left(x^T(t)+\int_0^{\tau}x^T(t-\theta)Q^T(\theta)d\theta\right)\times \\ &\times\begin{pmatrix} A_0\,x(t)+A_1\,x(t-\tau)+ \\ +\int_0^{\tau}\dot{Q}(\eta)x(t-\eta)d\eta - A_1\,x(t-\tau)+Q(0)x(t)\end{pmatrix}\end{aligned} \tag{22}$$

or:

$$\begin{aligned}\dot{V}(y(t)) = &\left(x^T(t)A_0^T + x^T(t)Q^T(0)+\int_0^{\tau}x^T(t-\theta)\dot{Q}^T(\theta)d\theta\right)\times \\ &\times\left(x(t)+\int_0^{\tau}Q(\eta)x(t-\eta)d\eta\right) \\ &+\left(x^T(t)+\int_0^{\tau}x^T(t-\theta)Q^T(\theta)d\theta\right)\times \\ &\times\left(A_0\,x(t)+Q(0)x(t)+\int_0^{\tau}\dot{Q}(\eta)x(t-\eta)d\eta\right)\end{aligned} \tag{23}$$

and, after some simple manipulations, follows:

$$\begin{aligned}\dot{V}\left(y(t)\right) &= x^T(t)\Big(\left(A_0^T + Q^T(0)\right) + \left(A_0 + Q(0)\right)\Big)x(t)\\ &+x^T(t)\int_0^\tau \left(A_0^T Q(\eta) + Q^T(0)Q(\eta) + \dot{Q}(\eta)\right)x(t-\eta)\,d\eta\\ &+\left(\int_0^\tau x^T(t-\theta)\left(Q^T(\theta)A_0 + Q^T(\theta)Q(0) + \dot{Q}^T(\theta)\right)d\theta\right)x(t)\\ &+\int_0^\tau\int_0^\tau x^T(t-\theta)\left(\dot{Q}^T(\theta)Q(\eta) + Q^T(\theta)\dot{Q}(\eta)\right)x(t-\eta)\,d\theta d\eta\end{aligned} \tag{24}$$

By the virtue of (10), one can get:

$$\begin{aligned}\dot{V}\left(y(t)\right) &= x^T(t)\,\Xi\,x(t)\\ &+x^T(t)\int_0^\tau \left(A_0^T + Q^T(0) + A_0 + Q(0)\right)Q(\eta)x(t-\eta)\,d\eta\\ &+\left(\int_0^\tau x^T(t-\theta)Q^T(\theta)\left(A_0^T + Q^T(0) + A_0 + Q(0)\right)d\theta\right)x(t)\\ &+\int_0^\tau\int_0^\tau x^T(t-\theta)\begin{Bmatrix} Q^T(\theta)\left(A_0^T + Q^T(0)\right)Q(\eta) + \\ +Q^T(\theta)\left(A_0 + Q(0)\right)Q(\eta)\end{Bmatrix}x(t-\eta)\,d\theta d\eta\end{aligned} \tag{25}$$

that is:

$$\begin{aligned}\dot{V}\left(y(t)\right) = x^T(t)\,\Xi\,x(t) &+ x^T(t)\,\Xi\int_0^\tau Q(\eta)x(t-\eta)\,d\eta\\ &+\left(\int_0^\tau x^T(t-\theta)Q^T(\theta)\,d\theta\right)\Xi\,x(t)\\ &+\int_0^\tau\int_0^\tau x^T(t-\theta)\left(Q^T(\theta)\,\Xi\,Q(\eta)\right)x(t-\eta)\,d\theta d\eta\end{aligned} \tag{26}$$

as well as:

$$\begin{aligned}\dot{V}\left(y(t)\right) &= x^T(t)\,\Xi\left(x(t) + \int_0^\tau Q(\eta)x(t-\eta)\,d\eta\right)\\ &+\left(\int_0^\tau x^T(t-\theta)Q^T(\theta)\,d\theta\right)\Xi\left(x(t) + \int_0^\tau Q(\eta)x(t-\eta)\,d\eta\right)\end{aligned} \tag{27}$$

and finally:

$$\dot{V}\left(y(t)\right) = x^T(t)\,\Xi\,y(t) + \left(\int_0^\tau x^T(t-\theta)Q^T(\theta)d\theta\right)\Xi\,y(t) \tag{28}$$

$$\dot{V}\left(y(t)\right) = \left[x^T(t) + \int_0^\tau x^T(t-\theta)Q^T(\theta)d\theta\right]\Xi\,y(t) \tag{29}$$

$$\dot{V}\left(y(t)\right) = y^T(t)\,\Xi\,y(t) \tag{30}$$

what completes the proof.

Previously derived result will be used to obtain the following stability condition.

Theorem 1. The time-delayed system (1)-(2) with is finite-time stable with respect to $\{\alpha,\beta,T\}$ if there exists a positive scalar $\wp$ such that:

$$\begin{gathered} x^T(t-\vartheta)x(t-\vartheta) < q\,x^T(t)x(t) \\ q>0,\quad \vartheta\in[-\tau,\,0],\quad \forall t\in[0,T] \end{gathered} \tag{31}$$

$$(1+\tau)(1+\psi)\left(1-\wp\psi-\frac{q\tau}{\wp}\right)^{-1} e^{\lambda_{\max}(\Xi)T} < \frac{\beta}{\alpha} \tag{32}$$

$$\begin{gathered} \wp\in\left(\max\{\wp_1,0\},\quad \wp_2\right), \\ \wp_{1,2} = \frac{1\pm\sqrt{1-4\psi\tau q}}{2\psi},\quad 4\psi\tau q<1 \end{gathered} \tag{33}$$

where:

$$R = A_0 + Q(0) \tag{34}$$

$$\Xi = R^T + R \tag{35}$$

$$\psi = \lambda_{\max}\left(Q(0)Q^T(0)\right)\frac{e^{2\mu_2(R)\tau}-1}{2\mu(R)} \tag{36}$$

$\mu_2(R)$ being matrix measure of matrix R and $Q(0)$ is any solution of the following nonlinear transcendental matrix equation:

$$e^{A_0+Q(0)\tau}Q(0) = A_1 \tag{37}$$

Proof. From (12) follows:

$$\dot{V}\left(y(t)\right) = y^T(t)\,\Xi\,y(t) \le \lambda_{\max}(\Xi)V\left(y(t)\right) \tag{38}$$

Integrating (38) from 0 to *t*, with $t \in [0, T]$, we have:

$$V\big(y(t)\big) < e^{\lambda_{\max}(\Xi)\cdot t} \cdot V(0) \tag{39}$$

From (8) , one can find:

$$\begin{aligned} V\big(y(0)\big) &= x^T(0)x(0) + 2\int_0^\tau x^T(0)Q(\vartheta)x(-\vartheta)d\vartheta \\ &+ \left[\int_0^\tau Q(\vartheta)x(-\vartheta)d\vartheta\right]^T \times \int_0^\tau Q(\vartheta)x(-\vartheta)d\vartheta \end{aligned} \tag{40}$$

Based on the known inequality (6), for $\Gamma = I$, one can get:

$$\begin{aligned} V\big(y(0)\big) &\le x^T(0)x(0) + \int_0^\tau x^T(0)Q(\vartheta)Q^T(\vartheta)x(0)d\vartheta \\ &+ \int_0^\tau x^T(-\vartheta)x(-\vartheta)\,d\vartheta + \left[\int_0^\tau Q(\vartheta)x(-\vartheta)d\vartheta\right]^T \times \int_0^\tau Q(\vartheta)x(-\vartheta)d\vartheta \end{aligned} \tag{41}$$

Using Jensen's inequality (4), we get:

$$\begin{aligned} V\big(y(0)\big) &\le x^T(0)x(0) + \int_0^\tau x^T(0)Q(\vartheta)Q^T(\vartheta)x(0)d\vartheta \\ &+ \int_0^\tau x^T(-\vartheta)x(-\vartheta)\,d\vartheta + \tau\int_0^\tau x^T(-\vartheta)Q^T(\vartheta)Q(\vartheta)x(-\vartheta)d\vartheta \end{aligned} \tag{42}$$

Introducing the general solution of (10), given with:

$$Q(\vartheta) = e^{R\vartheta}Q(0), \quad \vartheta \in [0, \tau], \quad R = A_0 + Q(0) \tag{43}$$

and by substituting (43) into (42), the following inequalities are obtained:

$$\begin{aligned} V\big(y(0)\big) &\le x^T(0)x(0) + \int_0^\tau x^T(0)e^{R\vartheta}Q(0)Q^T(0)e^{R^T\vartheta}x(0)d\vartheta \\ &+ \int_0^\tau x^T(-\vartheta)x(-\vartheta)\,d\vartheta + \tau\int_0^\tau x^T(-\vartheta)Q^T(0)e^{R^T\vartheta}e^{R\vartheta}Q(0)x(-\vartheta)d\vartheta \end{aligned} \tag{44}$$

$$\begin{aligned} V\left(y(0)\right) \leq {} & x^T(0)x(0) + \lambda_{\max}\left(Q(0)Q^T(0)\right)\int_0^{\tau}\lambda_{\max}\left(e^{R\vartheta}e^{R^T\vartheta}\right)x^T(0)x(0)\,d\vartheta \\ & + \int_0^{\tau}x^T(-\vartheta)x(-\vartheta)\,d\vartheta + \tau\int_0^{\tau}\lambda_{\max}\left(e^{R\vartheta}e^{R^T\vartheta}\right)x^T(-\vartheta)Q^T(0)Q(0)x(-\vartheta)\,d\vartheta \end{aligned} \tag{45}$$

$$\begin{aligned} V\left(y(0)\right) \leq {} & x^T(0)x(0) + x^T(0)x(0)\cdot\lambda_{\max}\left(Q(0)Q^T(0)\right)\int_0^{\tau}\lambda_{\max}\left(e^{R\vartheta}e^{R^T\vartheta}\right)d\vartheta \\ & + \int_0^{\tau}x^T(-\vartheta)x(-\vartheta)\,d\vartheta + \tau\,\lambda_{\max}\left(Q^T(0)Q(0)\right)\int_0^{\tau}\lambda_{\max}\left(e^{R\vartheta}e^{R^T\vartheta}\right)x^T(-\vartheta)\,d\vartheta \end{aligned} \tag{46}$$

Based on Definition 1, one can find:

$$\begin{aligned} V\left(y(0)\right) \leq {} & \alpha + \alpha\,\lambda_{\max}\left(Q(0)Q^T(0)\right)\int_0^{\tau}\lambda_{\max}\left(e^{R\vartheta}e^{R^T\vartheta}\right)d\vartheta \\ & + \alpha\,\tau + \alpha\,\tau\,\lambda_{\max}\left(Q^T(0)Q(0)\right)\int_0^{\tau}\lambda_{\max}\left(e^{R\vartheta}e^{R^T\vartheta}\right)d\vartheta \end{aligned} \tag{47}$$

From Coppell's inequality, Lemma 2, follows:

$$V\left(y(0)\right) \leq \alpha(1+\tau) + \alpha(1+\tau)\lambda_{\max}\left(Q(0)Q^T(0)\right)\int_0^{\tau}e^{2\mu(R)\vartheta}d\vartheta \tag{48}$$

or:

$$\begin{aligned} V\left(y(0)\right) & \leq \alpha\,(1+\tau)\left(1+\lambda_{\max}\left(Q(0)Q^T(0)\right)\left.\frac{e^{2\mu(R)\vartheta}}{2\mu(R)}\right|_{\vartheta=0}^{\vartheta=\tau}\right) \\ & = \alpha\,(1+\tau)\left(1+\lambda_{\max}\left(Q(0)Q^T(0)\right)\frac{e^{2\mu(R)\tau}-1}{2\mu(R)}\right) \end{aligned} \tag{49}$$

or finally:

$$V\left(y(0)\right) \leq \alpha(1+\tau)(1+\psi) \tag{50}$$

Based on (8)-(9), we have:

$$x^T(t)x(t) + 2\int_0^{\tau}x^T(t)Q(\eta)x(t-\eta)\,d\eta < V\left(y(t)\right) \tag{51}$$

or:

$$x^T(t)x(t) < V\big(y(t)\big) - 2\int_0^\tau x^T(t)Q(\eta)x(t-\eta)d\eta \tag{52}$$

Let us find the right second term in inequality (52). By using the inequality (7) for $\Gamma = pI > 0$ and by virtue of (31) and (43), one can find:

$$\begin{aligned}
&-2\int_0^\tau x^T(t)Q(\eta)x(t-\eta)d\eta \le \\
&\le \wp\int_0^\tau x^T(t)Q(\eta)Q^T(\eta)x(t)d\eta + \frac{1}{\wp}\int_0^\tau x^T(t-\eta)x(t-\eta)d\eta \\
&< \wp\int_0^\tau x^T(t)e^{R\eta}Q(0)Q^T(0)e^{R^T\eta}x(t)d\eta + \frac{q}{\wp}\int_0^\tau x^T(t)x(t)d\eta \\
&\le \wp\lambda_{\max}\big(Q(0)Q^T(0)\big)\int_0^\tau x^T(t)e^{R\eta}e^{R^T\eta}x(t)d\eta + \frac{q}{\wp}x^T(t)x(t)\int_0^\tau d\eta
\end{aligned} \tag{53}$$

or:

$$\begin{aligned}
&-2\int_0^\tau x^T(t)Q(\eta)x(t-\eta)d\eta \le \\
&\le \wp\lambda_{\max}\big(Q(0)Q^T(0)\big)x^T(t)x(t)\int_0^\tau \lambda_{\max}\Big(e^{R\eta}e^{R^T\eta}\Big)d\eta + \frac{q\tau}{\wp}x^T(t)x(t) \\
&< \wp\lambda_{\max}\big(Q(0)Q^T(0)\big)x^T(t)x(t)\int_0^\tau e^{2\mu(R)\eta}d\eta + \frac{q\tau}{\wp}x^T(t)x(t) \\
&= \wp\lambda_{\max}\big(Q(0)Q^T(0)\big)x^T(t)x(t)\left.\frac{e^{2\mu(R)\vartheta}}{2\mu(R)}\right|_{\eta=0}^{\eta=\tau} + \frac{q\tau}{\wp}x^T(t)x(t)
\end{aligned} \tag{54}$$

or:

$$\begin{aligned}
&-2\int_0^\tau x^T(t)Q(\eta)x(t-\eta)d\eta < \\
&< \wp\lambda_{\max}\big(Q(0)Q^T(0)\big)\frac{e^{2\mu(R)\tau}-1}{2\mu(R)}x^T(t)x(t) + \frac{q\tau}{\wp}x^T(t)x(t) \\
&= \left(\wp\psi + \frac{q\tau}{\wp}\right)x^T(t)x(t)
\end{aligned} \tag{55}$$

Thus, by using (39), (50), (52) and (55), leads to:

$$\begin{aligned} x^T(t)x(t) &< V(y(t)) + \left(\wp\psi + \frac{q\tau}{\wp}\right)x^T(t)x(t) \\ &\le e^{\lambda_{\max}(\Xi)t}V(y(0)) + \left(\wp\psi + \frac{q\tau}{\wp}\right)x^T(t)x(t) \\ &\le \alpha(1+\tau)(1+\psi)e^{\lambda_{\max}(\Xi)t} + \left(\wp\psi + \frac{q\tau}{\wp}\right)x^T(t)x(t) \end{aligned} \tag{56}$$

and:

$$\left(1 - \wp\psi - \frac{q\tau}{\wp}\right)x^T(t)x(t) < \alpha(1+\tau)(1+\psi)e^{\lambda_{\max}(\Xi)t}, \quad \forall t \in [0, T] \tag{57}$$

where:

$$1 - \wp\psi - \frac{q\tau}{\wp} > 0 \tag{58}$$

Obviously, if the condition (33) holds, then the inequality (58) is satisfied.

Finally, from the above inequality and (32), we get:

$$x^T(t)x(t) < \beta, \quad \forall t \in [0, T] \tag{59}$$

Remark 1. For the derived stability criteria, an existence of solution of the nonlinear algebraic matrix equation (37) is a necessary condition. In other words, the equation (37) must have at least one solution with respect to $Q(0)$, in order to Theorem 1 can be generally applied.

4 Numerical Example

Example 1. Given a system of the form:

$$\begin{aligned} &\dot{x}(t) = A_0x(t) + A_1x(t-0.1) \\ &\phi(t) = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}^T, \quad t \in [-0.1, 0] \\ &A_0 = \begin{bmatrix} -1.7 & 1.7 & 0 \\ 1.3 & -1 & 0.7 \\ 0.7 & 1 & -0.6 \end{bmatrix}, \quad A_1 = \begin{bmatrix} 1.5 & -1.7 & 0.1 \\ -1.3 & 1.5 & -0.3 \\ -0.7 & 1 & 0.1 \end{bmatrix} \end{aligned} \tag{60}$$

It is obvious that:

$$\phi^T(t)\phi(t) = 3 = \alpha, \quad t \in [-\tau, 0]$$

Figures 1-2 show the initial response $x(t)$ and the norm of state vector $x^T(t)x(t)$ of the system (60). Notice that the system (60) is not asymptotically stable. In addition, we determine upper bound of T such that the system (60) is FTS with respect $\{\alpha, \beta, T\}$.

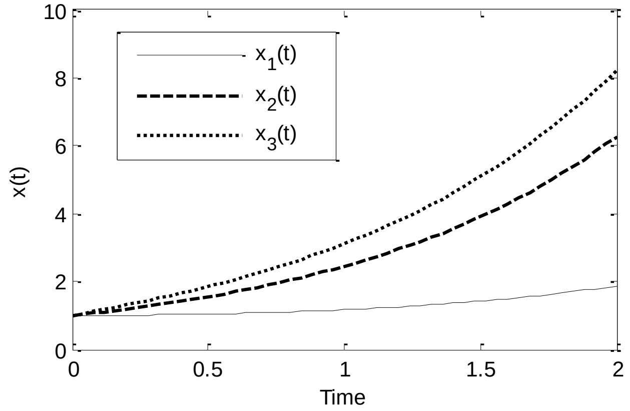


Figure 1
The state response $x(t)$ of the system (60)

Based on the initial response of the system (60), for following value of the parameter q can be adopted so (31) is valid: $q = 0.9$.

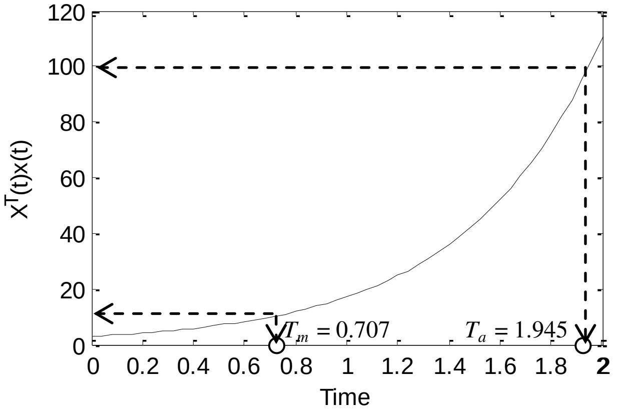


Figure 2
The norm $x^T(t)x(t)$ of the state vector of the system (60)

From (34) and (37) one can find:

$$Q(0)=\begin{bmatrix} 1.5279 & -1.7336 & 0.0994 \\ -1.2328 & 1.4145 & -0.2936 \\ -0.5249 & 0.8069 & 0.1575 \end{bmatrix}$$

$$R=\begin{bmatrix} -0.1721 & -0.0336 & 0.0994 \\ 0.0672 & 0.4145 & 0.4064 \\ 0.1751 & 1.8069 & -0.4425 \end{bmatrix},\ \Xi=\begin{bmatrix} -0.3442 & 0.0335 & 0.2745 \\ 0.0335 & 0.8290 & 2.2133 \\ 0.2745 & 2.2133 & -0.8849 \end{bmatrix}$$

so:

$$\lambda_{\max}\left\{Q(0)Q^T(0)\right\}=9.8109,\ \mu(R)=1.1789,\ \psi=1.1064,$$

$$\wp_1=0.1014,\ \wp_2=0.8025\ ,\ \wp\in(0.1014,\quad 0.8025),$$

$$4\psi\tau q=0.3983<1,\ \lambda_{\max}\{\Xi\}=2.3578.$$

Let $\beta\in\{100, 2000, 5000\}$ and find upper bound of T, T_m, so the system (60) is FTS. The results of the stability analysis, for different values of the parameter β, are shown in Table 1 using various methods: [17], [18], [21] and Theorem 1 (this paper). The actual values of parameter T, T_a , are estimated from the norm of state vector and shown in Table 1. Table 1 also lists the corresponding values of the parameter $\wp$.

Table 1
Upper bound of T , T_m

	$\beta=100$, $T_a=1.945$, see Fig. 2	$\beta=2000$, $T_a=3.525$	$\beta=5000$, $T_a=4.004$
[17]	0.585	1.085	1.238
[18]	0.448	0.842	0.962
[21] [22, without uncertainty]	1.225	2.517	2.939
Theorem 1	0.707, see Fig. 2 ($\wp=0.2865$)	1.978 ($\wp=0.2851$)	2.367 ($\wp=0.2837$)

From Table 1, it follows that Theorem 1 gives significantly better results than [17] and [18], but slightly poor results than [21] and [22]. However, unlike [21] and [22], which use LMI, Theorem 1 is based on algebraic inequalities, which can be solved without using appropriate optimization methods. Thus, compared to [21] and [22], the computational complexity of the presented stability criterion is significantly reduced.

Conclusion

This paper considers FTS of linear continuous time-delay systems. The combination of Lyapunov-like approach and two algebraic inequalities (Jensen's and Coppel's inequality) is used. The new sufficient, delay-dependent FTS criterion with algebraic inequality has been derived. The obtained result reduces the numerical computation.

Acknowledgment

This work was supported in part by the Ministry of Science and Technological Development of Serbia under the Project OI 174 001.

References

[1] Dorato P.: Short Time Stability in Linear Time-Varying System, Proc. IRE Internat. Conv. Rec. Part 4, New York, 1961, pp. 83-87

[2] Weiss L., Infante E. F.: Finite-Time Stability under Perturbing forces and on Product Spaces, IEEE Transaction on Automatic Control, Vol. 12, 1967, pp. 54-59

[3] Angelo H. D.: Linear Time-Varying Systems: Analysis and Synthesis, Allyn and Bacon, Boston, 1970

[4] Amato F., Ariola M., Dorato P.: State Feedback Stabilization Over a Finite-Time Interval of Linear Systems Subject to Norm Bounded Parametric Uncertainties, Proc. of the 36th Allerton Conference, Monticello, Sept. 23-25, 1998, pp. 499-505

[5] Amato F., Ariola M., Dorato P.: Finite-Time Control of Linear Systems Subject to Parametric Uncertainties and Disturbances, Automatica, Vol. 37 No. 9, 2001, pp. 1459-1463

[6] Amato F., Ariola M., Cosentino C., Abdallah C. T., Dorato P.: Necessary and Sufficient Conditions for Finite-Time Stability of Linear Systems, Proc. of American Control Conference, Denver, Colorado, June 2003, pp. 4452–4456

[7] Amato F., Ariola M., Dorato P.: Finite-Time Stabilization via Dynamic Output Feedback, Automatica, Vol. 42, 2006, pp. 337-342

[8] Moulay E., Perruquetti W.: Finite-Time Stability and Stabilization of a Class of Continuous Systems, Journal of Mathematical Analysis and Applications, Vol. 323, 2006, pp. 1430-1443

[9] Ming Q., Shen Y.: Finite-Time H∞ Control for Linear Continuous System with Norm-bounded Disturbance, Communications in Nonlinear Science and Numerical Simulation, Vol. 14, 2009, pp. 1043-1049

[10] G. Garcia, S. Tarbouriech, J. Bernussou, Finite-Time Stabilization of Linear Time-Varying Continuous Systems, IEEE Transaction on Automatic Control, Vol. 54, 2009, pp. 364-369

[11] Li P., Zheng Z.: Global Finite-Time Stabilization of Planar Nonlinear Systems with Disturbance, Asian Journal of Control, Vol. 14, No. 3, 2012, pp. 851-858

[12] Debeljkovic D. Lj., Nenadic Z. Lj., Koruga Dj., Milinkovic S. A., Jovanovic M. B.: On Practical Stability of Time-Delay Systems: New Results, Proc. 2nd ASCC97, Seoul, Korea, July 22-25, 1997, pp. 543-545

[13] Nenadic Z. Lj., Debeljkovic D. Lj., Milinkovic S. A.: On Practical Stability of Time Delay Systems, Proc. American Control Conference, Albuquerque, NM USA, 1997, pp. 3235-3235

[14] Lazarevic M. P., Debeljkovic D. Lj.: Finite Time Stability Analysis of Linear Autonomous Fractional Order Systems with Delayed State, Asian Journal of Control, Vol. 7, No. 4, 2005, pp. 440-447

[15] Lazarevic M. P., Debeljkovic D. Lj., Nenadic Z. Lj., Milinkovic S. A.: Finite-Time Stability of Delayed Systems, IMA Journal of Mathematical Control and Information, Vol. 17, No. 2, 2000, pp. 101-109

[16] Debeljkovic D. Lj., Lazarevic M. P., Koruga Dj., Milinkovic S. A., Jovanovic M. B.: Further Results on the Stability of Linear Nonautonomous Systems with Delayed State Defined Over Finite Time Interval, Proc. American Control Conference, Chicago, IL, USA, 2000, pp. 1450-1451

[17] Debeljkovic D. Lj., Buzurovic I.M., Nestorovic T., Popov D.: On Finite and Practical Stability of Time Delayed Systems: Lyapunov-Krassovski Approach: Delay Dependent Criteria, Proc. the 23rd IEEE Chinese Control and Decision Conference CCDC, Mianyang, China, March 23-25, 2011, pp. 331-337

[18] Debeljkovic D. Lj., Stojanovic S. B., Jovanovic A.: Further Results on Finite Time and Practical Stability of Continuous Time Delay Systems, *FME* Transactions, Vol. 41, 2013, pp. 241-249

[19] Stojanovic S. B., Debeljkovic D. LJ.: Delay-Dependent Stability Analysis for Discrete-Time Systems with Time Varying State Delay, Chemical Industry & Chemical Engineering Quarterly, Vol. 17 No. 4, 2011, pp. 497-504

[20] Stojanovic S. B., Debeljkovic D. LJ., Dimitrijevic N.: Finite-Time Stability of Discrete–Time Systems with Time-Varying Delay, Chemical Industry and Chemical Engineering Quarterly, Vol. 18, No 4/I, 2012, pp. 525-533

[21] Stojanovic S. B., Debeljkovic D. Lj., Antic D. S.: Finite Time Stability and Stabilization of Linear Time Delay Systems, Facta Universitatis, Series Automatic Control and Robotics, Vol. 11, No. 1, 2012, pp. 25-36

[22] Stojanovic S. B., Debeljkovic D. LJ., Antic D. S.: Robust Finite-Time Stability and Stabilization of Linear Uncertain Time-Delay Systems, Asian Journal of Control, Vol. 16, No. 2, 2013, DOI: 10.1002/asjc.689, pp. 1-7

[23] Lee T. N., Diant S.: Stability of Time-Delay Systems, IEEE Transactions on Automatic Control, Vol. 26, No. 4, 1981, pp. 951-953

New Vector Description of Kinetic Pressures on Shaft Bearings of a Rigid Body Nonlinear Dynamics with Coupled Rotations around No Intersecting Axes

Katica R. (Stevanović) Hedrih*, Ljljana Veljović**

*Mathematical Institute SANU, Belgrade, Serbia, and Faculty of Mechanical Engineering University of Niš, Serbia, E-mail: khedrih@eunet.rs

** Faculty of Mechanical Engineering University of Kragujevac, Serbia, E-mail: khedrih@sbb.rs

Abstract: New vector description of kinetic pressures on shaft bearings of a rigid body nonlinear dynamics with coupled rotations around no intersecting axes is first main result presented in this paper. Mass moment vectors and vector rotators coupled for pole and oriented axes, defined by K. Hedrih in 1991, are used for obtaining vector expressions for kinetic pressures on the shaft bearings of a rigid body dynamics with coupled rotations around no intersecting axes. A complete analysis of obtained vector expressions for kinetic pressures on shaft bearings give us a series of the kinematical vectors rotators around both directions determined by axes of the rigid body coupled rotations around no intersecting axes. As an example of defined dynamics, we take into consideration a heavy gyro-rotor-disk with one degree of freedom and coupled rotations when one component of rotation is programmed by constant angular velocity. For this system with nonlinear dynamics, series of graphical presentation transformations in realizations with changes of eccentricity and angle of inclination (skew position) of heavy rigid body-disk in relation to self rotation axis are presented, as well as in realization with changing orthogonal distance between axes of coupled rotations. Angular velocity of kinetic pressures components in vector form are expressed by using angular acceleration and angular velocity of component coupled rotations of gyrorotor-disk.

Keywords: coupled rotation; no intersecting axes; deviational mass moment vector; rotator; kinetic pressures; kinetic pressure components; nonlinear dynamics; gyrorotor-disk; eccentricity; angle of inclination, deviation kinetic couple; fixed point; graphical presentations; three parameter analysis

1 Introduction

No precisions and errors in the functions of gyroscopes caused by eccentricity and unbalanced gyro rotor body as well distance between axes of rotations are reason to investigate determined task as in the title of our paper.

The classical book [1] by Andonov, Vitt and Haykin contain a classical and very important elementary dynamical model of heavy mass particle relative motion along circle around vertical axis through it's center. Nonlinear dynamics and singularities lead to primitive model of the simple case of the gyro-rotor, which represent an useful dynamical and mathematical model of nonlinear dynamics.

Using K. Hedrih's (See Refs. [2-11]) mass moment vectors and vector rotators, some characteristics vector expressions of linear momentum and angular momentum and their derivatives for rigid body single rotation, were obtain physical and dynamical visible properties of the complex system dynamics and their kinetic parameters in vector form for single rotation. There are vector components of the shaft bearing kinetic pressures with opposite directions and same intensity that present deviational couple effect containing vector rotators, whose directions are same as kinetic pressure components on corresponding rotor shaft bearings (for detail see Refs. [2] and [5]).

The definitions of mass moment vectors coupled to the pole and the axis [2-9], [12] are introduced in the foundation of this vector method. The main vector is $\vec{\mathsf{J}}_{\vec{n}}^{(O)} \overset{def}{=} \iiint\limits_{V} [\vec{\rho},[\vec{n},\vec{\rho}]]dm$ of the body mass inertia moment at the point $A = O$ for the axis oriented by the unit vector and there is a corresponding vector $\vec{\mathsf{D}}_{\vec{n}}^{(O)}$ of the rigid body mass deviational moment for the axis through the point (see References [2] and [5-6]).

This vector approach is very suitable to obtain new view to the properties of dynamics of pure classical system dynamics investigated by numerous generations of the researchers and serious scientists around the world. We proof this approach in our published reference [12]. In Introduction of this paper [12] a short reviews of the basis of the subjective selected references about original research results of dynamics and stability of gyrostats was given. Then is reason that we didn't made any reviews of the papers about gyroscopes.

Passing through the content of the numerous published scientific paper, we can see that no results concerning behavior of the kinetic pressure directions and intensity depending of the nonlinear dynamic regimes. Then, our aim is to investigate kinetic pressures and deviation kinetic couple appearing to the shaft bearings of the rigid body coupled rotations around two no intersecting axes. Two our References [12] from (2008 and 2010) contain short presentation of the kinetic pressure to the gyrorotor self rotation bearings and rotators, as well as presentation of the nonlinear dynamics of the heavy gyro-rotor, but not completed.

This is reason that we take into a large consisderation and investigation three parameters analysis of the vector expressions of shaft bearing kinetic presures and their cmponents based on our previous results on applications vector method and published in our References [12]. This paper contan new rezults based on the previous our results.

Organizations of this paper based on the vector method applications with use of the mass moment vectors and vector rotators for describing vector expressions of kinetic pressures of the shaft bearings, of the rigid body coupled rotations around two no intersecting axes and corresponding kinetic deviation couple appearing by opposite kinetic pressures to shaft bearings and shaft bearing reactions.

Dynamics of a gyro-rotor with one degree of freedom and coupled rotations when one component of rotation is programmed by constant angular velocity is considered, as an example. For this system of nonlinear dynamics, the series of graphical presentation of the kinetic pressures of the shaft bearing of a rigid body self rotation are presented.

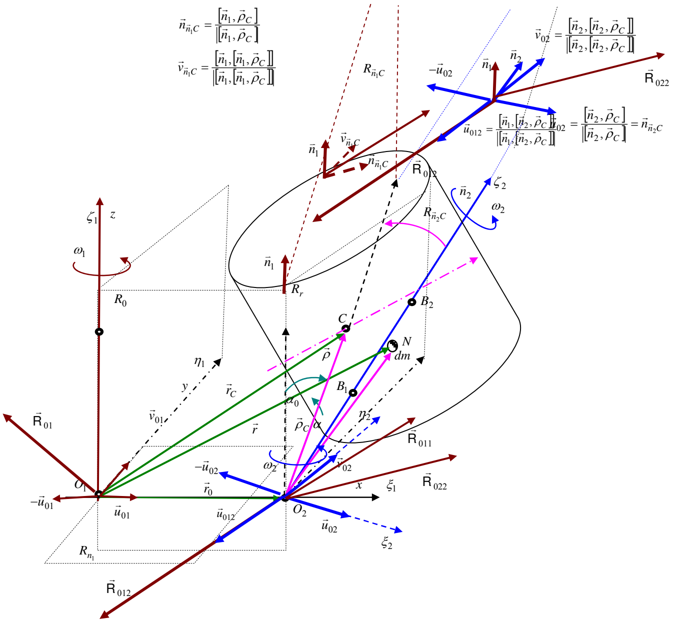


Figure 1

A rigid body coupled rotations around two no intersecting axes. System is with two degrees of mobility and one degrees of freedom, where φ_1 and φ_2 are rheonomic and generalized coordinates. Vector rotators $\vec{\mathsf{R}}_{01}$, $\vec{\mathsf{R}}_{011}$ and $\vec{\mathsf{R}}_{022}$ are presented.ssential connections

2 Model of a Rigid Body Rotation Around Two Axes without Intersection

Let us to consider rigid body coupled rotations around two no intersecting axes, presented in Figure 1. Ffirst axis is oriented by unit vector $\vec{n}_1$ with fixed position and second axis is oriented by unit vector $\vec{n}_2$ which is rotating around fixed axis with angular velocity $\vec{\omega}_1 = \omega_1 \vec{n}_1$. Axes of rotation are no intersecting axes. Rigid body is positioned on the moving rotating axis oriented by unit vector $\vec{n}_2$ and rotate around self rotating axis with angular velocity $\vec{\omega}_2 = \omega_2 \vec{n}_2$ and around fixed axis oriented by unit vector $\vec{n}_1$ with angular velocity $\vec{\omega}_1 = \omega_1 \vec{n}_1$. All geometrical parameters are presented in Fgure 1.

When any of three main central axes of rigid body mass inertia moments is not in direction of self rotation axis, then we can see that rigid body is scew positioned to the body self rotation axis. The angles of rigid body central main inertia axes inclinations acording self-lf rotation axis are β_i, $i = 1,2$. These angles are angles of scew position of rigid body to the body self rotation axis. When center C of the mass of rigid body is not on body self rotation axis of rigid body rotation, we can say that rigid body is scew positioned. Eccentricity of body position is normal distance between body mass center C and axis of self rotation and it is defined by $\vec{e} = [\vec{n}_2, [\vec{\rho}_C, \vec{n}_2]]$. Here $\vec{\rho}_C$ is vector position of mass center C with origin in point O$_2$, and position vector of mass center with fixed origin in point O$_1$ is $\vec{r}_C = \vec{r}_O + \vec{\rho}_C$.

3 Vector Equations of Dynamic Equlibrium of Rigid Body Coupled Rotations around Two No Intersecting Axes

By using theorems of linear momentum and angular momentum with respect to time, we can write two equations of dynamic equilibrium of the considered rigid body coupled rotations about two no intersecting axes, presented in Figure 1, in the following equations (for detail see Ref. [17] and Appendix):

$$\begin{aligned}\frac{d\vec{\mathsf{K}}}{dt} &= \vec{\mathsf{R}}_{01} \left\| [\vec{n}_1, \vec{r}_0] \right\| M + \vec{\mathsf{R}}_{011} \left| \vec{\mathsf{S}}_{\vec{n}_1}^{(O_2)} \right| + \vec{\mathsf{R}}_{022} \left| \vec{\mathsf{S}}_{\vec{n}_2}^{(O_2)} \right| + 2\omega_1\omega_2 \left[\vec{n}_1, \vec{\mathsf{S}}_{\vec{n}_2}^{(O_2)} \right] = \\ &= \vec{G} + \vec{F}_{AN1} + \vec{F}_{BN1} + \vec{F}_{Am} + \sum_{i=1}^{i=P} \vec{F}_i = \vec{G} + \vec{F}_{AN2} + \vec{F}_{BN2} + \vec{F}_{Am2} + \sum_{i=1}^{i=P} \vec{F}_i \end{aligned} \qquad (1)$$

$$\begin{aligned}\frac{d\vec{\mathsf{L}}_{O_1}}{dt} &= \vec{\chi}_{12}\left(\vec{r}_0, \vec{\rho}_C, M, \dot{\omega}_1, \dot{\omega}_2, \omega_1, \omega_2, \vec{n}_1, \vec{n}_2\right) + \dot{\omega}_1\vec{n}_1 r_0^2 M + 2\omega_1\omega_2\left[\vec{n}_1, \vec{\mathsf{J}}_{\vec{n}_2}^{(O_2)}\right] \\ &+ \dot{\omega}_1\left(\vec{n}_1, \vec{\mathsf{J}}_{\vec{n}_1}^{(O_2)},\right)\vec{n}_1 + \dot{\omega}_2\left(\vec{n}_2, \vec{\mathsf{J}}_{\vec{n}_2}^{(O_2)},\right)\vec{n}_2 + \vec{\mathsf{R}}_1\left|\vec{\mathsf{D}}_{\vec{n}_1}^{(O_2)}\right| + \vec{\mathsf{R}}_2\left|\vec{\mathsf{D}}_{\vec{n}_2}^{(O_2)}\right| = \\ &= \left[\vec{r}_0 + \vec{\rho}_{C,}\vec{G}\right] + \left[\vec{\rho}_{B1}, \vec{F}_{BN1}\right] + \sum_{i=1}^{i=P}\left[\vec{r}_0 + \vec{\rho}_{i,}, \vec{F}_i\right] = \left[\vec{r}_0 + \vec{\rho}_{C,}\vec{G}\right] + \\ &+ \left[\vec{r}_0 + \vec{\rho}_{A2}, \vec{F}_{AN2}\right] + \left[\vec{r}_0 + \vec{\rho}_{B2}, \vec{F}_{BN2}\right] + \left[\vec{r}_0 + \vec{\rho}_{A2}, \vec{F}_{Am2}\right] + \sum_{i=1}^{i=P}\left[\vec{r}_0 + \vec{\rho}_{i,}, \vec{F}_i\right]\end{aligned} \tag{2}$$

where $\vec{F}_i$, $i = 1,2,3...,P$ are active forces and $\vec{G}$ is weight of gyro rotor, $\vec{F}_{A1}$ and $\vec{F}_{B1}$ are reactive forces of fixed axis shaft bearing reactions and $\vec{F}_{A2}$ and $\vec{F}_{B2}$ are forces of self rotation shaft bearing reactions. From previous obtained vector equations, it is not difficult to obtain kinetic pressures to both shaft bearings, $\vec{F}_{A1}$ and $\vec{F}_{B1}$, as well as $\vec{F}_{A2}$ and $\vec{F}_{B2}$ on both shafts bearings as well as two differential equations along φ_1 and φ_2 of the rigid body coupled rotations about two no intersecting axes, and to obtain time solutions of unknown generalized coordinate φ_1 and φ_2, or if we know these coordinate to find unknown external active forces.

For the case that axes are perpendicular some terms in previous vector expressions and vector equations are equal to zero, but these equations are nonlinear along angle coordinates φ_1 and φ_2, and coupled by generalized coordinates, φ_1 and φ_2, and their derivatives, and also, by forces of shaft bearings reactions.

Two vector equates (1) and (2) are valid for rigid body coupled rotations around no intersecting axes, as well for the case intersecting axes as its special case. Also, these equations are valid for the system dynamics with two degrees of mobility, and for three different cases.

4 Vector Rotators of Rigid Body Coupled Rotations around Two No Intersecting Axes

We can see that in previous vector equations (1) and (2) terms for derivative of linear momentum and angular momentum contain two sets of the vector rotators:

$$\vec{\mathsf{R}}_{01} = \dot{\omega}_1\left[\vec{n}_1, \frac{\vec{r}_0}{r_0}\right] + \omega_1^2\left[\vec{n}_1, \left[\vec{n}_1, \frac{\vec{r}_0}{r_0}\right]\right], \qquad \vec{\mathsf{R}}_{011} = \dot{\omega}_1\frac{\vec{\mathsf{S}}_{\vec{n}_1}^{(O_2)}}{\left|\vec{\mathsf{S}}_{\vec{n}_1}^{(O_2)}\right|} + \omega_1^2\left[\vec{n}_1, \frac{\vec{\mathsf{S}}_{\vec{n}_1}^{(O_2)}}{\left|\vec{\mathsf{S}}_{\vec{n}_1}^{(O_2)}\right|}\right] \tag{3}$$

$$\left|\vec{\mathsf{R}}_{012}\right| = \mathsf{R}_{012} = 2\omega_1\omega_2 = 2\omega_1\omega_2 \tag{4}$$

$$\vec{\mathsf{R}}_{022} = \dot{\omega}_2 \frac{\vec{\mathsf{S}}^{(O_2)}_{\vec{n}_2}}{\left|\vec{\mathsf{S}}^{(O_2)}_{\vec{n}_2}\right|} + \omega_2^2 \left[\vec{n}_2, \frac{\vec{\mathsf{S}}^{(O_2)}_{\vec{n}_2}}{\left|\vec{\mathsf{S}}^{(O_2)}_{\vec{n}_2}\right|}\right], \quad \vec{\mathsf{R}}_{012} = 2\omega_1\omega_2 \frac{\left[\vec{n}_1, \vec{\mathsf{S}}^{(O_2)}_{\vec{n}_2}\right]}{\left|\left[\vec{n}_1, \vec{\mathsf{S}}^{(O_2)}_{\vec{n}_2}\right]\right|} \tag{5}$$

First two vector rotators $\vec{\mathsf{R}}_{01}$ and $\vec{\mathsf{R}}_{011}$ are orthogonal to the direction of the first fixed axis and third vector rotator is orthogonal to the self rotation axis. But, first vector rotator $\vec{\mathsf{R}}_{01}$ is coupled for pole O_1 on the fixed axis and second and third vector rotators, $\vec{\mathsf{R}}_{011}$ and $\vec{\mathsf{R}}_{022}$, are coupled for the pole O_2 on self rotation axis and for corresponding direction oriented by directions of component angular velocities of coupled rotations. Intensities of two first rotators are equal and are expressed by angular velocity and angular acceleration of the first component rotation, and intensity of third vector rotators is expressed by angular velocity and angular acceleration of the second component rotation, and they are in the following forms: $\mathsf{R}_{01} = \mathsf{R}_{011} = \sqrt{\dot{\omega}_1^2 + \omega_1^4}$ and $\mathsf{R}_{022} = \sqrt{\dot{\omega}_2^2 + \omega_2^4}$.

Lets introduce notation $\gamma_{01}, \gamma_{011}$ and γ_{022} denote difference between corresponding component angles of rotation φ_1 and φ_2 of the rigid body component rotations and corresponding absolute angles of rotation of pure kinematics vector rotators about axes oriented by unit vectors $\vec{n}_1$ and $\vec{n}_2$. These angular velocities of relative kinematics vectors rotators $\vec{\mathsf{R}}_{01}, \vec{\mathsf{R}}_{011}$ and $\vec{\mathsf{R}}_{022}$ which rotate about corresponding axis in relation to the component angular velocities of the rigid body component rotations are:

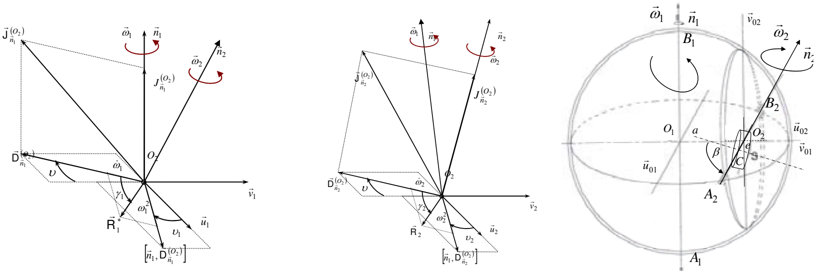
$$\dot{\gamma}_{01} = \dot{\gamma}_{011} = \frac{\dot{\varphi}_1\left(2\ddot{\varphi}_1 - \dot{\varphi}_1\dddot{\varphi}_1\right)}{\ddot{\varphi}_1^2 + \dot{\varphi}_1^4} \quad \text{and} \quad \dot{\gamma}_{02} = \frac{\dot{\varphi}_2\left(2\ddot{\varphi}_2 - \dot{\varphi}_2\dddot{\varphi}_2\right)}{\ddot{\varphi}_2^2 + \dot{\varphi}_2^4} \tag{6}$$

In Figure 1 Vector rotators $\vec{\mathsf{R}}_{01}$, $\vec{\mathsf{R}}_{011}$ and $\vec{\mathsf{R}}_{022}$ are presented.

We can see that in previous vector expression (2) for derivative of angular momentum are introduced vector rotators in the following vector form:

$$\vec{\mathsf{R}}_1 = \dot{\omega}_1 \frac{\vec{\mathsf{D}}^{(O_2)}_{\vec{n}_1}}{\left|\vec{\mathsf{D}}^{(O_2)}_{\vec{n}_1}\right|} + \omega_1^2 \left[\vec{n}_1, \frac{\vec{\mathsf{D}}^{(O_2)}_{\vec{n}_1}}{\left|\vec{\mathsf{D}}^{(O_2)}_{\vec{n}_1}\right|}\right], \qquad \vec{\mathsf{R}}_2 = \dot{\omega}_2 \frac{\vec{\mathsf{D}}^{(O_2)}_{\vec{n}_2}}{\left|\vec{\mathsf{D}}^{(O_2)}_{\vec{n}_2}\right|} + \omega_2^2 \left[\vec{n}_2, \frac{\vec{\mathsf{D}}^{(O_2)}_{\vec{n}_2}}{\left|\vec{\mathsf{D}}^{(O_2)}_{\vec{n}_2}\right|}\right]$$

$$\vec{\mathsf{R}}_{12} = 2\omega_1\omega_2 \frac{\left[\vec{n}_1, \vec{\mathsf{J}}^{(O_2)}_{\vec{n}_2}\right]}{\left|\left[\vec{n}_1, \vec{\mathsf{J}}^{(O_2)}_{\vec{n}_2}\right]\right|} = 2\omega_1\omega_2\vec{u}_{12} \tag{7}$$



a* b* c*

Figure 2

Vector rotators $\vec{R}_1$ (a*) and $\vec{R}_2$ (b*) in relations to corresponding mass moment vectors $\vec{J}^{(O_2)}_{\vec{n}_1}$ and $\vec{J}^{(O_2)}_{\vec{n}_2}$, and their corresponding deviational components $\vec{D}^{(O_2)}_{\vec{n}_1}$ and $\vec{D}^{(O_2)}_{\vec{n}_2}$ as well as to corresponding deviational planes. (c*) Model of heavy gyro rotor with two component coupled rotations around orthogonal axes without intersections

The first $\vec{R}_1$ is orthogonal to the fixed axis oriented by unit vector $\vec{n}_1$ and second $\vec{R}_2$ is orthogonal to the self rotation axis oriented by unit vector $\vec{n}_2$. Intensity of first rotator $\vec{R}_1$ is equal to intensity of previous defined rotator R_{01} and intensity of second rotator $\vec{R}_2$ is equal to intensity of previous defined rotator R_{022} defined by expressions (7). In Figure 2 vector rotators $\vec{R}_1$ (in Figure 2 a*) and $\vec{R}_2$ (in Figure 2.b*) in relations to corresponding mass moment vectors $\vec{J}^{(O_2)}_{\vec{n}_1}$ and $\vec{J}^{(O_2)}_{\vec{n}_2}$, and their corresponding deviational components $\vec{D}^{(O_2)}_{\vec{n}_1}$ and $\vec{D}^{(O_2)}_{\vec{n}_2}$ as well as to corresponding deviational planes are presented. Vector rotators $\vec{R}_1$ and $\vec{R}_2$ are pure kinematical vectors first presented in reference [18,19] as a function on angular velocity and angular accelerationin a form $\vec{R} = \ddot{\varphi}\vec{u} + \dot{\varphi}^2\vec{w} = R\vec{R}_0$.

Rotators from first set are rotated around through pole O_2 and axis in direction of first component rotation angular velocity and depend of angular velocity ω_1 and angular acceleration $\dot{\omega}_1$. There are two vectors of such type and all trees have equal intensity. Rotators from second set are rotated around axis in direction of second component rotation and depend of angular velocity ω_2 and angular acceleration $\dot{\omega}_2$. There are two vectors of such type and they have equal intensity.

Lets introduce notation γ_1, and γ_2 denote difference between corresponding component angles of rotation φ_1 and φ_2 of the rigid body component rotations and corresponding absolute angles of pure kinematics vector rotators about axes oriented by unit vectors $\vec{n}_1$ and $\vec{n}_2$ through pole O_2. These angular velocity of relative kinematics vectors rotators $\vec{\mathsf{R}}_1$ and $\vec{\mathsf{R}}_2$ which rotate about axes in corresponding directions in relation to the component angular velocities of the rigid body component rotations through pole O_2 are expressed as $\dot{\gamma}_1 = \dot{\gamma}_{01} = \dot{\gamma}_{011}$ and $\dot{\gamma}_2 = \dot{\gamma}_{02}$.

5 Vector Expressions of Kinetic Pressures (Kinetic Reactions) on Shaft Bearings of Rigid Body Coupled Rotations around Two No Intersecting Axes

Kinetic pressures (bearing reactions with out parts reactions induced by external forces) on fixed shaft bearings for the case that spherical bearing is at the pole O_1 and cylindrical in this fixed axis defined by vector position $\vec{\rho}_{B1} = \rho_{B1}\vec{n}_1$ are in the following form:

$$\vec{F}_{AN1} = -\vec{F}_{BN1} + \vec{\mathsf{R}}_{011}\left|\vec{\mathsf{S}}_{\vec{n}_1}^{(O_2)}\right| + \vec{\mathsf{R}}_{022}\left|\vec{\mathsf{S}}_{\vec{n}_2}^{(O_2)}\right| + 2\omega_1\omega_2\left[\vec{n}_1, \vec{\mathsf{S}}_{\vec{n}_2}^{(O_2)}\right] \tag{8}$$

$$\begin{aligned}\vec{F}_{BN1} = {} & \frac{1}{\rho_{B1}}\left[\vec{\mathsf{R}}_1, \vec{n}_{1,}\right]\left|\vec{\mathsf{D}}_{\vec{n}_1}^{(O_2)}\right| + \frac{1}{\rho_B}\left[\vec{\mathsf{R}}_2, \vec{n}_{1,}\right]\left|\vec{\mathsf{D}}_{\vec{n}_2}^{(O_2)}\right| + \frac{2\omega_1\omega_2}{\rho_B}\left[\left[\vec{n}_1, \vec{\mathsf{J}}_{\vec{n}_2}^{(O_2)}, \vec{n}_{1,}\right]\right] \\ & + \frac{1}{\rho_{B1}}\left[\vec{\chi}_{12}\left(\vec{r}_0, \vec{\rho}_C, M, \dot{\omega}_1, \dot{\omega}_2, \omega_1, \omega_2, \vec{n}_1, \vec{n}_2\right), \vec{n}_{1,}\right] + \frac{\dot{\omega}_2}{\rho_B}\left(\vec{n}_2, \vec{\mathsf{J}}_{\vec{n}_2}^{(O_2)},\right)\left[\vec{n}_2, \vec{n}_{1,}\right] = -\vec{F}_{BN1}^{dev}\end{aligned} \tag{9}$$

It is not difficult, by use system decomposition, to obtain kinetic pressures on body self rotation shaft bearings for the case that spherical bearing is at the pole O_2.

By analysis vector equations (1) and (2) and corresponding expressions (8) and (9) for kinetic pressures on the both shafts bearings, we can conclude that in the system to the both shaft bearings appear in the pair of bearings two opposite components of kinetic pressures with deviation couple. In fixed shaft bearings A_1 and B_1 appear the following opposite components: $\vec{F}_{BN1}$ and $\vec{F}_{BN1}^{dev}$ in vector relation: $\vec{F}_{BN1} = -\vec{F}_{BN1}^{dev}$, but in different points of appearance, bearings A_1 and B_1 with distance $\vec{\rho}_{B1}$ and build one couple, $\vec{\mathsf{M}}_{dev1} = \left[\vec{\rho}_{B1}, \vec{F}_{BN1}\right] = -\left[\vec{\rho}_{B1}, \vec{F}_{AN1}^{dev}\right]$, known under the name deviation couple, and identified in like our investigated system dynamics, for which we obtain the following vector expression:

$$\begin{aligned}\vec{\mathsf{M}}_{dev\,1} = \vec{\mathsf{R}}_1\left|\vec{\mathsf{D}}_{\vec{n}_1}^{(O_2)}\right| + \left[\vec{n}_1, \left[\vec{\mathsf{R}}_2, \vec{n}_1\right]\right]\left|\vec{\mathsf{D}}_{\vec{n}_2}^{(O_2)}\right| + 2\omega_1\omega_2\left[\vec{n}_1, \vec{\mathsf{J}}_{\vec{n}_2}^{(O_2)}\right] + \\ + \left[\vec{n}_1, \left[\vec{\chi}_{12}\left(\vec{r}_0, \vec{\rho}_C, M, \dot{\omega}_1, \dot{\omega}_2, \omega_1, \omega_2, \vec{n}_1, \vec{n}_2\right), \vec{n}_1\right]\right]\end{aligned} \qquad (10)$$

Also, it is possible to conclude for two opposite components of kinetic pressures to the self rotation shaft bearings $\vec{F}_{BN2}$ and $\vec{F}_2^{dev}$ in vector relation: $\vec{F}_{BN2} = -\vec{F}_{BN2}^{dev}$, but in different points of appearance, bearings A_2 and B_2 with distance $\vec{\rho}_{B2}$ and build one couple, $\vec{\mathsf{M}}_{dev\,2} = \left[\vec{\rho}_{B2}, \vec{F}_{BN\,2}\right] = -\left[\vec{\rho}_{B2}, \vec{F}_{AN\,2}^{dev}\right]$, known under the name deviation couple, and also identified in like our investigated system dynamics.

6 Dynamic of Rigid Body Coupled Rotations around Two Orthogonal No Intersecting Axes and with One Degree of Freedom

We are going to take into consideration special case of the considered heavy rigid body with coupled rotations about two axes without intersection with one degree of freedom, and in the gravitation field. For this case generalized coordinate φ_2 is independent, and coordinate φ_1 is programmed. In that case, we say that coordinate φ_1 is rheonomic coordinate and system is with kinematical excitation, programmed by forced support rotation by constant angular velocity. When the angular velocity of shaft support axis is constant, $\dot{\varphi}_1 = \omega_1 = const$, we have that rheonomic coordinate is linear function of time, $\varphi_1 = \omega_1 t + \varphi_{10}$, and angular acceleration around fixed axis is equal to zero $\dot{\omega}_1 = 0$. Special case is when the support shaft axis is vertical and the gyro-rotor shaft axis is horizontal, and all time in horizontal plane, and when axes are no intersecting at normal distance a. So we are going to consider that example presented in Figure 2c*. The normal distance between axes is a. The angle of self rotation around moveable self rotation axis oriented by the unit vector $\vec{n}_2$ is φ_2 and the angular velocity is $\omega_2 = \dot{\varphi}_2$. The angle of rotation around the shaft support axis oriented by the unit vector $\vec{n}_1$ is φ_1 and the angular velocity is $\omega_1 = constat$. The angular velocity of rotor is $\vec{\omega} = \omega_1\vec{n}_1 + \omega_2\vec{n}_2 = \dot{\varphi}_1\vec{n}_1 + \dot{\varphi}_2\vec{n}_2$. The angle φ_2 is generalized coordinates in case when, we investigate system with one degrees of freedom, but system have two degrees of mobility. Also, without loose of generality, we take that rigid body is a disk, eccentrically positioned on the self rotation shaft axis with eccentricity e, and that angle of skew inclined position between one of main axes of disk and self rotation axis is β, as it is visible in Figure 2c*.

For that example, differential equation of the heavy gyro rotor-disk self rotation of reviewed model in Figure 2 for the case coupled rotations about two orthogonal no intersecting axes by using (2), after multiplying scalar by $\vec{n}_2$, and taking into account orthogonal between axes of coupled rotations, we can obtain in the following form:

$$\ddot{\varphi}_2 + \Omega^2(\lambda - \cos\varphi_2)\sin\varphi_2 + \Omega^2\psi\cos\varphi_2 = 0 \qquad (11)$$

where

$$\Omega^2 = \omega_1^2 \frac{J^{(C)}_{\vec{u}_2} - J^{(C)}_{\vec{v}_2}}{J^{(C)}_{\vec{n}_2}},\ \lambda = \frac{mge\sin\beta}{\omega_1^2\left(J^{(C)}_{\vec{u}_2} - J^{(C)}_{\vec{v}_2}\right)},\ \psi = \frac{2mea\sin\beta}{J^{(C)}_{\vec{u}_2} - J^{(C)}_{\vec{v}_2}},\ \varepsilon = 1 + 4\left(\frac{e}{r}\right)^2 \qquad (12)$$

Here it is considered an eccentric disc (eccentricity is e), with mass m and radius r, which is inclined to the axis of its own self rotation by the angle β (see Figure 4), so that previous constants (12) in differential equation (11) become the following forms:

$$\Omega^2 = \omega_1^2 \frac{\left(\varepsilon\sin^2\beta - 1\right)}{\left(\varepsilon\sin^2\beta + 1\right)},\ \varepsilon = 1 + 4\left(\frac{e}{r}\right)^2,\ \lambda = \frac{g(\varepsilon - 1)\sin\beta}{e\omega_1^2\left(\varepsilon\sin^2\beta - 1\right)},\ \psi = \frac{2ea\sin\beta}{er\left(\varepsilon\sin^2\beta - 1\right)} \qquad (13)$$

Relative nonlinear dynamics of the heavy gyro-rotor-disk around self rotation shaft axis is possible to present by means of phase portrait method. Forms of phase trajectories and their transformations by changes of initial conditions, and for different cases of disk eccentricity and angle of its skew position, as well as for different values of orthogonal distance between axes of component rotations may present character of nonlinear oscillations.

For that reason it is necessary to find first integral of the differential equation (11). After integration of the differential equation (26), the non-linear equation of the phase trajectories of the heavy gyro rotor disk dynamics with the initial conditions $t_0 = 0,\ \varphi_1(t_0) = \varphi_{10},\ \dot{\varphi}_1(t_0) = \dot{\varphi}_{10}$, we obtain in the following form:

$$\dot{\varphi}_2^2 = \dot{\varphi}_{02}^2 + 2\Omega^2\left(\lambda\cos\varphi_2 - \frac{1}{2}\cos^2\varphi_2 + \psi\sin\varphi_2\right) - 2\Omega^2\left(\lambda\cos\varphi_{02} - \frac{1}{2}\cos^2\varphi_{02} + \psi\sin\varphi_{02}\right) \qquad (14)$$

The analyzed system is conservative and equation (14) is the energy integral.

In considered case for the heavy gyro-rotor-disk nonlinear dynamics in the gravitational field with one degree of freedom and with constant angular velocity about fixed axis, we have three sets of vector rotators.

Three of these vector rotators $\vec{\mathsf{R}}_{01}$, $\vec{\mathsf{R}}_{011}$ and $\vec{\mathsf{R}}_{1}$, from first set, are with same constant intensity $\left|\vec{\mathsf{R}}_{01}\right| = \left|\vec{\mathsf{R}}_{011}\right| = \left|\vec{\mathsf{R}}_{1}\right| = \omega_1^2 = cons\tan t$ and rotate with constant angular velocity ω_1 and equal to the angular velocity of rigid body component precession rotation about fixed axis, but two of these three vector rotators, $\vec{\mathsf{R}}_{011}$ and $\vec{\mathsf{R}}_{1}$ are connected to the pole O_2 on the self rotation axis, and are orthogonal to the axis parallel direction as direction of the fixed axis. All these three vector rotators $\vec{\mathsf{R}}_{01}$, $\vec{\mathsf{R}}_{011}$ and $\vec{\mathsf{R}}_{1}$ are in different directions (see Figures 1 and 3). Two of these vector rotators, $\vec{\mathsf{R}}_{022}$ and $\vec{\mathsf{R}}_{2}$, from second set, are with same intensity equal to $\mathsf{R}_{022} = \sqrt{\dot{\omega}_2^2 + \omega_2^4}$, and connecter to the pole O_2 and orthogonal to the self rotation axis oriented by unit vector $\vec{n}_2$ and rotate about this axis with relative angular velocity $\dot{\gamma}_2$ defined by expression (6), in respect to the self rotation angular velocity ω_2 (see Figures 2 a*, b* and c*).

By use expressions (3-5) and (7), we can list following series of vector rotators of the gyro-rotor–disk with coupled rotation around orthogonal no intersecting axes and with $\omega_1 = const$:

$$\vec{\mathsf{R}}_{01} = \omega_1^2 \vec{v}_{01}, \quad \left|\vec{\mathsf{R}}_{01}\right| = \omega_1^2, \quad \vec{\mathsf{R}}_{011} = -\omega_1^2 \frac{\left\langle \sin\widetilde{\beta} \sin\varphi_2 \vec{v}_{01} + \cos\widetilde{\beta} \vec{u}_{01} \right\rangle}{\sqrt{\cos^2\widetilde{\beta} + \sin^2\widetilde{\beta} \sin^2\varphi_2}}, \quad \left|\vec{\mathsf{R}}_{011}\right| = \omega_1^2$$

$$\vec{\mathsf{R}}_{022} = \dot{\omega}_2 \vec{v}_{02} - \omega_1^2 \vec{u}_{02}, \quad \left|\vec{\mathsf{R}}_{022}\right| = \sqrt{\dot{\omega}_2^2 + \omega_2^4}, \quad \vec{\mathsf{R}}_{012} = -2\omega_1\omega_2 \vec{u}_{01}, \tag{15}$$

$$\left|\vec{\mathsf{R}}_{012}\right| = \mathsf{R}_{012} = 2\omega_1\omega_2, \quad \vec{\mathsf{R}}_{1} = -\omega_1^2 \frac{\vec{u}_{01}\left\langle \sin\widetilde{\beta}\cos\varphi_2 \right\rangle + \vec{v}_{01}\left\langle \cos\widetilde{\beta} \right\rangle}{\sqrt{\cos^2\beta + \sin^2\beta\cos^2\varphi_2}}, \quad \left|\vec{\mathsf{R}}_{1}\right| = \omega_1^2$$

$$\vec{\mathsf{R}}_{2} = \dot{\omega}_2 \vec{v}_{02} - \omega_2^2 \vec{u}_{02}, \quad \left|\vec{\mathsf{R}}_{2}\right| = \sqrt{\dot{\omega}_2^2 + \omega_2^4}, \quad \vec{\mathsf{R}}_{12} = 2\omega_1\omega_2 \frac{\left[\vec{n}_1, \vec{\mathsf{J}}_{\vec{n}_2}^{(O_2)}\right]}{\left|\left[\vec{n}_1, \vec{\mathsf{J}}_{\vec{n}_2}^{(O_2)}\right]\right|} = 2\omega_1\omega_2 \vec{u}_{12},$$

$$\left|\vec{\mathsf{R}}_{12}\right| = \mathsf{R}_{12} = 2\omega_1\omega_2$$

in which $\widetilde{\beta}$ is angle between relative vector position $\vec{\rho}_C$ of rigid body mass center C and self rotation axis oriented by unit vector $\vec{n}_2$. One of the vectors rotators from the third set is $\vec{\mathsf{R}}_{012}$ with intensity $\left|\vec{\mathsf{R}}_{012}\right| = 2\omega_1\omega_2$ and direction:

$\vec{\mathsf{R}}_{012} = -2\omega_1\omega_2\vec{u}_{01}$. This vector rotator is connecter to the pole O_2 and orthogonal to the axis oriented by unit vector $\vec{n}_1$ and relative rotate about this axis. Intensity of this vector rotator expressed by generalized coordinate φ_2, angle of self rotation of heavy disk, taking into account first integral (29) of the differential equation (26) obtain the following form:

$$\left|\vec{\mathsf{R}}_{012}\right| = 2\omega_1\sqrt{\dot{\varphi}_{02}^2 + 2\Omega^2\left(\lambda\cos\varphi_2 - \frac{1}{2}\cos^2\varphi_2 + \psi\sin\varphi_2\right) - 2\Omega^2\left(\lambda\cos\varphi_{02} - \frac{1}{2}\cos^2\varphi_{02} + \psi\sin\varphi_{02}\right)} \tag{16}$$

Intensity R_{022} of two of these vector rotators, $\vec{\mathsf{R}}_{022}$ and $\vec{\mathsf{R}}_2$, from second set, depends on angular velocity ω_2 and angular acceleration $\dot{\omega}_2$. For the considered system of the heavy gyro-rotor-disk dynamics, for obtaining expressions of intensities of vector rotators, $\vec{\mathsf{R}}_{022}$ and $\vec{\mathsf{R}}_2$, from second set, in the function of the generalized coordinate φ_2, angle of self rotation of heavy disk self rotation, we take into account a first integral (14) of nonlinear differential equation (11), and by using these result and previous expressions (15) of vector rotator we can write:

*intensities of the vectors rotators, $\vec{\mathsf{R}}_{022}$ and $\vec{\mathsf{R}}_2$, connected for the pole O_2 and rotate around self rotation axis, in the following form:

$$\begin{aligned}\left|\vec{\mathsf{R}}_{022}\right| &= \left|\vec{\mathsf{R}}_{022}(\varphi_2)\right| = \\ &= \Omega^2\sqrt{\left[-(\lambda-\cos\varphi_2)\sin\varphi_2 + \psi\cos\varphi_2\right]^2 + \left[\dot{\varphi}_{02}^2 + 2\Omega^2\left(\lambda\cos\varphi_2 - \frac{1}{2}\cos^2\varphi_2 + \psi\sin\varphi_2\right) - 2\Omega^2\left(\lambda\cos\varphi_{02} - \frac{1}{2}\cos^2\varphi_{02} + \psi\sin\varphi_{02}\right)\right]^2}\end{aligned} \tag{17}$$

*vector rotators orthogonal to the self rotation axes are in the following vector forms:

$$\begin{aligned}\vec{\mathsf{R}}_{022}(\varphi_2) &= \Omega^2\left[-(\lambda-\cos\varphi_2)\sin\varphi_2 + \psi\cos\varphi_2\right]\frac{[\vec{n}_2,\vec{\rho}_C]}{\left|[\vec{n}_2,\vec{\rho}_C]\right|} + \\ &+ \Omega^2\left[\dot{\varphi}_{02}^2 + 2\Omega^2\left(\lambda\cos\varphi_2 - \frac{1}{2}\cos^2\varphi_2 + \psi\sin\varphi_2\right) - 2\Omega^2\left(\lambda\cos\varphi_{02} - \frac{1}{2}\cos^2\varphi_{02} + \psi\sin\varphi_{02}\right)\right]\frac{[\vec{n}_2,[\vec{n}_2,\vec{\rho}_C]]}{\left|[\vec{n}_2,\vec{\rho}_C]\right|}\end{aligned} \tag{18}$$

$$\begin{aligned}\vec{\mathsf{R}}_2(\varphi_2) &= \Omega^2\left[-(\lambda-\cos\varphi_2)\sin\varphi_2 + \psi\cos\varphi_2\right]\frac{\vec{\mathsf{D}}_{\vec{n}_2}^{(O_2)}}{\left|\vec{\mathsf{D}}_{\vec{n}_2}^{(O_2)}\right|} + \\ &+ \Omega^2\left[\dot{\varphi}_{02}^2 + 2\Omega^2\left(\lambda\cos\varphi_2 - \frac{1}{2}\cos^2\varphi_2 + \psi\sin\varphi_2\right) - 2\Omega^2\left(\lambda\cos\varphi_{02} - \frac{1}{2}\cos^2\varphi_{02} + \psi\sin\varphi_{02}\right)\right]\left[\vec{n}_2,\frac{\vec{\mathsf{D}}_{\vec{n}_2}^{(O_2)}}{\left|\vec{\mathsf{D}}_{\vec{n}_2}^{(O_2)}\right|}\right]\end{aligned}$$

7 Kinetic Pressures to Shaft Bearings of Rigid Body Coupled Rotations around Two Orthogonal No Intersecting Axes and with One Degree of Freedom

By use previous derived vector equations (1) and (2) and approach to obtaining vector expressions (8) and (9) for kinetic pressures, $\vec{F}_{A1}$ and $\vec{F}_{B1}$, to fixed shaft bearings of rigid body coupled rotations around two no intersecting orthogonal axes and for system with one degree of freedom, it is easy to obtain vector expressions for kinetic pressures $\vec{F}_{A2}$ and $\vec{F}_{B2}$ (including component reactions of the rigid body weight) to self rotation shaft bearings, A_2 and B_2, of rigid body coupled rotations around two orthogonal no intersecting axes and for considered particular example in the following form:

$$\vec{F}_{A\vec{n}_2}=\left[\left|\vec{\mathsf{S}}^{(O)}_{\vec{n}_1}\right|\left(\vec{\mathsf{R}}_1,\vec{n}_2\right)+\left|\vec{\mathsf{S}}^{(O)}_{\vec{n}_2}\right|\left(\vec{\mathsf{R}}_{21},\vec{n}_2\right)-\left(\vec{G},\vec{n}_2\right)\right]\vec{n}_2 \tag{19}$$

$$\begin{aligned}\vec{F}_{B2}=&\frac{1}{2}\left\{\left|\vec{\mathsf{S}}^{(O_2)}_{\vec{n}_2}\right|\left[\vec{n}_2,\left[\vec{\mathsf{R}}_2,\vec{n}_2\right]\right]+\left|\vec{\mathsf{S}}^{(O_2)}_{\vec{n}_1}\right|\left[\vec{n}_2,\left[\vec{\mathsf{R}}_1,\vec{n}_2\right]\right]+\left|\vec{\mathsf{S}}^{(O_2)}_{\vec{n}_2}\right|\left[\vec{n}_2,\left[\vec{\mathsf{R}}_{21},\vec{n}_2\right]\right]-\left[\vec{n}_2,\left[\vec{G},\vec{n}_2\right]\right]\right\}+\\&+\frac{1}{2\ell}\left\{\left|\vec{\mathsf{D}}^{(O_2)}_{\vec{n}_2}\right|\left[\vec{\mathsf{R}}^{*}_{02},\vec{n}_2\right]+\left|\vec{\mathsf{D}}^{(O_2)}_{\vec{n}_1}\right|\left[\vec{\mathsf{R}}^{*}_{02},\vec{n}_2\right]+\left[\vec{n}_1,\vec{n}_2\right]\dot{\omega}_1\left(\vec{n}_1,\vec{\mathsf{J}}^{(O_2)}_{\vec{n}_1}\right)-\left[\left[\vec{\rho}_C,\vec{G}\right],\vec{n}_2\right]\right\}+\\&+\frac{1}{2\ell}\left[\left[\left\{\left[\vec{n}_2,\vec{\mathsf{J}}^{(O_2)}_{\vec{n}1}\right]+\left[\vec{n}_1,\vec{\mathsf{J}}^{(O_2)}_{\vec{n}2}\right]+\mathsf{J}^{(O_2)}\left[\vec{n}_1,\vec{n}_2\right]\right\},\vec{n}_2\right],\vec{n}_1\right]\end{aligned} \tag{20}$$

$$\begin{aligned}\vec{F}_{AT2}=&\frac{1}{2}\left\{\left|\vec{\mathsf{S}}^{(O_2)}_{\vec{n}_2}\right|\left[\vec{n}_2,\left[\vec{\mathsf{R}}_2,\vec{n}_2\right]\right]+\left|\vec{\mathsf{S}}^{(O_2)}_{\vec{n}_1}\right|\left[\vec{n}_2,\left[\vec{\mathsf{R}}_1,\vec{n}_2\right]\right]+\left|\vec{\mathsf{S}}^{(O_2)}_{\vec{n}_2}\right|\left[\vec{n}_2,\left[\vec{\mathsf{R}}_{21},\vec{n}_2\right]\right]-\left[\vec{n}_2,\left[\vec{G},\vec{n}_2\right]\right]\right\}-\\&-\frac{1}{2\ell}\left\{\left|\vec{\mathsf{D}}^{(O_2)}_{\vec{n}_2}\right|\left[\vec{\mathsf{R}}^{*}_{02},\vec{n}_2\right]+\left|\vec{\mathsf{D}}^{(O_2)}_{\vec{n}_1}\right|\left[\vec{\mathsf{R}}^{*}_{01},\vec{n}_2\right]+\left[\vec{n}_1,\vec{n}_2\right]\dot{\omega}_1\left(\vec{n}_1,\vec{\mathsf{J}}^{(O_2)}_{\vec{n}_1}\right)-\left[\left[\vec{\rho}_C,\vec{G}\right],\vec{n}_2\right]\right\}-\\&-\frac{1}{2\ell}\left[\left[\left\{\left[\vec{n}_2,\vec{\mathsf{J}}^{(O_2)}_{\vec{n}1}\right]+\left[\vec{n}_1,\vec{\mathsf{J}}^{(O_2)}_{\vec{n}21}\right]+\mathsf{J}^{(O_2)}\left[\vec{n}_1,\vec{n}_2\right]\right\},\vec{n}_2\right],\vec{n}_1\right]\end{aligned} \tag{21}$$

where $\mathsf{J}^{(O_2)}$ is matrix of tensor of mass inertia moments for pole O_2. Previous expressions contain member which correspond to the bearing reactions of the rotor proper weight. After taking into account mass inertia moment vector for inclined disk and disk position with eccentricity of mass body center, we can write in scalar form components of kinetic pressures, $\vec{F}_{A2}$ and $\vec{F}_{B2}$ (including component reactions of the rigid body weight) to on bearings, A_2 and B_2, of the self rotation axis in the following form:

$$\begin{aligned}F_{Au2}=&\frac{1}{2}m(e\ddot{\varphi}_2\sin\beta-O_1O_2\omega_1^2\cos\varphi_2-e\omega_1^2\sin\beta\sin\varphi_2\cos\varphi_2)+\frac{1}{2}mg\left(1+\frac{e}{\ell}\cos\beta\right)\sin\varphi_2+\\&+\frac{1}{2\ell}\Big((J_u-J_v-J_n)\dot{\varphi}_2\omega_1+J_{vn}\omega_1^2\cos\varphi_2)\sin\varphi_2+J_{vn}\ddot{\varphi}_2\Big)-m\frac{e}{\ell}O_1O_2\omega_1^2\cos\beta\cos\varphi_2\end{aligned} \tag{22}$$

$$F_{Av2}=\frac{1}{2}m\left(e\dot{\varphi}_2^2\sin\beta+O_1O_2\omega_1^2\sin\varphi_2+e\omega_1^2\sin\beta\sin\varphi_2\sin\varphi_2\right)+\frac{1}{2}mg\left(1-\frac{e}{\ell}\cos\beta\right)\cos\varphi_{12}-$$

$$-\frac{1}{2\ell}\left((J_u-J_v+J_n)\omega_1\dot{\varphi}_2-J_{vn}\omega_1^2\cos\varphi_2)\cos\varphi_2+J_{vn}\dot{\varphi}_2^2\right)-m\frac{e}{\ell}O_1O_2\omega_1^2\cos\beta\sin\varphi_2 \quad (23)$$

$$F_{Bu2}=\frac{1}{2}m(e\ddot{\varphi}_2\sin\beta-O_1O_2\omega_1^2\cos\varphi_2-e\omega_1^2\sin\beta\sin\varphi_2\cos\varphi_2)+\frac{1}{2}mg\left(1-\frac{e}{\ell}\cos\beta\right)\sin\varphi_2-$$

$$-\frac{1}{2\ell}\left((J_u-J_v-J_n)\dot{\varphi}_2\omega_1+J_{vn}\omega_1^2\cos\varphi_2)\sin\varphi_2+J_{vn}\ddot{\varphi}_2\right)-m\frac{e}{\ell}O_1O_2\omega_1^2\cos\beta\cos\varphi_{12} \quad (24)$$

$$F_{Bv2}=\frac{1}{2}m\left(e\dot{\varphi}_2^2\sin\beta+O_1O_2\omega_1^2\sin\varphi_2+e\omega_1^2\sin\beta\sin\varphi_2\sin\varphi_2\right)+\frac{1}{2}mg\left(1-\frac{e}{\ell}\cos\beta\right)\cos\varphi_2+$$

$$+\frac{1}{2\ell}\left((J_u-J_v+J_n)\omega_1\dot{\varphi}_2-J_{vn}\omega_1^2\cos\varphi_2)\cos\varphi_2+J_{vn}\dot{\varphi}_2^2\right)-m\frac{e}{\ell}O_1O_2\omega_1^2\cos\beta\sin\varphi_2 \quad (25)$$

Previous obtained expressions (22)-(25) of the components of kinetic pressures, $\vec{F}_{A2}$ and $\vec{F}_{B2}$ (including component reactions of the rigid body weight) to bearings, A_2 and B_2 , of the self rotation axis in scalar form , is possible present in the following vector form:

$$\begin{aligned}\vec{F}_{B2}^{kin}&=F_{Bu2}^{kin}\vec{u}_2+F_{Bv2}^{kin}\vec{v}_2=(\frac{1}{2}me\sin\beta-\frac{1}{2\ell}J_{vn})(\ddot{\varphi}\vec{u}_2+\dot{\varphi}^2\vec{v}_2)+\\&+\frac{1}{2\ell}\left(J_{vn}\omega_1\cos\varphi_2-J_n\dot{\varphi}_2\right)\omega_1\left(\sin\varphi_2\vec{u}+\cos\varphi_2\vec{v}\right)-\\&-\frac{1}{2}m\omega_1^2(a+e\sin\beta\cos\varphi_2+2e\frac{a}{\ell}\cos\beta)\left(\cos\varphi_2\vec{u}-\sin\varphi_2\vec{v}\right)+\\&+\frac{1}{2\ell}\left(J_u-J_v\right)\omega_1\dot{\varphi}_2\left(\sin\varphi_2\vec{u}-\cos\varphi_2\vec{v}\right)\end{aligned} \quad (26)$$

$$\begin{aligned}\vec{F}_{A2}^{kin}&=F_{Au2}^{kin}\vec{u}_2+F_{Av2}^{kin}\vec{v}_2=(\frac{1}{2}me\sin\beta+\frac{1}{2\ell}J_{vn})(\ddot{\varphi}_2\vec{u}_2+\dot{\varphi}_2^2\vec{v}_2)+\\&+\frac{1}{2\ell}\left(J_{vn}\omega_2\cos\varphi_2+J_n\dot{\varphi}_2\right)\omega_1\left(\sin\varphi_2\vec{u}+\cos\varphi_2\vec{v}\right)-\\&-\frac{1}{2}m\omega_1^2(a+e\sin\beta\cos\varphi_2-2e\frac{a}{\ell}\cos\beta)\left(\cos\varphi_2\vec{u}-\sin\varphi_2\vec{v}\right)-\\&-\frac{1}{2\ell}\left(J_u-J_v\right)\omega_1\dot{\varphi}_2\left(\sin\varphi_2\vec{u}-\cos\varphi_2\vec{v}\right)\end{aligned} \quad (27)$$

Previous scalar expressions are suitable for analysis on the basis decompositions to the separate components with specific properties of intensity, directions, influence of some mass and geometrical properties and structure parameters, as well as angular velocities and other kinetic parameters of considered special example.

By introducing the following unit vectors $\vec{w}_1$, $\vec{w}_2$ and $\vec{w}_3$

$$\vec{w}_1 = \vec{u}_2 \sin\varphi_2 + \vec{v}_2 \cos\varphi_2\,,\ \vec{w}_2 = \vec{u}_2 \sin\varphi_2 - \vec{v}_2 \cos\varphi_2\,, \vec{w}_3 = -\vec{u}_2 \cos\varphi_2 + \vec{v}_2 \sin\varphi_2$$

pressures $\vec{F}_{Ag2}$ and $\vec{F}_{Bg2}$ (reactions of the rigid body weight) to bearings A_2 and B_2, of the self rotation axis is possible to express in the following vector form:

$$\vec{F}_{Ag2} = F_{Aug2}\vec{u}_2 + F_{Agv2}\vec{v}_2 = \frac{1}{2}mg\left(1+\frac{e}{\ell}\cos\beta\right)\sin\varphi_2\vec{u}_2 + \frac{1}{2}mg\left(1-\frac{e}{\ell}\cos\beta\right)\cos\varphi_2\vec{v}_2$$

$$\vec{F}_{Bg2} = F_{Bgu2}\vec{u}_2 + F_{Bgv2}\vec{v}_2 = \frac{1}{2}mg\left(1-\frac{e}{\ell}\cos\beta\right)\sin\varphi_2\vec{u}_2 + \frac{1}{2}mg\left(1-\frac{e}{\ell}\cos\beta\right)\cos\varphi_2\vec{v}_2 \tag{28}$$

From last forms of the pressures, $\vec{F}_{Ag2}$ and $\vec{F}_{Bg2}$ (reactions of the rigid body weight) to bearings, A_2 and B_2, of the self rotation axis, we can see that is possible to separate component with same intensity, and opposite directions, and also component with same angular velocity in one or in other directions.

In Figure 3 some of the introduced unit vectors $\vec{u}_2$, $\vec{v}_2$, $\vec{w}_1$, $\vec{w}_2$ and $\vec{w}_3$ for analysis kinetic pressures $\vec{F}_{B2}$ (including component reactions of the rigid body weight) to bearings, A_2 and B_2, of the self rotation axis used in expressions (28) schematically are presented with corresponding angular velocity and directions of rotations.

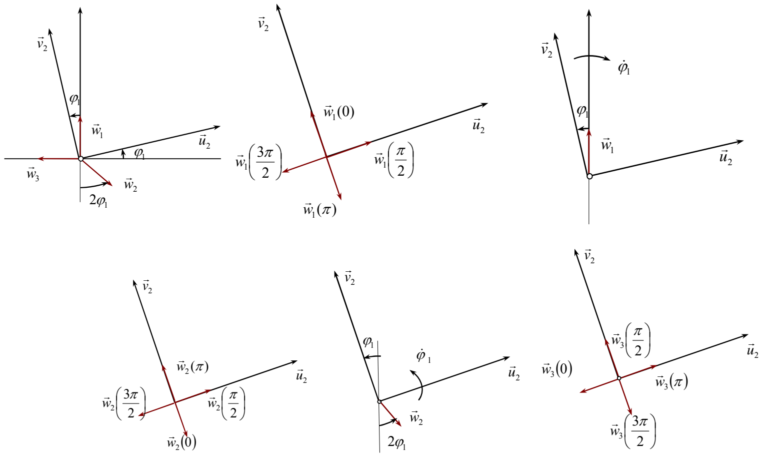


Figure 3

Schematically presentation of the unit vectors $\vec{u}_2$, $\vec{v}_2$, $\vec{w}_1$, $\vec{w}_2$ and $\vec{w}_3$, and their geometrical and kinematical relations with corresponding angular velocity and directions of rotations

Components F_{B21}^{kin}, F_{B22}^{kin}, F_{B23}^{kin} and F_{B24}^{kin} of pure kinetic pressure $\vec{F}_{B2}^{kin}$ to bearing B_2, of the self rotation axis are in the following forms:

$$F_{B21}^{kin} = \frac{1}{2\ell}\left(J_{vn}\omega_1\cos\varphi_2 - J_n\dot{\varphi}_2\right)\omega_1\ , \qquad F_{B22}^{kin} = \frac{1}{2\ell}\left(J_u - J_v\right)\omega_1\dot{\varphi}_2$$

$$F_{B23}^{kin} = \frac{1}{2}m\omega_1^2\left(a + e\sin\beta\cos\varphi_2 + 2e\frac{a}{\ell}\cos\beta\right), \qquad F_{B24}^{kin} = \frac{1}{2}me\sin\beta - \frac{1}{2\ell}J_{vn}$$

From previous expressions for components F_{B21}^{kin}, F_{B22}^{kin}, F_{B23}^{kin} and F_{B24}^{kin} of pure kinetic pressure $\vec{F}_{B2}^{kin}$ to bearing B_2, of the self rotation axis, we can conclude, that influence of disk position eccentricity is stronger to the components F_{B23}^{kin} of pure kinetic pressure $\vec{F}_{B2}^{kin}$, and that intensity of component F_{B22}^{kin} increase, and intensity of the component F_{B21}^{kin} decrease with increasing of disk eccentricity. Intensity of the pure kinetic pressure $\vec{F}_{B2}^{kin}$ increase with increasing of disk eccentricity.

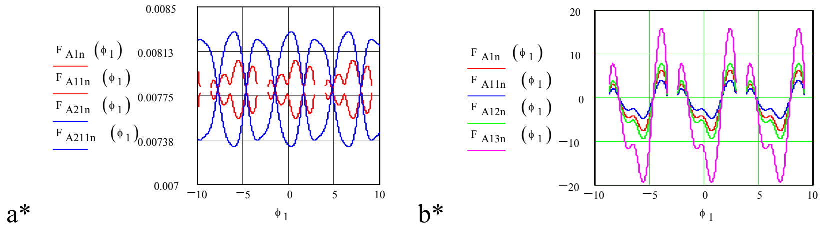


Figure 4

Intensity transformation of kinetic pressure component F_{A2n2}^{kin} to self rotation shaft spherical bearing A_2 of rigid body coupled rotations around two orthogonal no intersecting axes and for system with one degree of freedom, in direction of the self rotation shaft axis for different disk eccentricity

8 Graphical Presentation of Kinetic Pressures to Self Rotation Shaft Bearings of Rigid Body Coupled Rotations

By use previous listed expressions as well as other no listed heir, and MathCad as a software tool, a numerical experiment was followed for analysis properties of the kinetic pressures and their corresponding components to the both shaft bearings. Selected graphical presentation is done in the Figures 4-10. All graphical presentation are obtained by analytical expressions derived in previous chapters of this paper.

In Figure 4a* and b(graphical presentation of intensity transformation of kinetic pressure component F^{kin}_{A2n2} to self rotation shaft spherical bearing A_2 of rigid body coupled rotations around two orthogonal no intersecting axes and for system with one degree of freedom, in direction of the self rotation shaft axis and in function of self rotation relative angle φ_2, for different disk eccentricity, is presented.

In Figure 5 graphical presentation of intensity transformation of kinetic pressure component F^{kin}_{A2N2} to self rotation shaft spherical bearing A_2 of rigid body coupled rotations around two orthogonal no intersecting axes and for system with one degree of freedom, in orthogonal direction to the self rotation shaft axis, in function of self rotation relative angle φ_2, for different disk eccentricity, is presented.

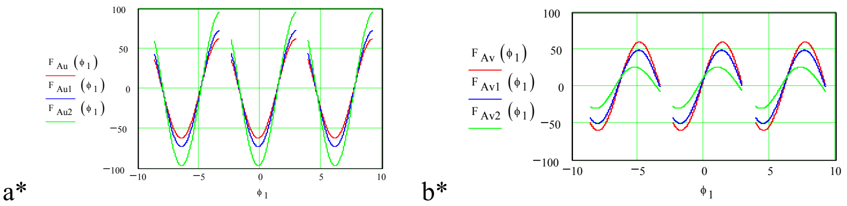


Figure 5

a* and b* Intensity of kinetic pressure component F^{kin}_{A2N2} to self rotation shaft spherical bearing A_2 of rigid body coupled rotations around two orthogonal no intersecting axes and for system with one degree of freedom, in orthogonal direction to the self rotation shaft axis, for different value of disk eccentricity

In Figure 6 (a*), (c*) and (d*) the intensity of kinetic pressure component of F^{kin}_{B2N2} to self rotation cylindrical bearing B_2 of rigid body coupled rotations around two orthogonal no intersecting axes in direction of $\vec{\mathrm{R}}_2$ and for system with one degree of freedom, in orthogonal direction to the self rotation shaft axis, for different value of disk angle β skew position is presented. In Figure 8 (b*) Intensity of the vector rotator $\vec{\mathrm{R}}_2$ in function of the value of disk angle β skew positions is presented.

In Figure 7 graphical presentation of intensity transformation of kinetic pressure component F^{kin}_{B2N2} to self rotation shaft cylindrical bearing B_2 of rigid body coupled rotations around two orthogonal no intersecting axes and for system with one degree of freedom, in orthogonal direction to the self rotation shaft axis, in function of self rotation relative angle φ_2, for different disk eccentricity, is presented.

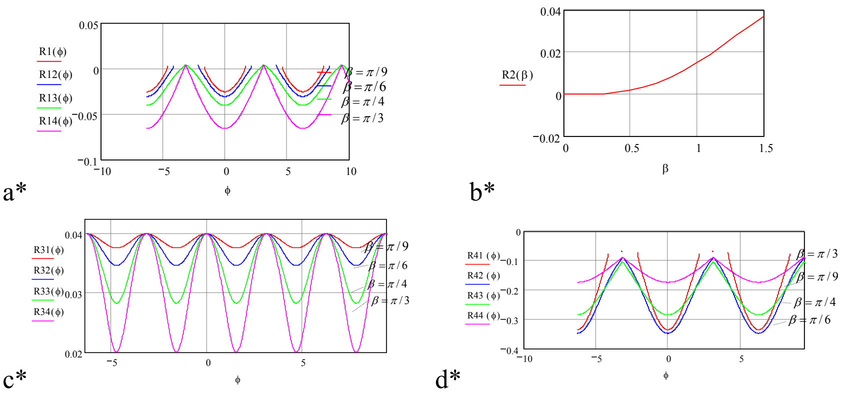


Figure 6

(a*), (c*) and (d*) Intensity of kinetic pressure component of F_{B2N2}^{kin} to self rotation cylindrical bearing B_2 of rigid body coupled rotations around two orthogonal no intersecting axes in direction of $\vec{R}_2$ and for system with one degree of freedom, in orthogonal direction to the self rotation shaft axis, for different value of disk angle β skew position..(b*) Intensity of the vector rotator $\vec{R}_2$ in function of the value of disk angle β skew position

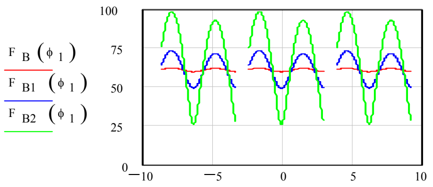


Figure 7

Intensity of kinetic pressure component F_{B2N2}^{kin} to self rotation shaft cylindrical bearing B_2 of rigid body coupled rotations around two orthogonal no intersecting axes and for system with one degree of freedom, in orthogonal direction to the self rotation shaft axis, for different value of disk eccentricity

In Figure 8 graphical presentation of intensity transformation of kinetic pressure component F_{B2N2}^{kin} to self rotation shaft cylindrical bearing B_2 of rigid body coupled rotations around two orthogonal no intersecting axes and for system with one degree of freedom, in orthogonal direction to the self rotation shaft axis, in function of self rotation relative angle φ_2, for different disk eccentricity, is presented. In Figure 9 intensities of kinetic pressure deviation couple to self rotation shaft bearings of rigid body coupled rotations around two orthogonal no intersecting axes and for system with one degree of freedom, in orthogonal direction to the self rotation shaft axis, for different value of disk eccentricity are presented.

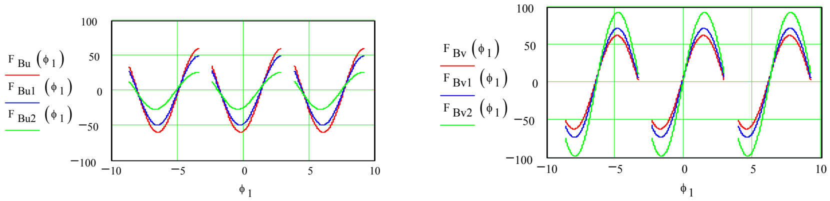


Figure 8

Intensity of kinetic pressure component F^{kin}_{B2N2} to self rotation shaft cylindrical bearing B_2 of rigid body coupled rotations around two orthogonal axes without intersection and for system with one degree of freedom, in orthogonal direction to the self rotation shaft axis, for different value of disk eccentricity.

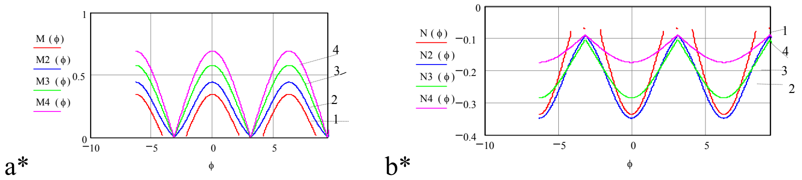


Figure 9

Intensity of kinetic pressure deviation couple to self rotation shaft bearings of rigid body coupled rotations around two orthogonal no intersecting axes and for system with one degree of freedom, in orthogonal direction to the self rotation shaft axis, for different value of disk eccentricity.

Concluding remarks

Complexity of the single rigid body motion with coupled rotations about no intersecting axes by vector method based on the mass moment vectors and vector rotators coupled for pole on selfrotation axis and component angular velocity axes is presented by sampler vector expressions them usually scalar forms in professional books in this area. New approach and new composition of this vector method open new way for applications to the multi-body system dynamics with coupled multi-rotations about nonintersecting axes. New vector expressions for linear momentum and angular momentum and their derivatives of the single rigid body complex motion by coupled rotations about nonintersecting axes expressed by new introduced mass moments vectors and their very elegant form open new possibility for generalizations these expressions for describing multi rigid body system complex motion by coupled multi-rotations about higher numbers of nonintersecting axes large present in many real mechanical engineering systems and robotic system dynamics with coupled multi-rotations.

Acknowledgement

Parts of this research were supported by the Ministry of Sciences and Technology of Republic of Serbia through Mathematical Institute SANU Belgrade Grant ON174001 Dynamics of hybrid systems with complex structures. Mechanics of materials and Faculty of Mechanical Engineering University of Niš and Faculty of Mechanical Engineering University of Kragujevac.

References

[1] Andonov, A. A., Vitt, A. A., Haykin, S. E., (1981), *Teoriya kolebaniy*, Nauka, Moskva, p. 568

[2] Hedrih (Stevanović) K., (1992), *On Some Interpretations of the Rigid Bodies Kinetic Parameters,* XVIIIth ICTAM HAIFA, Apstracts, pp. 73-74

[3] Hedrih (Stevanović) K., (1993), Same *Vectorial Interpretations of the Kinetic Parameters of Solid Material Lines*, ZAMM. Angew.Math. Mech. 73(1993) 4-5, T153-T156

[4] Hedrih (Stevanović) K., (1993), *The Mass Moment Vectors at n-dimensional Coordinate System*, Tensor, Japan, Vol 54 (1993), pp. 83-87

[5] Hedrih (Stevanović) K., (2001), V*ector Method of the Heavy Rotor Kinetic Parameter Analysis and Nonlinear Dynamics*, *University of Niš 2001,* Monograph, p. 252 (in English), YU ISBN 86-7181-046-1

[6] Hedrih (Stevanović) K., (1998), *Vectors of the Body Mass Moments*, Monograph paper, Topics from Mathematics and Mechanics, Mathematical institute SANU, Belgrade, Zbornik radova 8(16), 1998, pp. 45-104, Published in 1999 (in English), (Zentralblatt Review)

[7] Hedrih (Stevanović) K., (1998), *Derivatives of the Mass Moments Vectors with Applications*, Invited Lecture, Proceedings, 5th National Congress on Mechanics, Ioannina, 1998, pp. 694-705

[8] Hedrih (Stevanović) K., (1994), *Interpretation of the Motion of a Heavy Body around a Stationary Axis in the Field with Turbulent Damping and Kinetic Pressures on Bearing by Means of the Mass Moment Vector for the Pole and the Axis*, Facta Universitatis Series Mechanics, Automatic Control and Robotics, Vol. 1, No. 4, 1994, pp. 519-538

[9] Hedrih (Stevanović) K., (1998), *Vector Method of the Kinetic Parameters Analysis of the Rotor with Many Axes and Nonlinear Dynamics,* Parallel General Lecture, Third International Conference on Nonlinear Mechanics (ICNM III), August 17-20, 1998, Shanghai, China, pp. 42-47

[10] Hedrih (Stevanović) K., (2004), *Contribution to the Coupled Rotor Nonlinear Dynamics,* Advances in nonlinear Sciences, Monograph, Belgrade, Academy of Nonlinear Sciences, 229-259. ISBN 86-905633-0-X UDC 530-18299(082) 51-73:53(082) UKUP. STR. 261

[11] Hedrih (Stevanović) K., *Interpretation of the Motion Equations of a Variable Mass Object Rotating around a Atationary Axis by Means of the Mass Moment Vector for the Pole and the Axis*, Procedings of the 4th Greek National Congress on Mechanics, Vol. 1, Mechanics of Solids, Democritus University of Trace, Xanthi, 1995, pp. 690-696

[12] Hedrih (Stevanovic) K. and Veljovic Lj., Vector Rotators of Rigid Body Dynamics with Coupled Rotations around Axes without Intersection, Hindawi Publishing Corporation, Mathematical Problems in Engineering, Volume 2011, Article ID 351269, 26 pages, doi:10.1155/2011/351269

A New Method for Determining the Reliability Testing Period Using Weibull Distribution

Cristin Olimpiu Morariu, Sebastian Marian Zaharia

Transilvania University of Brasov,
Faculty of Technological Engineering and Industrial Management,
Department of Manufacturing Engineering
Colina Universitatii Nr. 1, 500036, Brasov, Romania
E-mail: zaharia_sebastian@unitbv.ro; c.morariu@unitbv.ro

Abstract: In this paper, we present a calculation methodology of the testing duration of the products' reliability, using the Weibull distribution, which allows the estimation of the mean duration of a censored and/or complete test, as well as of the confidence intervals for this duration. By using these values we can improve the adequate planning and allocation of material and human resources for the specific testing activities. The proposed methodology and the results' accuracy were verified using the Monte Carlo data simulation method.

Keywords: reliability; test plan; Weibull distribution; Monte Carlo simulation

1 Introduction

The reliability theory is a technological discipline closely related to the probability theory and mathematical statistics [1, 2, 3]. The data regarding the reliability of the products are obtained mainly through the following three methods: following the behavior of the products in real operation; during the laboratory tests; by using the data simulation through the Monte Carlo method. During the laboratory tests, we tried to emulate, as much as possible, the conditions in real operation, by reproducing the range of internal stresses, as well as the environmental stresses. The most important laboratory tests are the reliability tests [4, 5].

1.1 Background on Reliability Test

The reliability tests have a great importance, aiming either to determine, either to check the reliability characteristic of a product, if this is established in a predictive way. The reliability tests are extremely necessary and they have a decisive role in improving the technical solutions and in increasing the performances. The essential problem of reliability tests is the testing duration, which is generally comparable with the product's useful life time [6, 7, 8, 9].

The most used reliability tests are the following [7, 10, 11, 12, 13]:

- Complete tests (type *n* out of *n*) - in these tests *n* products of the same kind, the experiment being considered finished when all of the *n* products have failed.
- Censored tests (type *r* out of *n*) - are commonly used and they consist of subjection to testing of *n* products of the same type, the experiment being considered finished after the failure of $r<n$ tested products; obviously, the *r* number is previously determined, usually by technical, economical and statistical considerations [2, 14, 15].
- Truncated tests (with a fixed testing time) - a *n* number of products are subjected to testing, but the experiment doesn't stop according to the number of failed elements, but according to a $_{tr}$ time, previously set, a period during which the testing takes place.

The testing methodology about to be used has a direct economical impact, because in every test the following terms intervene: the cost of the tested product; the total cost of the experiment; the time consumed for testing and for the statistical processing of the data resulted from testing. Therefore, the selection of a specific type of test is a managerial decision that has to be taken by a responsible authority. Also, in the follow-up of the results of the statistical processing, we will propose certain corrective technical and economical actions, aimed directly at the quality and reliability of the product in question [3, 6, 16, 17, 18, 19, 20].

In order to realize reliability tests we must take into account the following aspects:

a. A previously determined n number of products that will be subjected to testing.

b. A testing plan that includes the following aspects: the selection of stress parameters, which will determine the failure mechanisms specific to the product.

c. Instructions regarding the adequate type of test and the methods of calculation, in order to estimate the reliability parameters.

d. A test chart where the experimental data are recorded and the statistical calculations, as well as the chronological recording of observations and interventions, are made.

e. Testing stands, testing equipment, auxiliary materials and qualified staff in order to realize the test.

1.2 Nomenclature

W(t,β,η) - the Weibull distribution, having the function of distribution:

$$F(t)=1-e^{-\left(\frac{t}{\eta}\right)^{\beta}};$$

β - the shape parameter of Weibull distribution;

η - the scale parameter of Weibull distribution;

B(x, n, p) - the binomial distribution, having the function of probability:

$$\Pr(X=x)=C_n^x\cdot p^x\cdot(1-p)^{n-x};$$

N - the number of simulations;

n - the sample volume;

r - the level of censorship;

$F_n\,(t_i)$ - the empirical function of distribution correspondent to the operation time until failure;

$T_{n/n}$ - the duration of a complete test realized on a volume sample *n*;

$T_{r/n}$ - the duration of a censored test al level *r*, realized on volume *n* sample;

1-α - confidence level;

α - significance level;

$Q_{p,\,v1,\,v2}$ - the p quantile of the Fisher - Snedecor, with v_1 and v_2 degrees of freedom;

U=1-α/2 - the index used to note the superior confidence limit; represents the value of the probability corresponding to the estimation of the superior limit of the testing duration, [%];

L=α/2 - the index used to note the inferior confidence limits; represents the value of the probability corresponding to the estimation of the inferior limit of the testing duration, [%].

1.3 Review on the Calculus of the Testing Period

When using the Weibull distribution in the modeling of the products' reliability, these is estimated, in the majority of situations, on the basis of experimental results obtained following their testing on stands, using censored tests or complete tests. The organization and the process of reliability tests of the products represent complex activities from an organizational standpoint and also big resources consumers. The calculation relations of the duration of a complete or a censored reliability test, found in the specialty literature, are based on a series of equations which allow the estimation of the empirical distribution function [9]:

$$F_n(t_r) = \frac{r}{n+1}. \tag{1}$$

The equation (1) gives mean values of the empirical distribution function. The mean values of the empirical distribution function can be obtained using:

$$F_n(t_r) = \frac{r-0.3}{n+0.4}, \tag{2}$$

or:

$$F_n(t_r) = \frac{i-0.30685-0.3863\cdot\left(\frac{r-1}{n-1}\right)}{n}, \tag{3}$$

for *n*>20 and:

$$F_n(t_r) = 1-2^{-\frac{1}{n}}\left(\frac{r-1}{n-1}\right)\cdot\left[2^{\left(1-\frac{1}{n}\right)}-1\right], \tag{4}$$

for *n*≤20.

The value of the duration of a reliability test censored at level r is obtained using the inverse function of distribution of the considered statistical model [4, 21, 22]:

$$\mathrm{T}_{r/n} = \eta\cdot\left\{\ln\left[\frac{1}{1-F_n(t_r)}\right]\right\}^{\frac{1}{\beta}}, \tag{5}$$

in this case being the Weibull distribution. The equation (5) results from the logarithmation of the Weibull distribution function, written as follows:

$$e^{-\left(\frac{t}{\eta}\right)^{\beta}} = 1-F(t). \tag{6}$$

we obtain:

$$\left(\frac{t}{\eta}\right)^{\beta} = \ln\left[\frac{1}{1-F(t)}\right]. \tag{7}$$

We notice that after a series of algebraic calculations, the equation (7) can be written in the form of equation (5). Also, in the equation (5), instead of F(t) we used the value determined by using one of the (1) ÷ (4) relations. Thus we obtained mean or median values of the duration of testing. For complete test case in relations (1) ÷ (4), parameter *r* is replaced by the value of the sample volume used *n*. Consequently, the objective of this paper is to present an estimation modality of the mean duration for censored and/or complete reliability tests, as well as of the confidence intervals for this duration. Knowing these values allows for the careful planning of the testing activities [4, 23, 24].

2 Statistical Calculation Model

The value obtained for a reliability test doesn't offer important data regarding the real duration of a test, because the time of operation until failure of a tested product represents a random variable.

For this situation, a favorable solution consists is the determination of the confidence intervals of the duration of the test. These intervals contain the real value of the test, with a 1- α probability [25, 26]:

$$\Pr\left(\mathrm{T}_L \le \mathrm{T}_{r/n} \le \mathrm{T}_U\right) = 1-\alpha. \tag{8}$$

The calculation of the confidence intervals is realized in the conditions of a Bernoulli extraction (the scheme of the urn with returned balls). Thus, the median value of probability at which from *n* products subjected to testing, a number of *r* products fail, results as a solution to the equation [4, 27, 28]:

$$\sum_{i=r}^{n} C_n^i \cdot F_{Me}^i \cdot \left(1-F_{Me}\right)^{n-i} = 0.5. \tag{9}$$

The difficulties in calculation which can occur solving the equation (9), depending on F_{Me}, can be eliminated by using an approximate value:

$$F_{Me} = \frac{1}{1+\dfrac{n-r+1}{r}\cdot Q_{0.5,2\cdot(n-r+1),2\cdot r}}. \tag{10}$$

The equation (10) represents the connecting relation that can be established between the binomial distribution and the Fisher-Snedecor distribution [4, 11, 26].

Using the F_{Me} solution, obtained by solving one of the (9) or (10) equations, along with to equation (5), leads to the obtaining of the duration of the reliability test. In fact, the equations (1) ÷ (4) are nothing more than regression relations of the solutions of equation (9), for different combinations of the parameters *n* and *r*.

Because, by definition, the distribution function is an ascending function, the confidence level of the duration of testing period results by using the solutions of the equations:

$$\sum_{i=r}^{n} C_n^i \cdot F_L^i \cdot (1 - F_L)^{n-i} = \frac{\alpha}{2}, \tag{11}$$

and

$$\sum_{i=r}^{n} C_n^i \cdot F_U^i \cdot (1 - F_U)^{n-i} = 1 - \frac{\alpha}{2}, \tag{12}$$

together with (5), namely:

$$\mathrm{T}_L = \eta \cdot \left\{ \ln\left[\frac{1}{1 - F_L} \right] \right\}^{\frac{1}{\beta}}, \tag{13}$$

and

$$\mathrm{T}_U = \eta \cdot \left\{ \ln\left[\frac{1}{1 - F_U} \right] \right\}^{\frac{1}{\beta}}, \tag{14}$$

A similar value of the F_L and F_U probabilities can be obtained by approximating the binomial distribution through the Fisher-Snedecor distribution:

$$F_L = \frac{1}{1 + \frac{n - r + 1}{r} \cdot Q_{1-\frac{\alpha}{2}, 2\cdot(n-r+1), 2\cdot r}}. \tag{15}$$

and

$$F_U = \frac{1}{1 + \frac{n - r + 1}{r} \cdot Q_{\frac{\alpha}{2}, 2\cdot(n-r+1), 2\cdot r}}. \tag{16}$$

For the case of the complete tests, in the calculation relations (10) and (15), (16) the parameter *r* is replaced with the value of the used sample volume *n*.

3 Simulation Study

3.1 Program Description

To verify the precision of reasoning and of the mathematic model proposed at point 2, we used the Monte Carlo simulation method. The method implies, in this case, generating a very high number of samples (N>>1000), that belong to a completely specified Weibull population, W(t,β,η). This database is then subjected to a statistical analysis that is aimed at the duration of a reliability test using a censored plan with the *n* and *r* parameters. For the development of this study we created a Mathcad calculus programme. The logical chart of this programme is presented in Figure 1. The running of the programme implies the determination of the following entry data: N, n, r, β, η and α. The program generates a matrix with n lines and N columns, using the generator of random, uniform and continuous numbers within the [0,1] interval. The values of simulated failure times are obtained by using the inverse function of distribution of the Weibull statistical:

$$t = \eta \cdot \ln\left(\frac{1}{1-rnd(1)}\right)^{\frac{1}{\beta}}. \tag{17}$$

Thus, we obtain a matrix with *n x N* dimensions, in which every column represents a reliability test. In order to determine the duration of censored tests at level *r*, the calculation program sets in ascending order the columns of the previously generated matrix.

Also, the *r*-th line of this matrix is extracted at the end. The *N* values contained in this line represent the simulated durations of the reliability tests ($t_{r,i}$). The calculation of the median and mean durations of the testing duration is made by determining the median and the mean of these values:

$$t_{Me} = \begin{cases} t_{\left(\frac{n+1}{2}\right)}, & \text{if } N \text{ is even} \\ \dfrac{t_{\left(\frac{n}{2}\right)} + t_{\left(\frac{n}{2}+1\right)}}{2}, & \text{if } N \text{ is odd} \end{cases} \tag{18}$$

and

$$m = \frac{\sum_{i=1}^{N} t_{r,i}}{N} \tag{19}$$

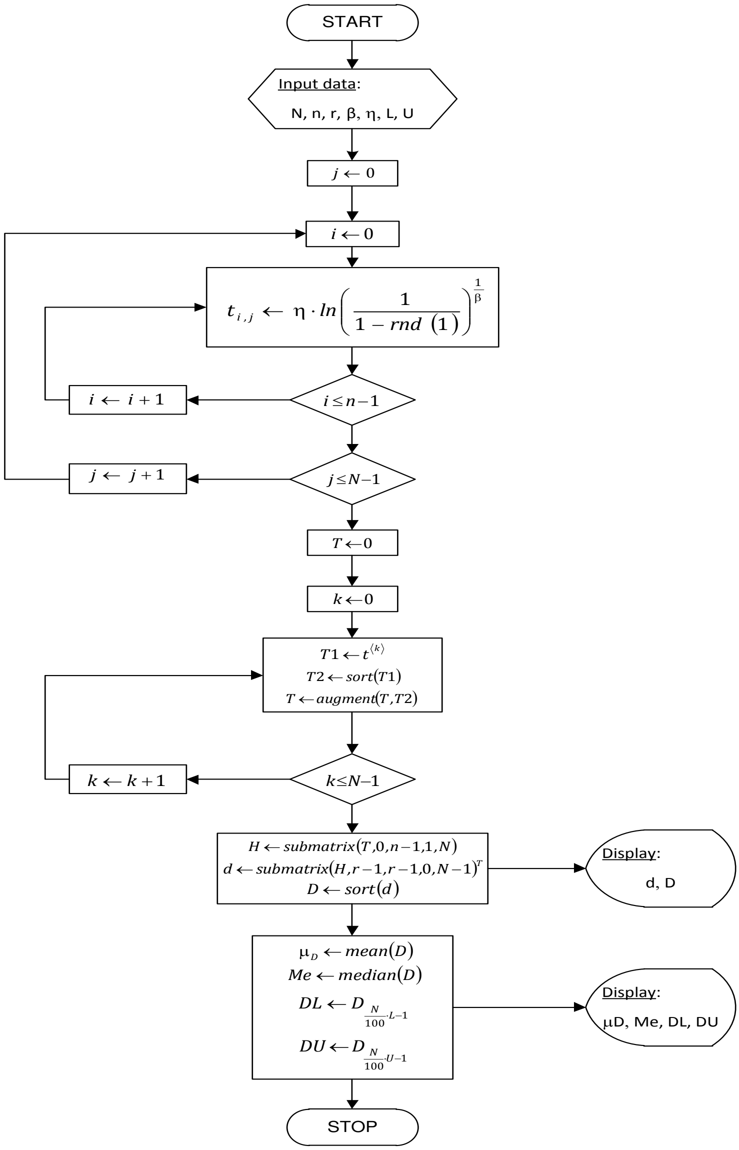

START
Input data:
N, n, r, β, η, L, U
$j \leftarrow 0$
$i \leftarrow 0$
$t_{i,j} \leftarrow \eta \cdot ln\left(\frac{1}{1 - rnd\ (1)}\right)^{\frac{1}{\beta}}$
$i \leq n-1$
$i \leftarrow i+1$
$j \leq N-1$
$j \leftarrow j+1$
$T \leftarrow 0$
$k \leftarrow 0$
$T1 \leftarrow t^{\langle k \rangle}$
$T2 \leftarrow sort(T1)$
$T \leftarrow augment(T,T2)$
$k \leq N-1$
$k \leftarrow k+1$
$H \leftarrow submatrix(T,0,n-1,1,N)$
$d \leftarrow submatrix(H,r-1,r-1,0,N-1)^T$
$D \leftarrow sort(d)$
Display:
d, D
$\mu_D \leftarrow mean(D)$
$Me \leftarrow median(D)$
$DL \leftarrow D_{\frac{N}{100} \cdot L-1}$
$DU \leftarrow D_{\frac{N}{100} \cdot U-1}$
Display:
μD, Me, DL, DU
STOP

Figure 1

The logical scheme of the Monte Carlo numerical simulation program

In the previous equations, we noted $t_{(p)}$ as the p quantile of the t variable. The determination of confidence limits for the duration of the tests is realized by determining the $t_{L/100}$ and $t_{U/100}$ quantiles of the truncation durations for the N simulated tests. The calculation method used for the determination of p quantiles applies the equation:

$$t_p = t_{(p\cdot(N+1))}. \tag{20}$$

If the $p\cdot(n+1)$ expression doesn't generate an integer value, then for the determination of the p quantile we recommend the use of linear interpolation. We assume that, after the evaluation of the $p\cdot(n+1)$ expression, we find that the value of the t_p quantile is included in the $[t_{(k)}, t_{(k+1)}]$.. To determine the value of the t_p quantile, we use the relation:

$$t_p = t_{(k)} + [p\cdot(n+1) - k]\cdot[t_{(k+1)} - t_{(k)}] \tag{21}$$

For high volumes of sample (such is the case with the realized application), instead of the previous relation we can use equation (22).

$$t_p = t_{[p\cdot(N+1)]}. \tag{22}$$

In equation (22), by $t_{[\mathrm{p}]}$ we noted the integer part of the value of the expression between brackets.

3.2 Monte Carlo Simulation Data

To demonstrate the way of using the calculation methodology presented in the third point of this paper, we present further several case studies, determined for different values of the Weibull distribution parameters β and η, as well as for different testing schemes n and r.

The solving of the equations (9), (11) and (12) was made using the specialized functions existent in Mathcad 14. The solving accuracy of these equations was established at 10^{-15}. In parallel, we presented the values obtained by using the approximate relations (10), (15) and (16).

The obtained F_{Me}, F_L și F_U probabilities are then used to determine the median duration of the reliability test, eq. (5) and the limits of the confidence interval (1-α) for this duration (T_L și T_U).

The values for these limits are obtained by using the equations (13) and (14). The significance level was established at the value of α =10%. In Table 1 values for different combinations of the Weibull distribution's parameters and different censored testing plans, *n/r* are presented.

The accuracy of the obtained results, by using the proposed calculation methodology, was verified using the Monte Carlo simulation.

For this purpose we used the MathCAD 14 software, which is described at point 3.1. The calculus programme was run for the same combinations of values of the Weibull distribution's parameters, as in the previous case. Also, the number of simulations was established at the value N=10000 and the confidence interval 1-α at 90%.

Table 1

The determination of the testing durations related with the reliability tests using the calculation relations

Test plan		The calculated duration of the reliability test					
		$T_{r/n}$		T_L		T_U	
n	r	eq. (8)	eq. (9)	eq. (10)	eq. (14)	eq. (11)	eq. (15)
β=1.5, η=50							
10	5	35.605	35.605	19.926	19.926	56.218	56.218
10	10	97.035	97.035	61.118	61.118	151.520	151.520
20	5	20.417	20.417	11.472	11.472	32.023	32.023
20	10	37.332	37.332	25.278	25.278	51.936	51.936
20	15	58.860	58.860	42.591	42.591	78.556	78.556
20	20	112.601	112.601	78.641	78.641	164.494	164.494
β=1.5, η=100							
10	5	71.211	71.211	39.854	39.854	112.438	112.438
10	10	194.069	194.069	122.236	122.236	303.040	303.040
20	5	40.834	40.834	22.945	22.945	64.045	64.045
20	10	74.664	74.664	50.556	50.556	103.871	103.871
20	15	117.720	117.720	85.182	85.182	157.112	157.112
20	20	225.201	225.201	157.282	157.282	328.989	328.989
β=2, η=50							
10	5	38.760	38.760	25.080	25.080	54.595	54.595
10	10	82.213	82.213	58.126	58.126	114.841	114.841
20	5	25.541	25.541	16.576	16.576	35.796	35.796
20	10	40.161	40.161	29.978	29.978	51.445	51.445
20	15	56.508	56.508	44.333	44.333	70.166	70.166
20	20	91.917	91.917	70.223	70.223	122.139	122.139
β=2, η=100							
10	5	91.917	91.917	70.223	70.223	122.139	122.139
10	10	164.425	164.425	116.252	116.252	229.681	229.681
20	5	51.082	51.082	33.152	33.152	71.592	71.592
20	10	80.322	80.322	59.956	59.956	102.890	102.890
20	15	113.015	113.015	88.667	88.667	140.332	140.332
20	20	183.835	183.835	140.446	140.446	244.279	244.279

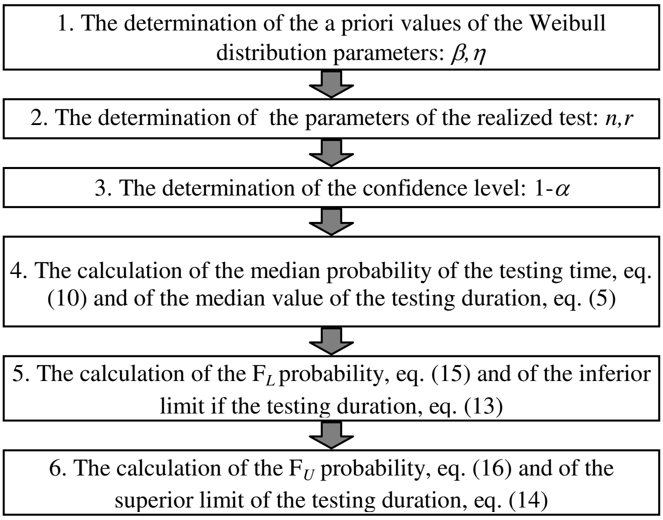
Under these conditions, we determined the median values, the mean values and the confidence intervals for the testing duration. The results obtained by Monte Carlo numerical simulation are presented in Table 2.

Table 2

The determination of the testing durations related to the reliability tests using the Monte Carlo method

Test plan		Values obtained by Monte Carlo simulation			
		$T_{r/n}$	T_L	T_U	$\check{T}_{r/n}$
n	r				
β=1.5, η=50					
10	5	35.744	19.998	55.945	36.562
10	10	96.820	61.412	152.861	100.609
20	5	20.385	11.680	31.748	20.885
20	10	37.301	25.362	52.071	37.843
20	15	58.992	42.416	78.349	59.594
20	20	112.252	78.947	165.037	115.808
β=1.5, η=100					
10	5	71.140	39.118	111.946	72.881
10	10	193.087	121.367	301.877	200.325
20	5	40.782	22.834	63.778	41.770
20	10	74.810	50.408	103.875	75.681
20	15	118.342	85.409	158.009	119.295
20	20	225.390	157.378	328.730	231.811
β=2, η=50					
10	5	38.614	25.034	54.654	39.101
10	10	82.338	58.557	115.534	83.971
20	5	25.502	16.507	35.797	25.793
20	10	40.211	29.926	51.526	40.393
20	15	56.463	44.296	70.287	56.759
20	20	91.957	70.221	121.616	93.419
β=2, η=100					
10	5	91.957	70.221	121.616	93.419
10	10	164.752	116.708	228.715	167.494
20	5	51.307	33.260	71.201	51.696
20	10	80.452	59.513	102.790	80.616
20	15	113.174	88.908	139.999	113.716
20	20	183.619	140.521	244.166	186.904

The using mode of this method for the estimation of the durations of the complete and/or at r level censored reliability tests is presented in Figure 2.



1. The determination of the a priori values of the Weibull distribution parameters: β, η

2. The determination of the parameters of the realized test: *n,r*

3. The determination of the confidence level: 1-α

4. The calculation of the median probability of the testing time, eq. (10) and of the median value of the testing duration, eq. (5)

5. The calculation of the F_L probability, eq. (15) and of the inferior limit if the testing duration, eq. (13)

6. The calculation of the F_U probability, eq. (16) and of the superior limit of the testing duration, eq. (14)

Figure 2

The calculation algorithm for the duration of the reliability tests

Conclusions

Based on the results presented in Tables 1 and 2, we will realize several comparative studies to show the correctness of the proposed calculus method. In Figure 3 the results of the median duration of a censored test at level *r* are presented, realized on a sample of *n* volume, on a graphical form for the case study β=1.5, η=50. The inferior and superior limits of the testing duration for the case study in question (β=1.5, η=50) are presented in Figures 4 and 5.

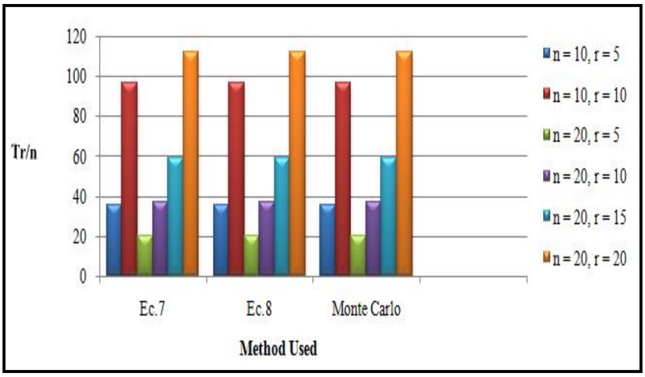


Figure 3

The duration of a censored test level *r*, realized on sample size *n* ($T_{n/r}$)

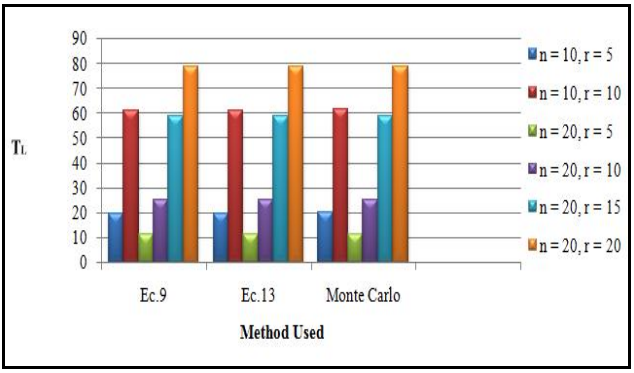


Figure 4
Limit inferior duration of a censored test level r, realized on sample size *n* (T_L)

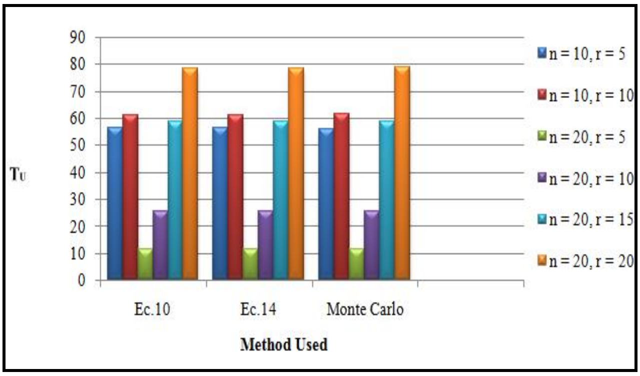


Figure 5
Limit superior duration of a censored test level r, realized on sample size *n* (T_U)

The presented calculation model allows the obtaining of accurate estimated values, because the differences towards the simulated values are very small. If the number of simulations would grow, the resulting differences would be insignificant. The approximate relations (10), (15) and (16) lead to the obtaining of some values, which, at the results' display accuracy of 10^{-3}, don't differ from the real values obtained through the equations (9), (11), (12). Based on the presented results we found the significant reductions in time that can be made by using the censored testing plans.

Given the powerful competition on the industrial market, we can no longer imagine the realization of a product without a rigorous quality and reliability control of the product, based on different types of tests, in all the stages of the products' existence, from the raw materials being used, up to their use. In these types of tests, we put special emphasis on the reliability tests. The testing laboratories, that possess modern testing equipment and highly-trained personnel, have acquired an ever increasing development. Today, almost every company has a reliability test laboratory, adequately equipped to the type of products it realizes. The application of censored testing plans, using the Monte Carlo method, determines the testing duration of the products in a shorter time and in conditions of economic efficiency.

The proposed method of calculation has applicability in laboratories that are specialized in testing the materials or the products from different fields of use. These fields are given by the particular versatility of the Weibull distribution's model:

- the tear resistance, the corrosion resistance, the wear resistance, the fatigue resistance and the contact fatigue resistance of textile and metallic materials;
- the modeling of the materials properties: steels, titanium, semi-conductor materials, tungsten, ceramic, glass, plastic materials, porcelain, graphite, paper, textile fibers, composite materials;
- the modeling of the durability of mechanical components: bearings, engines, motor vehicles' structures, tools;
- the modeling of the functioning times of relays, passive electronic components (resistors and capacitors) and active electronic components (transistors, integrated circuits);
- the modeling of the life times of subsystems, made of identical component elements, in series connected and whose behaviour is described using the gamma distribution.

Acknowledgement

This paper is supported by the Sectoral Operational Programme Human Resources Development (SOP HRD), financed from the European Social Fund and by the Romanian Government under the project number POSDRU/89/1.5/S/59323.

References

[1] D. W. Benbow. *The Certified Reliability Engineer Handbook,* ASQ Quality Press, Milwaukee, 2009

[2] B. Bertsche. Reliability in Automotive and Mechanical Engineering: Determination of Component and System Reliability, Springer, Berlin, 2010

[3] P. O'Connor & A. Kleyner. *Practical Reliability Engineering*, 5th Edition. Wiley & Sons, New Jersey, 2012

[4] D. B. Kececioglu. *Reliability Engineering Handbook*, Vol. I, PTR Prentice-Hall, New Jersey, 1991

[5] G. B. Yang. *Life Cycle Reliability Engineering*. New Jersey, Wiley, 2007

[6] I. Arizonoa, Y. Kawamuraa, Y. Takemotob. Reliability Tests for Weibull Distribution with Variational shape Parameter Based on Sudden Death Lifetime Data, *European Journal of Operational Research*, 189(2): 570-574, 2008

[7] E. E. Lewi. *Introduction to Reliability Engineering,* Wiley, New Jersey, 1995

[8] S. C. Saunders. *Reliability, Life Testing and the Prediction of Service Lives,* Springer, New York, 2007

[9] E. Zio. *An Introduction to the Basics of Reliability and Risk Analysis*, World Scientific, New Jersey, 2007

[10] A. Joarder, H. Krishna, D. Kundu. Inferences on Weibull Parameters with Conventional Type-I Censoring, *Computational Statistics & Data Analysis*, 55(1): 1-11, 2011

[11] D. B. Kececioglu. *Reliability & Life Testing Handbook,* Vol. II, PTR Prentice-Hall, New Jersey, 1994

[12] C. Kim, J. Jung, Y. Chung. Bayesian Estimation for the Exponentiated Weibull Model Under Type II Progressive Censoring, *Statistical papers*, 52(1): 53-70, 2011

[13] S. M. K. Quadri & N. Ahmad. Software Reliability Growth Modeling with New Modified Weibull Testing–effort and Optimal Release Policy, *International Journal of Computer Applications,*6(12):1-10, 2010

[14] M. I. Ageel. A Novel Means of Estimating Quantiles for 2 – Parameter Weibull Distribution under the Right Random Censoring Model, *Journal of Computational and Applied Mathematics*, 149(2): 373-380, 2002

[15] D. Kundu. Bayesian Inference and Life Testing Plan for the Weibull Distribution in Presence of Progressive Censoring, *Technometrics*, 50(2): 144-154, 2008

[16] H. Chin-Yu. Cost-Reliability-Optimal Release Policy for Software Reliability Models Incorporating Improvements in Testing Efficiency, *The Journal of Systems and Software,* 77 (2):139-155, 2005

[17] P. C. Jha, D. Gupta, B. Yang, P. K. Kapur. Optimal Testing Resource Allocation during Module Testing Considering Cost, Testing Effort and Reliability, *Computers & Industrial Engineering*, 57(3): 1122-1130, 2009

[18] A. Kleynera & P. Sandborn. Minimizing Life Cycle Cost by Managing Product Reliability via Validation Plan and Warranty Return Cost, *Int. J. Production Economics*, 112(2): 796-807, 2008

[19] Y. I. Kwon. A Bayesian Life Test Sampling Plan for Products with Weibull Lifetime Distribution Sold under Warranty, *Reliability Engineering & System Safety*, 53(1): 61-66, 1996

[20] S. M. Zaharia, I.Martinescu, C. O. Morariu. Life Time Prediction using Accelerated Test Data of the Specimens from Mechanical Element. *Eksploatacja i Niezawodnosc - Maintenance and Reliability*; 14(2): 99-106, 2012

[21] H. S. Alkutubi & H. M. Ali. Maximum Likelihood Estimators with Complete and Censored Data, *European Journal of Scientific Research*, 54(3): 407-410, 2011

[22] H. Panahi & S. Asadi. Estimation of the Weibull Distribution Based on Type-II Censored Samples, *Applied Mathematical Sciences*, 5(52): 2549-2558, 2011

[23] O. P. Yadav, N. Singh, P. S. Goel. Reliability Demonstration Test Planning: A Three Dimensional Consideration, *Reliability Engineering and System Safety,* 91(8): 882-893, 2006

[24] G. Yang & L. Jin. Best Compromise Test Plans for Weibull Distributions with Different Censoring Times, *Quality and Reliability Engineering International*, 10(5): 411-415, 1994

[25] C. O. Morariu & T. Păunescu. *Computer Applications in Engineering - Mathcad 2001*, Ed. Transilvania University of Braşov, 2004

[26] C. O. Morariu. *Applied Probability and Statistics*, Ed. Transilvania University of Braşov, 2010

[27] G. C. Perdona & F. Louzada – Neto. Interval Estimation for the Parameters of the Modified Weibull Distribution Model with Censored Data: a Simulation Study, *TEMA Tend. Mat. Apl. Comput*, 9(3): 437-446, 2008

[28] S. Vittal & R. Phillips. Uncertainty Analysis of Weibull Estimators for Interval-Censored Data, *Reliability and Maintainability Symposium – RAMS'07*, pp. 292-297, 22-25, Jan. 2007

Application of Fuzzy Multi-Criteria Decision Making Analysis for Evaluating and Selecting the Best Location for Construction of Underground Dam

Parviz Rezaei[1], Kamran Rezaie[2], Salman Nazari-Shirkouhi[2], Mohammad Reza Jamalizadeh Tajabadi[3]

[1]Roudbar Branch, Islamic Azad University, Roudbar, Iran

[2]Department of Industrial Engineering, College of Engineering, University of Tehran, Tehran, Iran
Email: krezaie@ut.ac.ir; snnazari@ut.ac.ir

[3]Faculty of Natural Resources, University Of Zabol, Zabol, Iran

Abstract: One crisis which human beings will probably face in the upcoming decades is the water crisis. The crisis in arid and semi-arid regions covering a large part of Iran would be much more severe. Thus, using novel methods of water collection such as construction of underground dam is so important. Decision making and selection of an appropriate option in construction of such dams is one basic challenge. The major issue in construction of such dams is selecting an appropriate location. Selecting the best location for building underground dams is a challenge due to involvement of a wide range of influential factors. In this paper, analytic hierarchy process (AHP), one of multi-criteria decision making (MCDM) techniques in fuzzy environment is applied to select the optimal alternative for construction of an underground dam in a case study. Results show that using AHP in the fuzzy environment improves decision making through considering more important factors in decision making.

Keywords: underground dam; multi-criteria decision making; fuzzy theory; AHP

1 Introduction

Water shortage in arid and semi-arid regions is one of the problems of policy makers. Various solutions have been used to overcome this problem around the world. One of such solutions is construction and use of underground dams. In recent years, efforts have been made at national level to use dams more because of increase in severity, extent, and frequency of droughts. Thus, steps were taken quickly so as to facilitate construction of more such dams in the country. Since construction and operations of these dams is a new technique in water resource management in Iran, the present paper attempts to compare application of two methods of fuzzy analytic hierarchy process and analytic hierarchy process. The aim of this comparison is to familiarize experts with these methods and to specify the strengths and weaknesses of these two methods.

Underground dams are built for different purposes such as prevention of saltwater and freshwater interference (Garagunis, 1981), avoidance of underground water penetration in the mines (Gupta et al., 1987), prevention of seawater into freshwater aquifers (Onder and yilmaz, 2000), and holding water for operation (Nilsson, 1988). Underground dams are usually constructed in the bed of a water stream where signs of underground water is seen. For underground dams to be able to extract water, their construction should be justified but considering some factors as follows: bedrock depth in the water stream, water stream width, impenetrable walls, suitable sediments, the volume of useable water, suitable context for using extracted water, social issues, economic justification and so on (Nilsson, 1988). Evaluation of factors requires detailed studies to be carried out before determining initial appropriate options. Therefore, the first step in constructing underground dams is to find suitable options.

Basically, several factors influence on selection of an alternative for construction of underground dams. Taking into account all of these factors makes the decision making problem so complex. Thus, multi-criteria decision making (MCDM) methods are applied to tackle this problem.. One widely used MCDM method is the analytic hierarchy process (AHP) which has been used in various managerial areas from hydrogen production methods (Pilavachi et al., 2009) to motor cleaning systems and many others. In addition, AHP has been applied for water resource management in many studies such as Anagnostopoulos et al. (2005, 2007), Srdjevic (2007), and Mei et al. (1989). Mei et al. (1989), in applying multi-criteria decision making methods for water resource management in China, stated that analytic hierarchy process method specifies not only relative importance of each factor, but also it specifies combined weights matched to initial goal. Akpinar et al. (2005) believed that multi-criteria decision making methods are useful in planning for issues in which many factors are involved. They used this method in determining agricultural land-use types in Turkey and approved successfulness of analytic hierarchy process method in priority setting in agricultural land-use types. Montazar and Behbahani (2007) developed an optimized irrigation system selection model using analytic hierarchy process. Their findings indicated that results obtained from this model are consistent with those obtained from field evaluations. In addition, evaluation of methods showed that results provided by this model were more reliable than ordinary weighting methods. Okada et al. (2008) applied analytic hierarchy process for improvement of irrigation project. They found out that the first priority for irrigation management planners is the water delivery. In fact, they considered appropriate water allocation and control as the main factor for these planners. Montazar and Zadbagher (2010) used an analytic hierarchy model for assessing global water productivity in irrigation networks. They stressed that AHP is a practical and comprehensive tool for improving effectiveness of such systems.

Standard hierarchical analysis process is not effective to solve more complicated problems. Therefore, some modifications are necessary for this method. Combining fuzzy methods with analytic hierarchy process is one approach for solving the complicated problems. Fuzzy analytic hierarchy process (fuzzy AHP) has been applied in different problems as follows: in geographical information system (GIS) application (Vahidnia et al., 2008), risk evaluation of information technology projects (Iranmanesh et al., 2008), water management plans assessment (Srdjevic and Medeiros, 2008), and eco-environmental vulnerability assessment (Li et al., 2009). Kong and Liu (2005) applied fuzzy analytic hierarchical process to evaluate success factors of e-commerce. They stressed that fuzzy AHP has qualifications of both subjective and thematic factors in the decision making process. Stirn (2006) integrated the fuzzy AHP with dynamic programming approach for determining the optimal forest management decisions so that he could maximize economic, ecological and social benefits. Results indicated that this method can be successful in problems where different criteria are involved in decision making. Ascough et al. (2008) stated that decision making in natural environment is difficult due to inherent complexity of the environment and different interests of decision makers and operators. They proposed solutions to overcome this problem which are based on using fuzzy systems. They found out combining fuzzy systems with other decision making methods useful. Alias et al. (2009) applied fuzzy analytic hierarchy process for logical use of Johor River in Malaysia. The considered different dimensions of the area and concluded fuzzy method with triangular fuzzy numbers can be successful for ambiguous data. Opricovic (2011) applied fuzzy AHP with fuzzy VIKOR for water resources planning. Tsiko and Haile (2011) used GIS and fuzzy AHP in modeling optimum sites for locating water reservoirs.

Locating underground dam construction projects is a complex problem due to existence of uncertainty in factors influencing on it. Since the real world is full of ambiguities and imprecise and vague terms, most decision makers in field of underground dam construction know using linguistic terms more practical and feasible. Zadeh (1965) introduced fuzzy sets theory as a powerful tool to dominate these ambiguities, vagueness and uncertainties when there is a special complexity and lack of complete information on experts' opinions. In the current study, a useful and practical methodology is presented for group decision making on the location of underground dams construction based on the AHP and fuzzy theory.

The rest of the paper is structured as follows: Section 2 describes the proposed methodology; in Section 3 the proposed methodology is applied to locate the underground dam construction as an experiment and results are provided; in Section 4, the proposed methodology is tested for the verification and validation purposes; finally Section 5 includes conclusions of the present work.

2 Proposed Methodology

In this section, the proposed fuzzy AHP based methodology is presented for evaluating and selecting the best location for underground dam construction location. The steps of the proposed methodology are illustrated in Figure 1. The steps will be implemented in a case study and described in great details.

2.1 Fuzzy Analytic Hierarchy Process

AHP is a decision making method for decomposing the hierarchical problem and can apply to solve a complex multi-criteria decision problem (Saaty, 1980). In the literature, AHP has widely been applied to solve the different MCDM problems. Many times decision makers are only able to provide a subjective and uncertain answer rather than an exact value (Shaw et al., 2012). Hence, such answers need to be quantified. Conventional methods of AHP cannot be used for decision making problem in the real world when fuzziness and vagueness is observed in data of problems. To handle such uncertainties and vagueness, fuzzy sets theory, initially introduced by Zadeh (1965), can be applied. Therefore, incorporation of the fuzzy concept with AHP can be more applicable and more effective than the conventional AHP in the real world problems. This issue has attracted many researchers to apply fuzzy AHP in different fields such as risk and disaster management (Takács, 2010), work safety evaluation (Zheng et al., 2012), green initiatives in the fashion supply chain (Wang et al., 2012).

Figure 1 shows the proposed fuzzy AHP based methodology for decision making on selection of the best location form underground dam construction. The steps of the proposed methodology are as follows:

Step1: ***Determining Criteria and Alternatives and Establish hierarchal structure***

The first step of our methodology is to determine the criteria which are going to be affected for making a decision about underground dam construction location.

Step 2: ***Collecting experts' judgments based on fuzzy scale and establish fuzzy pair-wise comparison matrices***

The sample questionnaire by Azadeh et al. (2010, 2011) and Nazari-Shirkouhi et al. (2011) can be applied to collect the experts' judgments based on fuzzy scales. In the present paper, the triangular fuzzy numbers (TFNs) for fuzzy membership function applied to enable the decision maker to make easier decisions (Kaufmann and Gupta, 1988). The membership function of a TFN is shown in Equation (1). The TFN is usually shown with $\tilde{A} = (l, m, u)$, where $l \leq m \leq u$

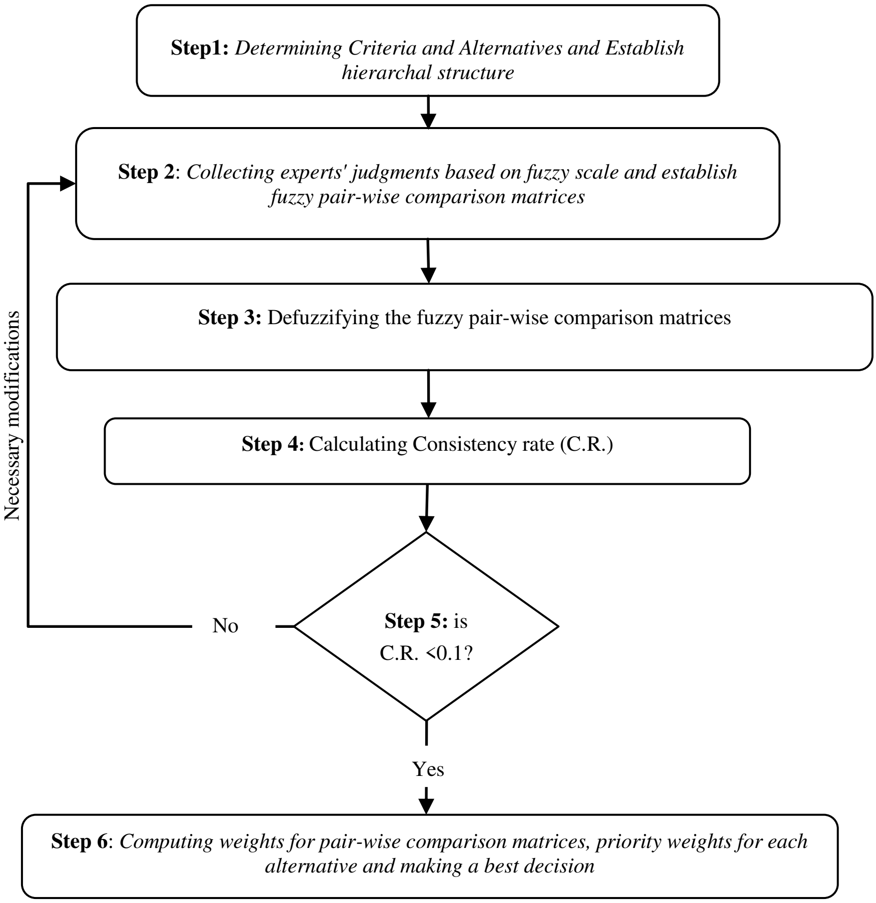

Step1: *Determining Criteria and Alternatives and Establish hierarchal structure*
Step 2: *Collecting experts' judgments based on fuzzy scale and establish fuzzy pair-wise comparison matrices*
Step 3: Defuzzifying the fuzzy pair-wise comparison matrices
Step 4: Calculating Consistency rate (C.R.)
Step 5: is C.R. <0.1?
No
Yes
Necessary modifications
Step 6: *Computing weights for pair-wise comparison matrices, priority weights for each alternative and making a best decision*

Figure 1

The proposed methodology based on fuzzy AHP

$$\mu_{\tilde{A}}(x)=\begin{cases}\dfrac{x-l}{m-l} & l\le x\le m\\ \dfrac{u\text{-}x}{u\text{-}m} & m\le x\le u\\ 0 & x\prec l \ \text{ or } \ x\succ u\end{cases} \qquad (1)$$

Where $-\infty\prec x\prec\infty$; $\mu_{\tilde{A}}(x)$ is a continuous mapping from *R* to the interval [0,1]. For two TFNs $\tilde{A}$ = (l_1, m_1, u_1); $\tilde{B}$ = (l_2, m_2, u_2), some of the main mathematical operations can be expressed in equation (2) as bellow:

$$\begin{aligned}
&\tilde{A}+\tilde{B}=\left(l_1+l_2,\ m_1+m_2,\ u_1+u_2\right)\\
&\tilde{A}-\tilde{B}=\left(l_1-u_2,\ m_1-m_2,\ u_1-l_2\right)\\
&\tilde{A}*\tilde{B}=\left(l_{1*}l_2\ ,\ m_{1*}m_2,\ u_{1*}u_2\right)\\
&\tilde{A}/\tilde{B}=\left(l_1/u_2,\ m_1/m_2,\ u_1/l_2\right)\\
&k\tilde{A}=(kl_1,km_1,ku_1),k>0,k\in R
\end{aligned} \tag{2}$$

The AHP method uses pair-wise comparisons and related matrix is shown in Equation (3).

$$\tilde{A}^k=[\tilde{a}_{ij}]^k=\begin{array}{c} \\ C_1\\ C_2\\ \vdots\\ C_n\end{array}\begin{array}{c}\begin{array}{cccc} C_1 & C_2 & \cdots & C_n \end{array}\\ \begin{bmatrix} \tilde{1} & \tilde{a}_{12} & \cdots & \tilde{a}_{1n}\\ 1/\tilde{a}_{12} & \tilde{1} & \cdots & \tilde{a}_{2n}\\ \vdots & \vdots & \ddots & \vdots\\ 1/\tilde{a}_{1n} & 1/\tilde{a}_{2n} & \cdots & \tilde{1}\end{bmatrix}^k\end{array} \tag{3}$$

Where, $\tilde{a}_{ij}^k=(1,1,1):\forall i=j;\ \tilde{a}_{ij}^k=\dfrac{1}{\tilde{a}_{ij}^k}:\forall i\neq j$.

$\tilde{A}^k$ is the fuzzy judgment matrix of *k*th expert, $\tilde{a}_{ij}^k$ is a the fuzzy evaluation between criterion *i* and criterion *j* of *k*th expert, $\tilde{a}_{ij}=(l_{ij}^k,m_{ij}^k,u_{ij}^k)$.

To aggregate the experts' judgments, Buckley (1985)'s method is applied here. As is shown in equations (4-7) *l, m,* and *u* show the minimum possible, most likely and the maximum possible value of a fuzzy number, respectively. TFN $\tilde{A}^k$ is defined as the following:

$$\tilde{A}_{ij}=(l_{ij},m_{ij},u_{ij}):l_{ij}\le m_{ij}\le u_{ij},l_{ij},m_{ij},u_{ij}\in[1/9,9] \tag{4}$$

$$l_{ij}=\min(a_{ijk}) \tag{5}$$

$$m_{ij}=\sqrt[K]{\prod_{k=1}^{K}(a_{ijk})} \tag{6}$$

$$u_{ij}=\max(a_{ijk}) \tag{7}$$

Which, a_{ijk} shows the relative importance of criteria C_i and C_j given by expert *k*.

The linguistic scale and underlying TFNs are illustrated in Table 1 based on Azadeh et al. (2011) and Nazari-Shirkouhi et al. (2011).

Table 1

The linguistic scale and underlying TFN

Fuzzy number	Linguistic scales	Scale of fuzzy number
$\tilde{1}$	Equally important	(1, 1, 1)
$\tilde{3}$	Weakly important	(2, 3, 4)
$\tilde{5}$	Essentially important	(4, 5, 6)
$\tilde{7}$	Very strongly important	(6, 7, 8)
$\tilde{9}$	Absolutely important	(7, 8, 9)
$\tilde{2},\tilde{4},\tilde{6},\tilde{8}$	Intermediate values ($\tilde{x}$)	(*x*- 1, *x*, *x*+1)
$1/\tilde{x}$	Between two adjacent judgments	(1/(*x*+ 1), 1/*x*, 1/ (*x*- 1))

Step 3: ***Defuzzifying the fuzzy pair-wise comparison matrices***

There are various methods to defuzzify fuzzy numbers. In this paper, we applied the Liou and Wang (1992) s' method to defuzzify fuzzy matrix $\tilde{A}$ into crisp matrix $g_{\alpha,\mu}$:

$$g_{\alpha,\mu}(\tilde{a}_{ij}) = [\mu.f_{\alpha}(l_{ij}) + (1-\mu).f_{\alpha}(u_{ij})],\ 0 \le \alpha, \mu \le 1 \tag{8}$$

$$g_{\alpha,\mu}(\tilde{a}_{ij}) = 1/\, g_{\alpha,\mu}(\tilde{a}_{ji}),\ \ 0 \le \alpha, \mu \le 1 : i > j \tag{9}$$

$f_{\alpha}(l_{ij}) = (m_{ij} - l_{ij}).\alpha + l_{ij}$ is the left-hand value $\alpha - cut$ for $\tilde{a}_{ij}$ and $f_{\alpha}(u_{ij}) = u_{ij} - (u_{ij} - m_{ij}).\alpha$ is the right-hand value $\alpha - cut$ for $\tilde{a}_{ij}$.

The range of uncertainty can be shown by α index. In other words, α index can indicate stable or unstable conditions. The larger value of α index indicates the lower degree of uncertainties. The μ index can be viewed as the degree of pessimism of a decision maker for the judgment matrix $\tilde{A}^{k}$. The larger value of μ index indicates the lower degree of optimism (decision maker is pessimistic).

Therefore, the defuzzified pair wise comparison matrix can be expressed as equation (10).

Step 4***: Calculating Consistency rate (C.R.)***

$$g_{\alpha,\mu}(\tilde{A}) = g_{\alpha,\mu}([\tilde{a}_{ij}]) =$$
$$\begin{array}{c} \\ C_1 \\ C_2 \\ \vdots \\ C_n \end{array} \begin{array}{c} \begin{array}{cccc} C_1 & C_2 & \dots & C_n \end{array} \\ \begin{bmatrix} 1 & g_{\alpha,\mu}(\tilde{a}_{12}) & \cdots & g_{\alpha,\mu}(\tilde{a}_{1n}) \\ 1/g_{\alpha,\mu}(\tilde{a}_{12}) & 1 & \cdots & g_{\alpha,\mu}(\tilde{a}_{2n}) \\ \vdots & \vdots & \ddots & \vdots \\ 1/g_{\alpha,\mu}(\tilde{a}_{1n}) & 1/g_{\alpha,\mu}(\tilde{a}_{2n}) & \cdots & 1 \end{bmatrix} \end{array} \tag{10}$$

Saaty (1980) suggests a consistency test to verify conformity of the calculation results. To calculate of consistency rate (C.R.), eigenvalue ($\lambda_{\max}$) of the single pair-wise comparison matrix $g_{\alpha,\mu}(\tilde{A})$ should be determined first. $\lambda_{\max}$ is calculated by equation (11).

$$\det(g_{\alpha,\beta}(\tilde{A}) - \lambda_{\max}) = 0 \tag{11}$$

After finding $\lambda_{\max}$, values of consistency index (*C.I.*) and *C.R.* can be obtained from equations (12-13):

$$C.I. = \frac{\lambda_{\max} - n}{n-1} \tag{12}$$

$$C.R. = \frac{C.I.}{R.I._n} \tag{13}$$

The value of Random index (*R.I.*) depends on the value of *n* and is the average consistency index for randomly generated entries (Saaty, 1980).

Step 5*: is C.R. <0.1?*

According to Saaty (1980), $C.R. < 0.1$ is acceptable scope; otherwise, for the comparison matrix modifications are necessary and new matrix must be solicited.

Step 6: *Computing weights of pair-wise comparison matrices, priority weights for each alternative and making a best decision*

The *W* is the weight of pair-wise comparison matrix $g_{\alpha,\beta}(\tilde{A})$. On the other hand, the *W* is eigenvector of matrix $g_{\alpha,\beta}(\tilde{A})$ and can be defined as equation (14).

$$[g_{\alpha,\beta}(\tilde{A}) - \lambda_{\max}].W = 0 \tag{14}$$

After calculating the weights for all pair-wise comparison matrices of the proposed hierarchical structure, in this step the final weight of the alternatives can be calculated and then the best decision made. The weights can be sorted decreasingly and the best alternative is selected finally.

3 Experiment and Results

In this section, the proposed methodology is implemented on an actual case in one of the biggest provinces (Kerman province) in Iran to select the best location for construction of an underground dam. Following successive droughts in the province and the benefits of underground dams in utilization of unconventional waters, the expert team suggested several options for selecting and evaluating the best location for building the underground dam Construction in the city of Rafsanjan. Figure 2 shows position of selected options over the city of Rafsanjan. Selection of the best location should be done based on criteria in such a way that all important technical factors are considered. The best location for underground dam construction can provide appropriate amount of water for agriculture in this region.

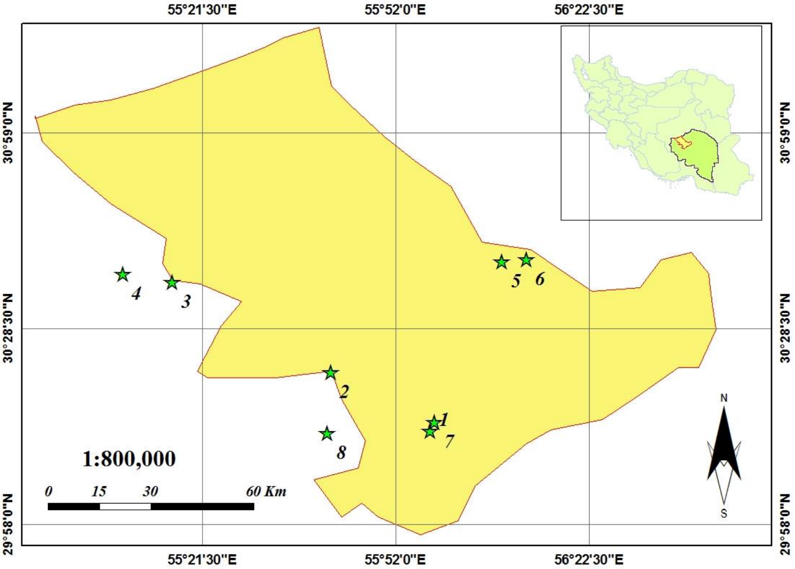


Figure 2

The position of 8 selected locations over Rafsanjan city and the Kerman province

The steps of the proposed methodology to select the best location for underground dam construction are described as follows:

Step1: *Determining criteria and alternatives and establish hierarchal structure*

The expert team should firstly determine the related criteria to evaluate the alternatives. The criteria and alternatives should be able to describe the existing difficult decision problem. Thus, considering these criteria and alternatives are very important for the decision makers' team in selecting the best location for underground dam construction.

The selected criteria according to the methodology of studying the physical specifications are as follows: bed width, utilization land area, distance to utilization location, bed slope, wall material which are extracted from the topographical maps. The, data are evaluated by experts and field studies to ensure the precision of data. After final approval, the proposed methodology is used to select the best location of project and its priorities.

Each criterion used for priority setting of a location has optimal values and conditions which should be met. For the slope, if it is high, it causes ejection of reserved water in the reservoir and thus water accumulation on its surface which, in turn, leads to subsequent problems. On the other hand, very low slope causes that there is a long distance when the reserved water is transferred and when it is transferred from the depth to the bed. Therefore, the best slope for selecting an option is about -12% (Nilsson, 1988). The minimum width is the most appropriate bed width. Of course, the less this width is, fewer water will be reserved. Thus, here it is assumed that bed width does not influence on the upstream reservoir. The third important factor is the wall material. The stronger and more impenetrable the walls are (and have fewer seams and cracks), the more appropriate they are considered. The stratification direction is also important which should be perpendicular to the flow. Another criterion is the distance between water extraction location and water utilization location. If this distance is shorter, construction cost will be lower. The last criterion is the area of agricultural lands which need using accumulated water in the dam. If the area of lands is large, justification for dam construction will be more logical. Related matrixes were built and calculations were performed following converting amounts of criteria intro measurable values.

After reviewing the literature related criteria, the experts' team considered eight candidate locations to evaluate with regard the expert's judgment who had worked in related field. Finally, the eight candidates are Goor choopan (Alternative 1), Khezr (Alternative 2), Bayaz (Alternative 3), Tezerj (Alternative 4), Uderj (Alternative 5), Joz (Alternative 6), Givdari (Alternative 7), and Dahaneh abolfazl (Alternative 8). The position of eight candidates over the city of Rafsanjan and Kerman province are shown in Figure 2.

After determining the criteria and alternatives, decision makers will setup hierarchical structure. The hierarchical structure should be able to break the existing complex decision problem into manageable components of different layers/levels (Nazari-Shirkouhi et al., 2011). The selected criteria can determine the levels of hierarchical structure. Level #1 (target level) addresses target (selecting the best location for underground dam construction). Level #2 (criterion level) addresses different factors impacting on locating decisions for underground dam construction. In the present paper, five criteria are considered. Finally, the latter level usually consists of alternatives. Different levels of the hierarchy structure for locating the underground dam construction are sketched in Figure 3.

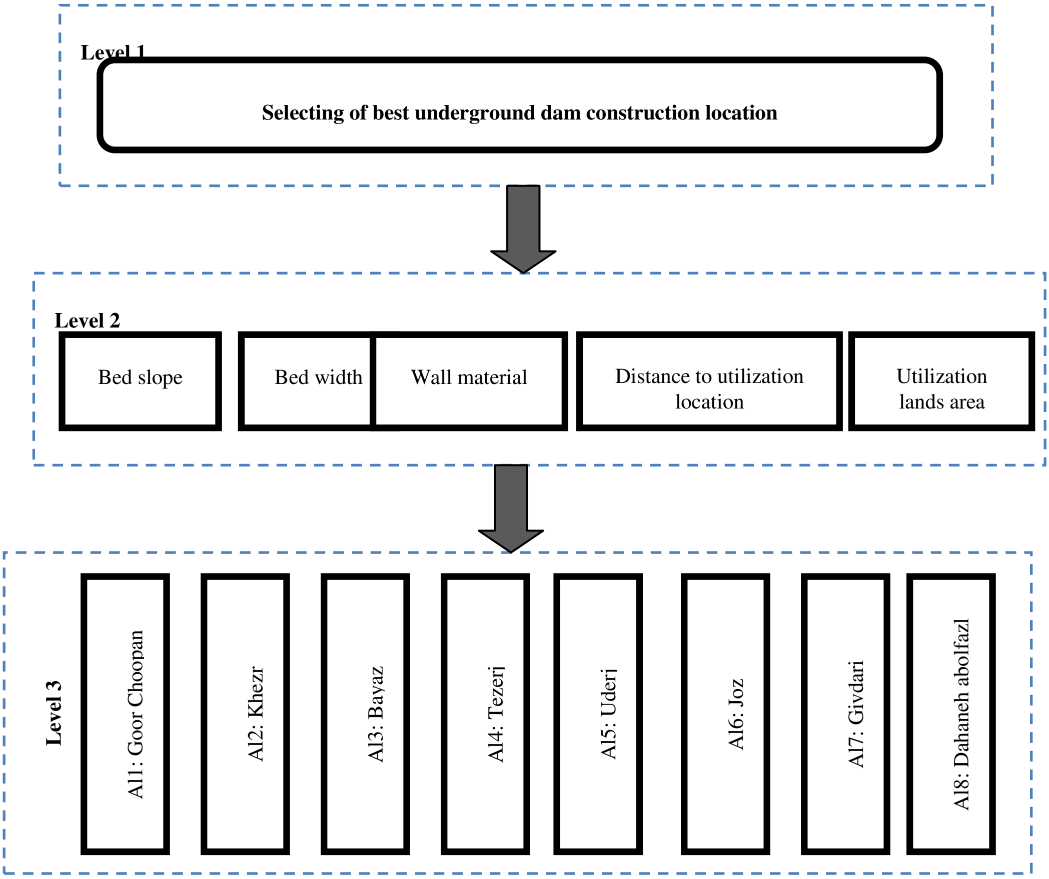


Figure 3

Hierarchical structure for underground dam construction

Step 2: *Collecting experts' judgments based on fuzzy scale and establish fuzzy pair-wise comparison matrices*

Because the problem of locating underground dam construction can be modeled based on expert's judgment, experts play an important role on the reliability and accuracy of evaluating locations of underground dam construction. In this case study, the project manager decided to consider the problem of underground dam construction depending on the judgments by seven experts.

The sample questionnaire (see Nazari-Shirkouhi et al., 2011) is applied to find the weights of the criteria using experts' judgments in the form of fuzzy numbers shown in Table 1. According to the linguistic scale, underlying TFN in Table 1 and equations (4-7), the fuzzy decision matrix for criteria with respect to goals are achieved from a questionnaire filled by experts. Then, the fuzzy decision matrices are converted to fuzzy numbers in a way explained in Azadeh et al. (2011) and Nazari-Shirkouhi et al. (2011). Table 2 shows the aggregated fuzzy decision matrix of criteria (level 2).

Table 2
Aggregated fuzzy comparison of criteria (level 2) with respect to goal

Goal	Bed slope			Bed width			Wall material			Distance to utilization location			Utilization lands area		
Bed slope	1	1	1	1	2.884	5	0.167	1.474	9	1	4.121	8	0.167	0.776	8
Bed width				1	1	1	0.200	1.310	4	1	3.476	8	0.250	1.260	4
Wall material							1	1	1	0.167	1.063	4	0.250	0.693	3
Distance to utilization location										1	1	1	0.125	0.189	0.5
Utilization lands area													1	1	1

Step 3: *Defuzzifying the fuzzy pair-wise comparison matrices*

After making the fuzzy matrices for all levels, the matrices are defuzzified. Using equations (8-9) and setting α and μ to 0.5, the final defuzzified matrix (Table 2) is shown in Table 3.

Table 3
Defuzzified matrix of criteria (level 2) with respect to goal

Goal	Bed slope	Bed width	Wall material	Distance to utilization location	Utilization lands area
Bed slope	1	2.942	3.029	4.311	2.4296
Bed width	0.340	1	1.705	3.988	1.6925
Wall material	0.330	0.586	1	1.573	1.1592
Distance to utilization location	0.232	0.251	0.636	1	0.2510
Utilization lands area	0.412	0.591	0.863	3.984	1

Step 4: *Calculating Consistency rate (C.R.)*

The consistencies of fuzzy judgment matrices are evaluated using equations (12-13). Equation (11) is used to determine the maximum eigenvalue ($\lambda_{\max}$). After solving $\lambda_{\max}$ equals to 5.1703.

Step 5: is *C.R.* <0.1?

The results indicate that *C.R.* is lower than 0.1 and the decision matrix for the second level of the hierarchical structure is consistent. The *C.R.s* of all the matrices are below 0.1 which show their consistency.

Step 6: *Computing weights for pair-wise comparison matrices, priority weights for each alternative and making a best decision*

After solving equation (14), weights of the five criteria in level 2 (*W*) are shown in Table 4.

Table 4
The weights of five criteria of level 2

Criteria	Bed slope	Bed width	Wall material	Distance to utilization location	Utilization lands area
Weight	0.4163	0.219	0.1344	0.0667	0.1636

Table 5
Summaries of results for level 2 to level 3

Criteria	Weights for level 2	Weights for level 3							
		Al1: Goor Choopan	Al2: Khezr	Al3: Bayaz	Al4: Tezerj	Al5: Uderj	Al6: Joz	Al7: Givdari	Al8: Dahaneh abolfazl
Bed slope	0.4163	0.1649	0.2017	0.1264	0.0309	0.1261	0.1006	0.1612	0.0883
Bed width	0.219	0.1451	0.0755	0.0956	0.2179	0.182	0.1406	0.0259	0.1175
Wall material	0.1344	0.1968	0.1892	0.0619	0.1597	0.1878	0.0718	0.0891	0.0437
Distance to utilization location	0.0667	0.1695	0.1714	0.0202	0.1714	0.0905	0.042	0.1675	0.1675
Utilization lands area	0.1636	0.1974	0.1898	0.0617	0.1482	0.1885	0.0721	0.0985	0.0438
Final Weight		0.0132	0.0127	0.0041	0.0099	0.0126	0.0048	0.0066	0.0029

The local weights of the alternatives are calculated using equation (14). The final weights of all alternatives are shown in Table 5. The final weights of the alternatives using data of Table 5 are as follows: 0.0132, 0.0127, 0.0041, 0.0099, 0.0126, 0.0048, 0.0066, and 0.0029 for Al_1 to Al_8, respectively.

According to results, the first alternative has the highest weight and is the most proper location according to the experts' judgment in the fuzzy environment. "Goor Choopan" and "Dahaneh abolfazl" locations are suggested as the first and last options, respectively.

4 Validation and Verification

For validation and verification of results, the pair-wise comparison matrices are run in the crisp state (standard AHP). The local weights of criteria in the second hierarchical level (AHP) are shown in Table 6.

Table 6

The weights of five criteria of level 2 (AHP)

Criteria	Bed slope	Bed width	Wall material	Distance to utilization location	Utilization lands area
Weight	0.42	0.22	0.12	0.06	0.16

Table 7

Summaries of results (AHP)

Criteria	Weights for level 2	Weights for level 3							
		Al1: Goor Choopan	Al2: Khezr	Al3: Bayaz	Al4: Tezerj	Al5: Uderj	Al6: Joz	Al7: Givdari	Al8: Dahaneh abolfazl
Bed slope	0.42	0.17	0.20	0.13	0.02	0.13	0.09	0.17	0.09
Bed width	0.22	0.13	0.07	0.09	0.23	0.19	0.15	0.03	0.12
Wall material	0.12	0.20	0.20	0.06	0.14	0.20	0.07	0.09	0.04
Distance to utilization location	0.06	0.16	0.17	0.02	0.17	0.10	0.04	0.17	0.17
Utilization lands area	0.16	0.20	0.20	0.06	0.14	0.20	0.07	0.09	0.04
Final Weight		0.167	0.170	0.092	0.112	0.161	0.094	0.111	0.088

As we can see in Table 4 and Table 6), the criterion 1 (Bed slope) and the criterion 4 (Distance to utilization location) are the most important and least important criteria according to their weights in both AHP and Fuzzy AHP methods, respectively. The final weights of all alternatives (AHP) are shown in Table 7.

The results of two runs (fuzzy AHP and AHP) have been compared and shown in Table 8.

Table 8

Comparison of ranks between AHP and Fuzzy AHP

Alternatives	AHP		Fuzzy AHP	
	Weight	Rank	Weight	Rank
Al1: Goor Choopan	0.167	2	0.0132	1
Al2: Khezr	0.17	1	0.0127	2
Al3: Bayaz	0.092	7	0.0041	7
Al4: Tezerj	0.112	4	0.0099	4
Al5: Uderj	0.161	3	0.0126	3
Al6: Joz	0.094	6	0.0048	6
Al7: Givdari	0.111	5	0.0066	5
Al8: Dahaneh abolfazl	0.088	8	0.0029	8

After ranking the alternatives in two states of AHP and Fuzzy AHP, the only difference is in ranks 1 and 2. In AHP method, the fist alternative (Khezr) is the best location and in the fuzzy AHP method, the second alternative (Goor Choopan) is the best underground dam construction location. As we can observe in the Table 8 not only all weights have changed but also the ranks of alternatives (locations) have changed. Using fuzzy theory for selecting the best location for underground dam construction can reduce vagueness and uncertainty that are inherent in problem.

Conclusion

In this paper, a holistic fuzzy AHP approach was proposed as a multi criteria decision making tool for evaluating and selecting the best location of underground dam construction Fuzzy sets theory was applied for selecting the best location of underground dam construction to reduce ambiguities and uncertainties inherent in the selection criteria. Bed slope, bed width, wall material, distance to utilization location, and utilization lands area were considered as the criteria. Eighth different alternatives for the location underground dam construction were considered in an actual case study. Based on the goal of underground dam construction, the proposed hierarchical structure may vary slightly. Finally an experiment and actual case has been conducted to apply the proposed methodology in evaluating and selecting the best underground dam construction location as a case by using judgments of six experts who had worked in the underground dam construction field and then the results were represented. As a result of the empirical study, we found that the fuzzy AHP is practical and holistic approach for ranking the candidates in terms of their overall performance regarding multiple criteria. In this case, fuzzy AHP provides a very useful decision-making tool to rank underground dam construction locations. It is expected that the present paper will serve as guideline for future studies and applications of locating in underground dam construction. Also, the proposed approach can be applied for other regions.

Acknowledgement

Special thanks go to Watershed Research Center, Agriculture and Natural Resources Division of the Kerman Province for cooperation in preparing this paper. In addition, helps by Ms. Haj-Seyyed-Alikhani, expert in Watershed Research Center for preparing related maps are also highly appreciated.

References

[1] Akpinar, N., Talay, I., & Gun, S. (2005). Priority Setting in Agricultural Land-Use Types for Sustainable Development. *Renewable Agriculture and Food Systems*, *20*(03), 136-147

[2] Alias, M. A., Hashim, S. Z. M., & Samsudin, S. (2009). Using Fuzzy Analytic Hierarchy Process for Southern Johor River Ranking. *Int J Adv Soft Comp Appl*, *1*(1), 62-76

[3] Anagnostopoulos, K. P., Gratziou, M., & Vavatsikos, A. P. (2007). Using the Fuzzy Analytic Hierarchy Process for Selecting Wastewater Facilities at Prefecture Level. *Journal of European Water*, *19*(20), 15-24

[4] Anagnostopoulos, K. P., Petalas, C., & Pisinaras, V. (2005). Water Resources Planning Using The Ahp And Promethee Multicriteria Methods: The Case Of Nestos River-Greece. *The 7th Balkan Conference on Operational Research (BACOR 00), Constanta, May 2000, Romania*

[5] Ascough, J. C., Maier, H. R., Ravalico, J. K., & Strudley, M. W. (2008). Future Research Challenges for Incorporation of Uncertainty in Environmental and Ecological Decision-Making. *ecological modelling*, *219*(3-4), 383-399

[6] Azadeh, A., Nazari-Shirkouhi, S., Hatami-Shirkouhi, L., & Ansarinejad, A. (2011). A Unique Fuzzy Multi-Criteria Decision Making: Computer Simulation Approach for Productive Operators' Assignment in Cellular Manufacturing Systems with Uncertainty and Vagueness. *The International Journal of Advanced Manufacturing Technology*, *56*(1), 329-343

[7] Azadeh, A., Shirkouhi, S. N., & Rezaie, K. (2010). A Robust Decision-Making Methodology for Evaluation and Selection of Simulation Software package. The International Journal of Advanced Manufacturing Technology, 47(1), 381-393

[8] Buckley, J. J. (1985). Fuzzy Hierarchical Analysis. *Fuzzy sets and systems*, *17*(3), 233-247

[9] Garagunis, C. N. (1981). Construction of an impervious diaphragm for improvement of a subsurface water-reservoir and simultaneous protection from migrating salt water. *Bulletin of Engineering Geology and the Environment*, *24*(1), 169–172

[10] Gupta, R. N., Mukherjee, K. P., & Singh, B. (1987). Design of underground artificial dams for mine water storage. *Mine Water and the Environment*, *6*(2), 1–14

[11] Iranmanesh, H., Shirkouhi, S. N., & Skandari, M. R. (2008). Risk evaluation of information technology projects based on fuzzy analytic hierarchal process. *International Journal of Computer and Information Science and Engineering, 2*(1), 38–44

[12] Kaufmann, A., & Gupta, M. M. (1988). *Fuzzy mathematical models in engineering and management science*. Elsevier Science Inc, Netherlands

[13] Kong, F., & Liu, H. (2005). Applying fuzzy Analytic Hierarchy Process to evaluate success factors of e-commerce. *International Journal of Information and Systems Sciences*, *1*(3-4), 406–412

[14] Li, L., Shi, Z. H., Yin, W., Zhu, D., Ng, S. L., Cai, C. F., & Lei, A. L. (2009). A fuzzy analytic hierarchy process (FAHP) approach to eco-environmental vulnerability assessment for the danjiangkou reservoir area, China. *Ecological Modelling*, *220*(23), 3439–3447

[15] Liou, T. S., & Wang, M. J. J. (1992). Ranking fuzzy numbers with integral value. *Fuzzy sets and systems*, *50*(3), 247–255

[16] Mei, X., Rosso, R., Huang, G. L., & Nie, G. S. (1989). Application of analytical hierarchy process to water resources policy and management in Beijing, China. *Closing the Gap between Theory and Practice*, Proceedings of the Baltimore Symposium, IAHS Publ., pp.73-83

[17] Montazar, A., & Behbahani, S. M. (2007). Development of an optimised irrigation system selection model using analytical hierarchy process. *Biosystems Engineering*, *98*(2), 155–165

[18] Montazar, A., & Zadbagher, E. (2010). An analytical hierarchy model for assessing global water productivity of irrigation networks in Iran. *Water resources management*, *24*(11), 2817–2832

[19] Nazari-Shirkouhi, S., Ansarinejad, A., Miri-Nargesi, S., Dalfard, V. M., & Rezaie, K. (2011). Information Systems Outsourcing Decisions Under Fuzzy Group Decision Making Approach. *International Journal of Information Technology & Decision Making (IJITDM)*, *10*(06), 989–1022

[20] Nilsson, A. (1988). Groundwater dams for small-scale water supply, *Intermediate Technology Publications Ltd. London*, pp. 69

[21] Okada, H., Styles, S. W., & Grismer, M. E. (2008). Application of the Analytic Hierarchy Process to irrigation project improvement: Part II. How

professionals evaluate an irrigation project for its improvement. *Agricultural Water Management*, *95*(3), 205–210

[22] Onder, H. and yilmaz, M. (2000). Underground dams: A tool of sustainable development and management of groundwater resources. European Water, 11(12), 30-40

[23] Opricovic, S. (2011). Fuzzy VIKOR with an application to water resources planning. *Expert Systems with Applications*, *38*(10), 12983–12990

[24] Pilavachi, P. A., Chatzipanagi, A. I., & Spyropoulou, A. I. (2009). Evaluation of hydrogen production methods using the Analytic Hierarchy Process. *International Journal of hydrogen energy*, *34*(13), 5294–5303

[25] Saaty, T. L. (1980). The analytic hierarchy process. 1980. McGraw-Hill, New York

[26] Shaw, K., Shankar, R., Yadav, S. S., & Thakur, L. S. (2012). Supplier selection using fuzzy AHP and fuzzy multi-objective linear programming for developing low carbon supply chain. Expert Systems with Applications, 39(9), 8182–8192

[27] Srdjevic, B. (2007). Linking analytic hierarchy process and social choice methods to support group decision-making in water management. *Decision Support Systems*, *42*(4), 2261–2273

[28] Srdjevic, B., & Medeiros, Y. D. P. (2008). Fuzzy AHP assessment of water management plans. *Water Resources Management*, *22*(7), 877–894

[29] Stirn, L. (2006). Integrating the fuzzy analytic hierarchy process with dynamic programming approach for determining the optimal forest management decisions. *Ecological modelling*, *194*(1), 296–305

[30] Takács, M. (2010). Multilevel Fuzzy Approach to the Risk and Disaster Management. *Acta Polytechnica Hungarica*, *7*(4), 91–102

[31] Tsiko, R. G., & Haile, T. S. (2011). Integrating Geographical Information Systems, Fuzzy Logic and Analytical Hierarchy Process in Modelling Optimum Sites for Locating Water Reservoirs. A Case Study of the Debub District in Eritrea. *Water*, *3*(1), 254–290

[32] Vahidnia, M. H., Alesheikh, A., Alimohammadi, A., & Bassiri, A. (2008). Fuzzy analytical hierarchy process in GIS application. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *37,* 593-596

[33] Wang, X., Chan, H. K., Yee, R. W. Y., & Diaz-Rainey, I. (2012). A two-stage fuzzy-AHP model for risk assessment of implementing green initiatives in the fashion supply chain. *International Journal of Production Economics*, *135*(2), 595–606

[34] Zadeh, L. A. (1965). Fuzzy sets. *Information and control*, *8*(3), 338–353

[35] Zheng, G., Zhu, N., Tian, Z., Chen, Y., & Sun, B. (2012). Application of a trapezoidal fuzzy AHP method for work safety evaluation and early warning rating of hot and humid environments. *Safety Science*, *50*(2), 228–239